CIRJE-F-452

# Implementing Nonparametric and Semiparametric Estimators

Hidehiko Ichimura
University of Tokyo

Petra E. Todd
University of Pennsylvania

December 2006

**Abstract**

This chapter reviews recent advances in nonparametric and semiparametric estimation, with an emphasis on applicability to empirical research and on resolving issues that arise in implementation. It considers techniques for estimating densities, conditional mean functions, derivatives of functions and conditional quantiles in a flexible way that imposes minimal functional form assumptions.

The chapter begins by illustrating how flexible modeling methods have been applied in empirical research, drawing on recent examples of applications from labor economics, consumer demand estimation and treatment effects models. Then, key concepts in semiparametric and nonparametric modeling are introduced that do not have counterparts in parametric modeling, such as the so-called curse of dimensionality, the notion of models with an infinite number of parameters, the criteria used to define optimal convergence rates, and "dimension-free" estimators. After defining these new concepts, a large literature on nonparametric estimation is reviewed and a unifying framework presented for thinking about how different approaches relate to one another. Local polynomial estimators are discussed in detail and their distribution theory is developed. The chapter then shows how nonparametric estimators form the building blocks for many semiparametric estimators, such as estimators for average derivatives, index models, partially linear models, and additively separable models. Semiparametric methods offer a middle ground between fully nonparametric and parametric approaches. Their main advantage is that they typically achieve faster rates of convergence than fully nonparametric approaches. In many cases, they converge at the parametric rate.

The second part of the chapter considers in detail two issues that are central with regard to implementing flexible modeling methods: how to select the values of smoothing parameters in an optimal way and how to implement "trimming" procedures. It also reviews newly developed techniques for deriving the distribution theory of semiparametric estimators. The chapter concludes with an overview of approximation methods that speed up the computation of nonparametric estimates and make flexible estimation feasible even in very large size samples.

Key words

Flexible modeling, Nonparametric estimation, semiparametric estimation, local polynomial estimators, smoothing parameter choice, convergence rates, asymptotic distribution theory

Other key words (less significant)

Additively separable models, index models, average derivative estimator, maximum score estimator, least absolute deviations estimator, semiparametric least squares estimator, trimming, binning algorithms

# Implementing Nonparametric and Semiparametric Estimators[1]

Hidehiko Ichimura and Petra E. Todd

November 28, 2006

[1]Prepared for Handbook of Econometrics Volume 6. This research was supported by NSF grant # SBR-9730688, ESRC grant RES-000-23-0797, and JSPS grant 18330040. We thank Yoichi Arai, James Heckman, and Whitney Newey for helpful comments. We also thank Jennifer Boober for detailed editorial comments.

# Contents

# 1 Introduction

In the last two decades significant progress has been made in the study of nonparametric and semiparametric models. This chapter describes recent advances with special emphasis on their applicability to empirical research and on issues that arise in implementation. As the coverage of the chapter is broad, and our ability limited, our discussion provides only an overview. It covers mostly cross sectional analysis emphasizing methods which have rigorous theoretical justifications, albeit in most cases only in first order asymptotic forms from frequentists' view point.[1] Nevertheless, we hope the chapter captures the basic motivations and ideas behind the developments and serves as a guide to using the methods appropriately. We begin by briefly summarizing the nature of recent progress, implications for empirical research, and some implementation issues.

## 1.1 The nature of recent progress

A major motivation for work on flexible models is the desire to avoid masking important features of the data by use of parametric models.[2] Recent progress has provided many new ways of modeling and estimating different aspects of a conditional probability distribution. For example, there are now a number of alternatives to linear regression model for modeling and estimating the conditional mean function as well as methods available for examining other features of distributions, such as conditional quantiles. Another area of advance has been in the study of models with limited dependent variables. In the early eighties, the standard approach with such models was to specify the error distribution parametrically and employ parametric maximum likelihood (ML) estimation. Recent research has shown that parametric specification of the error term is often unnecessary for consistent estimation of slope parameters. Models with simultaneity problems can also now be analyzed under weaker functional form assumptions. In these contexts and in others, model specification is beginning to be made more flexible. These developments enable empirical work to be carried out under fewer restrictions than was deemed possible twenty years ago.

Another important motivation for research on flexible models is the pursuit of a classical theme in econometrics: the study of the trade-off between efficiency and allowing for less restrictive models. We often wish to identify a parameter within the broadest class of models possible, but broadening a class sometimes comes at the expense of less efficient estimation. Recent research has clarified the trade-offs in terms of convergence rates and attainable efficiency bounds between specifying more or less restrictive models.

## 1.2 Benefits of flexible modeling approaches for empirical research

From an empirical perspective, the primary benefit of recent work in flexible modeling is a provision of new estimation methods with a better understanding of the efficiency loss associated with different modeling approaches. Another benefit is that the departure from the traditional linear

[1] For developments in studying panel data, see Arellano and Honoré (2001).

[2] See McFadden (1985). For brevity, we refer to nonparametric and semiparametric models as flexible models.

modeling framework decreases the tendency to focus on the conditional mean function as the sole object of interest. Using flexible models provides a natural way of considering other aspects of the probability distribution that may be of interest, such as conditional quantiles.[3] Research on limited dependent variable models has shown that quantile restrictions provide sharper restrictions than conditional mean restrictions for identifying model parameters.[4]

When we construct an econometric model of a dependent variable, either explicitly or implicitly, we model the form of a conditional distribution function. Sometimes the conditional distribution function is the parameter of interest, but more often we are interested in particular aspects of it, such as the conditional mean function, conditional quantile function, or derivatives of these functions, as we will see in the next section. When data on the dependent variable given some conditioning variables are directly observed for a random sample of the population, then the nonparametric methods discussed later in this chapter can be directly applied. However, often the application is not straightforward, because the conditional distribution that is observed differs from the conditional distribution in a random population. This can arise in variety of modeling situations, such as with limited dependent variable models, with models with measurement error, and with simultaneity. For example, a demand function can be represented as a conditional distribution of demand given price, but the distribution of the observed quantity-price data may differ from the conceptual conditional distribution we wish to study, because the supply side can affect the observed quantity and price as well.

When the conditional distribution of interest differs from the conditional distribution that can be measured directly from the data, there are two different approaches taken in the literature. One is to search for a source of variation in the data that can be used to identify the conceptual distribution of interest. This may require using data generated from a randomized experiment or from a so-called "natural experiment".[5] When variation of this sort is available in the data, the methods described in this chapter can often be directly applied. An alternative approach is to explicitly model the relationship between the observed distribution and the conceptual distribution of interest and then try to identify some aspects of the distribution of interest from the observed distribution. Much work has been done towards extending nonparametric methods to account for limited dependent variables, sample selectivity, and simultaneity. Section 2.2.3 provides some examples of applications of semiparametric selection models.

An additional benefit of using flexible models is that they allow for a more direct connection between the parameters of interest and the identification restrictions being exploited in estimation. For example, consider the linear regression model with the conditional mean restriction $E(y|x) = x'\beta_0$. Here $\beta_0$ represents a vector that defines the conditional mean function and also a vector that defines the derivative of the conditional mean function. Generally, in a restricted framework conceptually different parameters may coincide and there can be a discrepancy between the parameter of interest and the source of variation in the data used to estimate the parameter.

[3] See e.g. Buchinsky (1995), Chamberlain (1995), Buchinsky and Hahn (1998).

[4] Powell (1984), Manski (1985), Chamberlain (1986a), and Cosslett (1987).

[5] See Rosenzweig and Wolpin (2000) for a discussion of the use of natural experiments in economics.

Using flexible models makes more transparent the source of variation in the data that should be used to estimate the parameter of interest. For example, it is natural to estimate $\beta_0$ by ordinary least squares when it represents a vector defining the conditional mean function and to estimate it by an average derivative estimator, when it represents a vector defining the derivative of the conditional mean function. Average derivative estimators are discussed below in section 5. Actual implementation may require using a more restricted model for the curse of dimensionality problem we will discuss, however.

Finally, flexible models provide a systematic way of addressing concerns about model specification. First, they require fewer modeling assumptions, which directly eliminates the need for some specification testing. Second, they provide a formal framework for conducting the specification search. In parametric models, searches often proceed piece-meal, leaving the selection of which models to examine and the order in which to examine them up to the researcher. The route by which a particular model is chosen is often not made explicit, which makes it difficult to obtain general results about the properties of the estimators. Another difficulty is that there is no formal language for effectively communicating the domain of search, and the description of the domain is usually left up to the researcher's conscious effort. With nonparametric estimators, the class of models for which the estimation is valid is *a priori* specified, so that the domain is clear and the process by which a particular model is chosen is more transparent.

Careful researchers have always been aware of potential drawbacks of parametric models and have guarded against misspecification by examining the sensitivity of empirical results to alternative specifications and using imaginative ways of checking model restrictions.[6] The recent progress in flexible modeling makes it easier for researchers to address concerns about model specification and also to assess the variability of estimation procedures. The progress represents an important step towards replacing what has been characterized as the difficult art of model specification with a simpler, more systematic approach.

### 1.3 Implementation issues

So far we have emphasized the benefits of using flexible models. To fully realize these benefits, however, there are still some questions that need to be resolved regarding how to choose a model and an estimation method that is well suited to a particular application and how to implement the chosen estimation method.

A key consideration in using a flexible model is that greater flexibility often comes at a cost of a slower convergence rate. Thus, understanding the trade-off between flexibility and efficiency is important to choosing an appropriate estimation strategy. A barrier to implementing the new estimators is how to choose from a bewildering array of available estimators. A first impression from studying nonparametric literature is the richness in the variety of methods. In this chapter,

[6] Various formal specification tests and model selection rules have been developed. See for example Davidson and MacKinnon (1981), Hansen (1982), Hausman (1978), Newey (1985, 1987), Tauchen (1985), White (1980), and Wu (1974).

we attempt to pick up some common threads among different methods, to highlight differences and commonalities, and to discuss how each method has been theoretically justified.

Another consideration is that there is a degree of arbitrariness in many of the available estimation procedures that takes the form of unspecified parameters. The arbitrariness is not problematic for certain theoretical questions of interest, such as the question of whether a particular level of convergence rate is achievable. But the arbitrariness poses a problem when we implement the method, because different ways of specifying these parameters can greatly affect the estimates. For example, parameter estimates or asymptotic variance estimates can be highly sensitive to the choice of smoothing parameters or to different ways of trimming the data.[7] One focus of this chapter is on how to choose the values of these unspecified parameters.

A third problem we address is how to assess the variability of nonparametric and semiparametric estimators. In many empirical applications, the model used and methods applied deviate in some respects from the prototypical models and methods studied in the theoretic literature. Hence, it is important for researchers to be able to modify theories according to their needs and to derive the properties of modified versions of the estimators. For models and estimators based on moment conditions with finite dimensional parameters, Hansen (1982) and Pakes and Pollard (1989) provide results that are sufficiently general to accommodate many different kinds of modifications. For semiparametric models, some progress has also been made along similar lines. See Andrews (1994), and Newey and McFadden (1994) Ai and Chen (2003) and Chen, Linton and van Keilegom (2003), and Ichimura and Lee (2006).

Finally, another obstacle in applying flexible estimators is that they can be computationally intensive, particularly for large data sets. Because of slower rates of convergence, the methods are ideally suited for larger data sets. Yet it is precisely when sample sizes are large, say on the order of 100,000, when the computational burden of these methods can make them impractical. We discuss approximation methods that speed up estimation and provide great gains in speed, making it feasible to analyze even very large samples.

## 1.4 Overview of chapter

In section 2, we illustrate through examples drawn from different empirical literatures how flexible estimation methods have been used as an alternative or as a supplement to more traditional estimation approaches. Section 3 describes some concepts in semiparametric and nonparametric modeling and makes precise how new developments in the literature broaden the kinds of models and parameters of interest that can be considered in empirical research.

Section 4 discusses nonparametric estimation of densities, conditional mean functions, and derivatives of functions. Although fully nonparametric analysis are not often practical because of slow rates of convergence, we begin with nonparametric estimators because they serve as building

[7] "Trimming" is the practice of excluding a fraction of observations in local nonparametric estimation. Trimming is required when the density of the data is low at these observations and a nonparametric estimate would be unreliable. See Section 6.

blocks for many semiparametric estimators. We discuss how apparently different estimators are in some ways closely related and present a unifying framework for thinking about nonparametric density and conditional mean estimators.

Section 5 considers estimation of the same parameters of interest (densities, conditional mean functions, and derivatives of functions) using semiparametric modeling methods that overcome the problem of slow-convergence of fully nonparametric estimators. We describe a variety of semiparametric approaches to estimating densities and conditional mean functions. Although there are many estimators proposed for a variety of semiparametric and nonparametric models in the literature, we only discuss a subset of them. The models we cover are additively separable models, index models, and partially linear models as well as nonparametric models.

Section 6 focuses on the question of how to choose smoothing parameters and trimming methods in estimators applicable to nonparametric and semiparametric models. The problem of choosing the values of these unspecified parameters is similar to a model selection problem in a parametric context. For each estimator, we summarize existing research on how to choose the values of these parameters and describe the evidence on the effectiveness of various smoothing parameter selection methods, some of which comes from our own Monte Carlo studies.

Section 7 discusses how to assess the variability of different estimation procedures. Section 8 examines the problem of how to compute local nonparametric estimates in large samples. We describe binning algorithms that speed up computation through accurate approximation of nonparametric densities and conditional mean functions.

Section 9 concludes with a discussion of other issues left for future research.

### 1.5 Related literature

There are many useful surveys in the literature to which we will at times refer in this chapter. For excellent introduction to nonparametric literature in the book form we recommend Silverman (1986) and Fan and Gijbels (1996). Surveys by Blundell and Duncan (1998), Härdle and Linton (1994), and Yatchew (1998) cover nonparametric methods compactly. Useful surveys for semiparametric models are given by Arellano and Honoré (2001), Delgado and Robinson (1992), Linton (1995), Matzkin (1994), Newey and McFadden (1994), Powell (1994), and Robinson (1988).

Books by Bierens (1985), Härdle (1990), Prakasa-Rao (1983), Scott (1992) cover nonparametric density or regression function estimation methods. Books by Horowitz (1998), Lee (1996), Pagan and Ullah (1999), Stoker (1991), Ullah and Vinod (1993), and Yatchew (2003) cover both nonparametric and semiparametric methods. Deaton (1996) describes how nonparametric and semiparametric models are used in substantively important issues of household bahavior and policy analysis in developing countries.

Efficiency issues are dealt with concisely by Newey (1990, 1993) and in detail by Bickel, Klaassen, Ritov, and Wellner (1993). Most of the probabilistic techniques are explained by van der Vaart (1998) and van der Vaart and Wellner (1996).

## 2 Applications of Flexible Modeling Approaches in Economics

We first illustrate through several examples how flexible models have been used in empirical work, either as an alternative to more traditional estimation approaches or as a supplement to them. The examples are drawn from the literatures on estimating consumer demand functions, estimating the determinants of worker earnings, correcting for sample selection bias, and evaluating the effects of social programs. Our examples are chosen to highlight different kinds of parameters that may be of interest in empirical studies, such as densities, conditional mean and quantile functions and averages of the functions.

### 2.1 Density estimation

In many empirical studies, researchers are interested in analyzing the distribution of some random variable. Nonparametric density estimators provide a straightforward way of estimating densities. One nonparametric estimator that has already gained widespread use is the histogram estimator, which estimates the density by the fraction of observations falling within a specified bin divided by the bin width. In section 4, we discuss how the histogram relates to other nonparametric density estimators and how to optimally choose the bin width. We also present alternatives to the histogram estimator that have superior properties, such as the Nadaraya-Watson kernel density estimator for particular choices of kernel functions, which can be viewed as a generalized version of the histogram estimator.

An innovative empirical application of nonparametric density estimation methods is given by DiNardo, Fortin and Lemieux (1996), which investigates the effects of institutional and labor market factors on changes in the U.S. wage distribution over time. DiNardo et. al. (1996) write the overall wage density at time $t$, $f_w(w|t)$, in terms of the conditional wage densities, where conditioning is on a set of labor market or institutional factors, $z$, whose effects on earnings they analyze:

$$f_w(w|t) = \int_Z f_w(w|z,t) f_z(z|t) dz.$$

In their study, $z$ includes variables indicating union status, industrial sector, and whether the wage falls above or below the minimum wage. Counter-factual wage densities are then constructed by replacing $f_z(z|t)$ by a different hypothetical conditional density, $g_z(z|t)$, for the purpose of inferring the effect of changes in elements of $z$ on the wage distribution.

A traditional parametric approach to simulating wage distributions would specify a parametric functional form for the $w$ and $z$ distributions, in which case inference would only be valid within the class of models specified. The approach taken in DiNardo, Fortin and Lemieux (1996) is to estimate the densities nonparametrically, using a nonparametric kernel density estimator that will be discussed in section 4 of this chapter. Using a flexible modeling approach makes inference valid for a broader class of models and avoids the need to search for an appropriate parametric model specification for $f_w$ and $f_z$.

## 2.2 Conditional mean and conditional quantile function estimation

### 2.2.1 Earnings function estimation

In addition to studying the shape of the earnings distribution, economists are often interested in examining how changes in individual characteristics, such as education or years of labor market experience, affect some aspect of the distribution, such as the mean. An earnings specification that is widely used in empirical labor research is that of Mincer (1974), which writes log earnings as a linear function of years of schooling ($s$) and as a quadratic in years of work experience (exp) and other control variables ($z$):

$$\ln y = \alpha_0 + \rho s + \beta_1 \exp + \beta_2 \exp^2 + z'\gamma + \varepsilon.$$

This simple parametric specification captures several empirical regularities, such as concavity of log earnings-age and experience profiles and steeper profiles for persons with more years of education.[8] However, Mincer's model was derived under some strong assumptions, so it is of interest to also consider more general specifications of the earnings equation such as

$$\ln y = g(s, \exp, z) + \varepsilon,$$

where $g$ is a function that is continuous in the continuous variable (experience). Usually the $g$ function is interpreted as the conditional mean function. In Heckman, Lochner and Todd (2005), nonparametric regression methods are applied to estimate the above equation and to examine the empirical support for the parametric Mincer model. Their study finds substantial support for the parametric specification in decennial Census data from 1940-1960 but not in more recent decades.[9] Figure 1 shows the nonparametrically estimated log earnings-experience relationship for alternative schooling classes for adult males from the 1960 U.S. decennial census (the same data analyzed by Mincer, 1974). Nonparametric estimation was performed using local linear regression methods that are described in section 4 of this chapter.

{*Figure 1: Earnings-Experience Profiles by Education Level Estimated Nonparametrically by a Local Linear Regression Estimator*}

One can also interpret the $g$ function to be the conditional quantile function, in which case the nonparametric or semiparametric quantile estimation methods (Koenker and Basset (1978) and Koenker (2005)) can be applied. For example, Buchinsky (1994) applies semiparametric conditional quantile estimation methods to study changes in the U. S. Wage Structure from 1963-1987, using data from the Current Population Survey. He estimates a model of the form:

$$Y = X\beta_\theta + u_\theta,$$

where $\beta_\theta$ is a parameter that characterizes the conditional quantile. The model is estimated under the restriction that the $\theta th$ conditional quantile of $Y$ given $X = x$ is $x'\beta_\theta$.

[8] See Willis (1986) for a discussion of the use of the Mincer model in labor economics.

[9] Data from the 1940, 1950, 1960 show support for the model, but data from 1970, 1980 and 1990 show important deviations from the model, which Heckman et. al. (2005) attribute in part to changing skill prices over recent decades, which violates an assumption of the traditional Mincer model.

The estimation yields a time series of the estimated returns to education and experience at different quantiles of the earnings distribution. Buchinsky (1994) finds that the mean returns to education and experience and the returns at different quantiles generally follow similar patterns. Analysis of the spreads of the distributions reveals large changes in the 0.75-0.25 spread and that changes in inequality come mainly from longer tails at both ends of the wage distribution.

### 2.2.2 Analysis of consumer demand

Several recent studies in consumer demand analysis have made use of flexible estimation techniques in estimating Engel curves, which relate a consumer's budget share or expenditure on a good to total expenditure or income. Economic theory does not place strong restrictions on functional forms for Engel curves, so earlier research addressed the question of model specification mainly by adopting flexible parametric functional forms. Recent research by Banks, Blundell and Lewbel (1997), Blundell and Duncan (1998), Deaton and Paxson (1998), Härdle, Hildebrand and Jerison (1991) and Schmalensee and Stoker (1999), and Blundell, Browning and Crawford (2003) consider nonparametric and semiparametric estimation of Engel curves. The basic modeling framework is

$$y = g(x, z) + u,$$

where $y$ is the budget share of a good, $x$ is total expenditure or income, and $z$ represents other household or individual characteristics included as conditioning variables. Typically $g(x, z)$ is assumed to be the conditional mean function of $y$ given $x$ and $z$ so that $E(u|x, z) = 0$.

The traditional approach to estimating conditional mean functions specifies the functional form of $g$ up to some finite number of parameters. In consumer demand analysis, the Engel curve function is often assumed to be linear or quadratic in $\ln x$ and $z$ and the coefficients on the conditioning variables are estimated by ordinary least squares (OLS). A nonparametric estimation approach places no restrictions on the $g(x, z)$ relationship other than assuming that the $g(\cdot)$ function lies within a class of smooth functions (such as the class of twice continuously differentiable functions).

As discussed in section 3, with a large number of regressors fully nonparametric estimators converge at a rate that is too slow to be practical in conventional size samples. Semiparametric modeling approaches provide a more practical alternative. These methods achieve a faster rate of convergence by allowing some aspects of the $g(x, z)$ relationship to be flexible while imposing some parametric restrictions. For example, the approach taken in Banks, Blundell and Lewbel (1997), Blundell and Duncan (1998), and Deaton and Paxon (1998) is to model the budget-share-log-income relationship nonparametrically under the parametric restriction that other $z$ covariates enter in a linear, additively separable way. This yields a *partially linear model*[10]:

$$y = g(x) + z\gamma + u.$$

Engle, Granger, Rice, and Weiss (1986) considered electricity demand setting $x$ to be the temperature and $z$ captures household characteristics. In this application, the parameter of interest was

[10]Schmalensee and Stoker (1999) adopt a similar but slightly more general specification.

$g(x)$, how the electricity demand peaked as temperature varied. When $z$ are discrete variables, assuming that they enter in a linear fashion imposes only the assumption of additive separability.[11] Analogous to the Mincer example, the partially linear model may also be regarded as a conditional quantile function. Blundell, Chen, and Kristensen (2003) has considered the model allowing for endogeneity of income variable.

A variety of semiparametric estimators that allow for flexibility in different model components have been proposed in the econometrics and statistics literatures. Several classes of estimators will be discussed in section 5 of this chapter.

### 2.2.3 Analysis of sample selection

A leading area of application of flexible estimation methods in economics is to the sample selection problem. In fact, several estimators for the partially linear model were developed with the sample selection model in mind.[12] In the sample selection problem, an outcome is observed for a nonrandom subsample of the population and the goal is to draw inferences that are valid for the full population. For example, in the analysis of labor supply the outcome equation corresponds to the market wage, observed only for workers, and the selection equation corresponds to the decision to participate in the labor force. The wage model takes the form

$$w = w(x, \theta_1) + u$$

where $x$ denotes individual characteristics, $w$ is observed if the wage exceeds the individual's reservation wage, $w_r$, which is the minimum wage the individual would be willing to accept.

Under sample selection, the above model leads to the wage model of the form:

$$w = w(x, \theta_1) + \varphi(x, z) + u$$

where $\varphi(x, z) = E(u|w > w_r, x)$ is the so-called *control function* that needs to be estimated along with parameter $\theta_1$.[13] Clearly, in the above equation the functions $w(x, \theta_1)$ and $\varphi(x, z)$ could not be nonparametrically separately identified without some additional restrictions. Section 5 of this chapter considers alternative estimators for the sample selection model under different kinds of restrictions.

There have been numerous applications of the partially linear sample selection model. For example, Newey, Powell and Walker (1990) and Buchinsky (1998) apply the model to study female labor force participation. Stern (1996) uses it to study labor force participation among disabled workers. Olley and Pakes (1996) use the partially linear model to control for nonrandom firm

[11] In more recent work, Ai, Blundell and Chen (2000) consider the consumer demand model of the form

$$y = g(x + z\gamma) + z\gamma + u$$

and show that including the term $z\gamma$ both in the $g(\cdot)$ function and in the linear term is necessary to make the Engel curve consistent with a consumer demand system.

[12] The sample selection model is developed by Gronau (1973), Heckman (1976), and Lewis (1974).

[13] See Heckman (1980).

exit decisions in a study of productivity in the telecommunications industry. Some additional applications are discussed in section 5.

### 2.3 Averages of functions: evaluating effects of treatments

A common problem that arises in economics as well as many other fields is that of determining the impact of some intervention or treatment on some measured outcome variables. For example, one may be interested in estimating the effect of a job training program on earnings or employment outcomes.[14] In evaluating social programs, the average effect of the program for people participating in it (known as the mean impact of treatment on the treated) is a key parameter of interest on which many studies focus.

Let $(y_1, y_0)$ denote the outcomes for an individual in two hypothetical states of the world corresponding to with and without receiving treatment. Let $d$ be an indicator variable that takes the value 1 if treatment is received and 0 else. The outcome observed for each individual can be written as $y = dy_1 + (1-d)y_0$. The mean effect of the program for program participants with characteristic $z$ is given by $E(y_1 - y_0|d = 1, z)$. The average of this parameter for the treated $(d = 1)$ population is $E(y_1 - y_0|d = 1)$.

Clearly the first parameter is more informative than the second. However, as discussed in detail in the next section and in section 6, the conditional on $z$ parameter can be estimated nonparametrically less accurately than the second parameter can be estimated nonparametrically.

A variety of estimators have been put forth in the literature to estimate $E(y_1 - y_0|D = 1)$. One class of estimators are so-called matching estimators, which impute no-treatment outcomes for treated persons by matching each treated person to one or more observably similar untreated persons. Heckman, Ichimura and Todd (1997, 1998b) develop nonparametric matching estimators that use local polynomial regression methods to construct matched outcomes. Local polynomial regression estimators are discussed in section 4. The application of these estimators in program evaluation settings is considered in this handbook by Abbring, Heckman and Vytlacil.

## 3 Convergence rates, asymptotic bias, and the curse of dimensionality

A key motivation for developing flexible models is to achieve a closer match between the functional form restrictions suggested by economic theory, which are typically weak, and the functional forms used in empirical work. To study aspects of the conditional distribution functions, such as the conditional mean function and the conditional quantile function, the linear in parameter model is traditionally used. Let the conditioning finite dimensional random vector be $X$, and a *known* finite dimensional vector-valued function evaluated at $X = x$ be $r(x)$. Then the linear in parameter model

[14] See, e.g. Ashenfelter (1978), Bassi (1984), Ashenfelter and Card (1985), Fraker and Maynard (1987), Heckman and Hotz (1989), Heckman and Smith (1995), Heckman, Ichimura, Smith and Todd (1998a), and Heckman, Ichimura, and Todd (1997, 1998b), and Smith and Todd (2001, 2005).

specifies the conditional mean function or the conditional quantile function of a dependent random variable $Y$ by $r(x)'\theta$ for some *unknown finite dimensional vector* $\theta$. For example $x = (x_1, x_2)'$ and $r(x) = \left(1, x_1, x_1^2, x_2, x_2^2, x_1 \cdot x_2\right)'$. The ordinary least squares (OLS) estimator estimates the conditional mean function and the quantile regression estimator estimates the conditional quantile function. Alternatively, the most flexible model would specify $\theta(x)$ for some *unknown function* $\theta(\cdot)$. The unknown function itself or its derivative could be the parameter of interest.

The specification of parametric models involves two difficulties: which variables to include in the model and what functional form to use. Although nonparametric methods do not resolve the first difficulty, they do resolve the second. Thus if $\theta(\cdot)$ could be estimated with the same accuracy as that for the finite dimensional case, then there would be no reason to consider a finite dimensional parameter model. Unfortunately, that is not the case.

Recall that under very general regularity conditions, including the random sampling, most of the familiar estimators–the OLS estimator, the generalized method of moment (GMM) estimator, and the maximum likelihood (ML) estimator–have the property that $n^{1/2}(\hat{\beta} - \beta)$ converges in distribution to the mean zero random vector with some finite variance-covariance matrix as the sample size $n$ goes to infinity, where $\hat{\beta}$ denotes the estimator and $\beta$ the target parameter. This implies not only that $\hat{\beta} - \beta$ converges to 0 in probability, but that the difference is bounded with arbitrarily high probability (i.e. stochastically bounded) even when it is blown up by the increasing sequence $n^{1/2}$. In this case, we say that the difference converges to 0 with rate $n^{-1/2}$, that the estimator is $n^{1/2}$-consistent and that its convergence rate is $n^{-1/2}$. More generally, if an estimator has the property that $r_n(\hat{\beta} - \beta)$ is stochastically bounded, then the estimator is said to be $r_n$-consistent or to have convergence rate is $1/r_n$. If $r_n/n^{1/2}$ converges to zero, then the $r_n$-consistent estimator converges to $\beta$ slower than the $n^{1/2}$-consistent estimator does. When two estimators of the same parameter have different convergence rates, the one that approaches to the target faster is generally more desirable asymptotically.[15]

As discussed, there are estimators of the regression coefficient $\theta$, such as the OLS estimator, that converges with rate $n^{-1/2}$, so that $r(x)'\theta$ can be estimated with the same rate. But in the context of estimating the conditional mean function, Stone (1980, 1982) showed that any estimator of the regression function $\theta(\cdot)$ converges slower than $n^{-1/2}$.

To state the Stone's results, we need to clarify two complications that arise because the target parameter is a function rather than a point in a finite dimensional space $\mathbb{R}^d$ for some positive integer $d$. Note first that we need to define what we mean by an estimator to converge to a function. If we consider a function at a point, then the convergence rate can be considered in the same way discussed above. If we want to consider a convergence of an estimator of a regression function as a whole to the target regression function, then we need to define a measure of distance between two functions. There are different ways we can define the distance between the functions and the discussion about the convergence rate will generally depend on the distance used. Typically a norm is used to define the distance.

[15]Note that this is an asymptotic statement and the finite sample performance may be different. Clearly, it would also be desirable to have a better understanding about the sample size at which one estimator dominates the other.

To define a few examples of the norms used, let $k = (k_1, ..., k_d)$ where $k_j$ is a nonnegative integer for each $j = 1, ..., d$, and define $D^k\theta(x) = \partial^{k_1+\cdots+k_d}\theta(x)/\partial x_1^{k_1}\cdots\partial x_d^{k_d}$. Leading examples of the norms used are the $L_q$-norm for $1 \leq q < \infty$ ($\|\cdot\|_q$), the sup-norm ($\|\cdot\|_\infty$), and more generally the Sobolev norm ($\|\cdot\|_{\alpha,q}$ or $\|\cdot\|_{\alpha,\infty}$):

$$\left[\sum_{0\leq k_1+\cdots+k_d\leq\alpha}\int_{\mathcal{X}}\left|D^k\left(\hat{\theta}_n(x)-\theta(x)\right)\right|^q d\mu(x)\right]^{1/q} \quad \text{or} \quad \max_{0\leq k_1+\cdots+k_d\leq\alpha}\sup_{x\in\mathcal{X}}\left|D^k\left(\hat{\theta}_n(x)-\theta(x)\right)\right|.$$

Note that $\|\cdot\|_{0,q} = \|\cdot\|_q$ and $\|\cdot\|_{0,\infty} = \|\cdot\|_\infty$.[16]

Once a norm is defined, then consistency and hence the rate of convergence concept can be defined using one of the three standard consistency concepts, convergence in probability, convergence almost surely, and the $q$-th order moment convergence by how fast the distance between the estimator and the target function converges to 0.[17]

Which norm is more appropriate will depend on how the estimator is going to be used. For example if a function value at a point or its derivative is of interest, then $L^q$-norm is not useful because there are many functions close to a function in $L^q$-sense which does not determine the value at that point or the derivative values may be rather different. For these type of applications, the sup-norm may be used.

For any two norms, $\|\cdot\|_1$ and $\|\cdot\|_2$ in a finite dimensional space $\Theta$ there exist positive constants $C_H$ and $C_L$ such that for any $\theta$ and $\theta' \in \Theta$,

$$C_L\left\|\theta-\theta'\right\|_1 \leq \left\|\theta-\theta'\right\|_2 \leq C_H\left\|\theta-\theta'\right\|_1.$$

Hence, consistency using one norm implies consistency using another norm on the same space. For infinite dimensional spaces, this is no longer the case without any restriction on the class of functions under consideration. Thus we need to be more explicit about which norm is used to define consistency.[18]

Next we need to define the class of functions under consideration. When the target parameter is a point in $\mathbb{R}^d$, the class to which the parameter belongs is well defined. When the target parameter is a function, however, we need to be more specific about the class of functions to which the target belongs.

Stone specified a set of differentiable functions restricting the highest order derivative to be Hölder continuous. Let $\lfloor p \rfloor$ denote the maximum integer that is strictly smaller than $p$ and $\Theta_{p,C}$ be a class of functions which are $\lfloor p \rfloor$-times continuously differentiable with their $\lfloor p \rfloor$th derivative

[16] Clearly we need to restrict the class of functions so that the written objects are well defined.

[17] More generally one can define a metric on a relevant space of functions, but that generality may not be useful as we typically want the distance between $\hat{m}(x)$ and $m(x)$ and that between $\hat{m}(x)+c(x)$ and $m(x)+c(x)$ for any $c(x)$ to be the same. It is easy to see that the distance between $\hat{m}(x)$ and $m(x)$ only depends on $\hat{m}(x)-m(x)$.

[18] One might wonder if a point-wise consistency concept can be regarded as a consistency concept using a metric or a norm. Whether this is possible will depend on what the domain of $m(x)$ is and what the set of functions is. Without any restriction this is not possible.

being Hölder continuous with exponent $0 < \gamma \leq 1$: denoting $p = \lfloor p \rfloor + \gamma$

$$\Theta_{p,C} = \left\{ f; \max_{k_1+\cdots+k_d=\lfloor p \rfloor} \left| D^k f(x) - D^k f(x') \right| \leq C \cdot \left\| x - x' \right\|^{\gamma} \right\}$$

for some positive $C$.

Denote the distribution of the dependent variable $Y$ conditional on $X$ by $h(y|x,t)\,\phi(dy)$, where $\phi$ is a measure on $\mathbb{R}$ and $t$ is an unknown real-valued parameter in an open interval $J$, and $t$ is the mean of $Y$ given $X$ so that

$$\int y h(y|x,t)\,\phi(dy) = t \text{ for } x \in \mathbb{R}^d \text{ and } t \in J.$$

By the construction, $t$ varies with $x$ according to $t = \theta(x)$, where $\theta(x) \in \Theta$.

Stone (1980, 1982) considers a model with some regularity conditions which imply: (1) $t$ does not shift the support of $h$ or some other aspects of the conditional distribution than the mean, (2) the effect of a change in $t$ on the log-density is smooth (3) $h$ is bounded away from 0 at relevant points and for the case of the global case (4) $h$ has at most an exponential tail and (5) the region defining the $L^q$-norm is compact.

For the model which satisfies the regularity conditions, Stone shows that the optimal convergence rate for estimating the $m$th order derivative of $\theta(\cdot)$ point-wise or with $L^q$-norm for any $q$ with $0 < q < \infty$ depends on the dimension of the number of continuous conditioning variables $d$ and the smoothness $p$ ($p > m$) of $\theta(\cdot)$. Let $r = (p-m)/(2p+d)$. In particular he shows that the optimal rate of convergence is $n^{-r}$. For the sup-norm, he shows that the optimal rate is $(\log n/n)^r$. Note that $r < 1/2$ so that Stone's results imply that the optimal rate for estimating a regression function within a very general class of functions specified by $\Theta_{p,C}$ is slower than $n^{-1/2}$. Stone also shows that an analogous result holds for the estimation of Lebesgue densities.

If we specify a different class of functions in place of $\Theta_{p,C}$, then the optimality result may change. For example, the neural network literature considers a class of functions $\Theta_C$ representable by an inverse Fourier transform formula with finite absolute first moment:

$$\Theta_C = \left\{ \theta; \theta(x) = \int e^{i\omega \cdot x} \tilde{F}(d\omega) \text{ for some complex measure } \tilde{F} \text{ with } \int_{\mathbb{R}^d} |\omega| \left| d\tilde{F}(\omega) \right| d\omega \leq C \right\}.$$

See, for example, Barron (1993). For this class of functions, Chen and White (1999) constructs an estimator which converges in mean square with rate

$$(n/\ln n)^{-(1+2/(d+1))/[4(1+1/(d+1))]}.$$

Whether this is the best rate for $\Theta_C$ is an open question. This rate is better than the Stone's optimal rate when $p < d/2 + d/(d+1)$. This implies that not all functions which are less smooth than $d/2 + d/(d+1)$ is in $\Theta_C$. Let $[s]$ denote the largest integer which is less or equal to $s$. Barron (1993) has shown that if the partial derivatives of $\theta(x)$ of order $[d/2]+2$ are continuous on $\mathbb{R}^d$, then those functions can be considered to be in $\Theta_C$.[19]

[19] It will be useful to clarify the relationship between $\Theta_{p,C}$ and $\Theta_C$ more completely.

That the optimal rate may be slower than the regular $n^{-1/2}$-rate may be intuitive. Consider estimating the conditional mean function $\theta(x) = E(y|x)$ at a point $x$. If $X$ has a probability mass at $x$, then we can use data whose corresponding $X$ equals $x$ and construct the conditional mean function estimator at point $x$. However, if $X$ has continuous distribution and if we do not wish to presume any particular functional form in the conditional mean function, all we can make use of are data that lie close to $x$. Let it be an $\varepsilon$-neighborhood of $x$. In general we will have sample size of order $n\varepsilon^d$ if the underlying density is bounded away from 0 and finite. This implies that the variance of the sample mean will decrease with rate $1/\left(n\varepsilon^d\right)$ under i.i.d. sampling.[20] If we are to construct a consistent estimator for a large set of functions, we will have to make $\varepsilon$ smaller as sample size increases, because without making $\varepsilon$ smaller we will not be able to guarantee the estimator to be consistent for a broad class of functions specified in the set. This consideration separates nonparametric estimators from more restricted estimation. That $\varepsilon$ converges to zero implies that the variance will decrease with rate slower than $n^{-1}$ which in turn implies the estimator to converge at rate slower than the $n^{-1/2}$-rate.

This intuition can be used to gain more insight to the formula obtained by Stone. As we discussed the variance of an estimator of the mean in an $\varepsilon$-neighborhood is of order $\left(n\varepsilon^d\right)^{-1}$. On the other hand, if $\theta(\cdot)$ has smoothness $p$, then a parametric assumption of polynomial of order $\lfloor p \rfloor$ in the neighborhood will result in the bias of order $\varepsilon^p$ if we are to consider all functions in set $\Theta_{p,C}$. Thus the mean square error to the first order is, for some constants $C_1$ and $C_2$

$$\frac{C_1}{n\varepsilon^d} + C_2 \cdot \varepsilon^{2p}.$$

Minimizing this expression over $\varepsilon$ yields $r = p/(2p+d)$. If the target function is the $m$-th order derivative of $\theta(x)$, note that the bias changes to something of order $\varepsilon^{p-m}$. The variance also changes because the target changes to the difference of means divided by something of order $\varepsilon^m$.[21] Since the number of observations is still of order $n\varepsilon^d$, the mean square error expression changes to

$$\frac{C_1}{n\varepsilon^{d-2m}} + C_2 \cdot \varepsilon^{2(p-m)}.$$

Minimizing this expression with respect to $\varepsilon$ yields $r = (p-m)/(2p+d)$.

The result means that if we can only restrict ourselves to conditional functions with a certain degree of smoothness, then we can estimate the function with a slower rate than the $n^{-1/2}$-rate which depends on three factors: the number of continuous regressors, underlying smoothness of the target function, and the order of the derivative of the target function itself. The result is in sharp contrast to the situation where we obtain the convergence rate $n^{-1/2}$ regardless of these factors in estimating regression function or its derivatives under random sampling.

[20] An uncritical assertion we take for granted is that the mean of $y$ whose corresponding regressors are in the neighborhood is the best estimator of the $\theta(x_0)$.

[21] For example

$$\lim_{\epsilon \to 0} \frac{f(x+\epsilon) - 2f(x) + f(x-\epsilon)}{\epsilon^2} = f''(x).$$

The above analysis makes clear the reason the convergence rate for the non-parametric case is slower than for the parametric case. It is because we need to make $\varepsilon$ converge to zero to reduce the potential bias for a broad class of functions and the number of data points in the shrinking $\varepsilon$-neighborhood grows slower than the sample size. The sample size within an $\varepsilon$-neighborhood also grows more slowly when the dimension is high. When the underlying function is smooth, $\varepsilon$ can be shrunk less rapidly to reduce the potential bias. The fact that the standard error decreases with the square root of the relevant sample size (sample size within $\varepsilon$-neighborhood) does not change.

In the above discussion, we observed that the extent to which the small neighborhood approximates the underlying function depends on the smoothness of the function itself. That the function is only an approximation and thus there is an approximation error even in the neighborhood distinguishes nonparametric or semiparametric approach from the parametric approach. For example, consider estimating a one dimensional regression function. One flexible estimator that could be used is a nonparametric power series expansion estimator (described in section 5), which estimates the regression function by a finite power series. For the estimator to be consistent, the order of the polynomials must increase with the sample size to cover all potential models. But for any finite sample size, the number of polynomial terms used is fixed so that superficially the estimator appears to be the same as a standard regression problem. The key distinction between whether we have a parametric or a nonparametric model in mind is whether the estimator is considered to have a negligible bias relative to the rate of convergence or not. If we regard the estimator only as an approximation to the true regression function, then the model is nonparametric and there is a bias that needs to be taken into account in conducting inference which results in slower convergence rate. Admitting the possibility of misspecification leads us to use a more conservative standard error as the convergence rate is slower than the standard $n^{-1/2}$ rate and the form of variance will be different as well.

The dependence of the convergence rate on the dimension in particular is often referred to as *the curse of dimensionality*, which limits our ability to examine conditional mean functions or Lebesgue densities in a completely flexible way. This limitation of fully flexible models has motivated the development of semiparametric modeling methods, which offer a middle ground between fully parametric and fully nonparametric approaches. For a clarifying discussion of the definition of semiparametric models we refer the readers to Powell (1994) section 1.2. Note that the discussion so far concentrated on the estimation of the conditional mean function but results would be analogous for the estimation of the conditional quantile function.

Interestingly, not all nonparametric estimation of functions face the curse of dimensionality. A leading example is the cumulative distribution function. As it can be expressed as the mean of a random variable defined using an indicator function, the finite dimensional cumulative distribution can be estimated with $n^{-1/2}$-rate nonparametrically. We will briefly discuss a necessary condition for the $n^{1/2}$-consistent estimability of the parameter under consideration in the next sub-section as a part of the discussion of how the curse of dimensionality has been addressed in the literature.

## 3.1 Semiparametric approaches

The curse of dimensionality has been addressed using one of the following three approaches: by restricting the class of considered models, by changing the target parameter, and by changing the stochastic assumption maintained. We shall see that all three approaches can be understood within a single framework, but first we will discuss each of the concrete approaches in turn.

### 3.1.1 Using semiparametric models

The first approach is to impose some restrictions on the underlying models. Leading semiparametric models of the conditional functions are the additive separable model, the partially linear model, and the single and the multiple index models. These models provide ways to strike a balance between the flexibility and the curse of dimensionality.

The additive separable model is

$$\theta(x) = \phi_1(x_1) + \cdots + \phi_k(x_k),$$

where $\phi_j$ $(j = 1, ..., k)$ are unknown functions and $x_j$ are sub-vectors of $x$ with different dimension.

For the additive model for the conditional mean function, Linton and Nielsen (1995), Linton (1997), and Huang (1998) constructed an estimator of $\phi_j(x_j)$ which converges with the rate that depends only on the number of continuous regressors and smoothness of $\phi_j(\cdot)$. Thus the convergence rate for estimating $\theta(x)$ is driven by the maximum number of continuous regressors in $\{\phi_j(\cdot)\}$ assuming the same degree of smoothness for each of the component functions. For the conditional quantile function, Horowitz and Lee (2005) has constructed an estimator with analogous properties.

The partially linear model is

$$\theta(x) = r(x_0)'\beta + \phi(x_1),$$

where $r(\cdot)$ is a known function, $\phi(\cdot)$ is an unknown function and $x_0$ and $x_1$ are sub-vectors of $x$. Robinson (1988) shows that when $\theta(x)$ is the conditional mean function, $\beta$ can be estimated with $n^{-1/2}$-rate regardless of the number of regressors in $x_1$ and constructs an estimator of $\phi$ which performs as if $\beta$ were known. Thus the convergence rate for estimating $\theta(x)$ is driven by the number of continuous regressors in $x_1$ and smoothness of $\phi(\cdot)$.

The multiple index model is

$$\theta(x) = r_0(x_0)'\beta_0(\theta) + \phi\left(r_1(x_1)'\beta_1(\theta), \ldots, r_k(x_k)'\beta_k(\theta)\right),$$

where $r_j$ $(j = 0, 1, ..., k)$ are known functions and $\phi(\cdot)$ is an unknown function. Note that the multiple index model reduces to the partially linear model when $\beta_j(\theta)$ $(j = 1, ..., k)$ are known. Ichimura (1993) constructed an estimator of $\beta_1$ in a single-index model without $\beta_0$. Using the same idea, Ichimura and Lee (1991) shows that $\theta$ can be estimated with $n^{-1/2}$-rate regardless of the dimension of unknown function $\phi$. It is straightforward to show that the estimation of $\phi$ can be done as if $\beta_j(\theta)$ for $j = 0, ..., k$ are known. For the single index model, Blundell and Powell

(2003) develops a method to allow for an endogenous regressor and Ichimura and Lee (2006) studies the asymptotic property of Ichimura's estimator under general misspecification. Ichimura and Lee (2006) also examines the single index model under a quantile restriction, rather than the conditional mean restriction and shows that results analogous to the conditional mean restriction case hold.

We will discuss these models and accompanying estimation methods in some detail in section 5. The advantage of using these models is clear. Because the parameters are estimated without being subject to the curse of dimensionality and because these models typically include the linear in parameter specification as a special case, they permit examining the conditional mean and quantile functions under less stringent conditions than previously thought possible.

There are at least two limitations in using semiparametric models. First, we do not know which of these three or an alternative semiparametric model to use. Second, there could be a discrepancy between the parameter we want to estimate and the variation we would use to estimate the parameter. As Powell (1994) has emphasized, the defining characteristic of a semiparametric model is that there are different ways to express the same parameters. For example, consider the partially linear model with $r(x_0) = x_0$ and assume that $x_0$ and $x_1$ do not have a common variable and that all relevant moments are finite. In this case, $\beta$ is the partial derivative of $\theta(x)$ with respect to $x_0$. When $\theta(x) = E(y|x)$, $\beta$ is also a solution to minimizing $E[Var(y - x_0'b|x_1)]$ with respect to $b$ and at the same time $\beta$ is a solution (corresponding to $b$) to minimizing $E\left[(y - x_0'b - f(x_1))^2\right]$ with respect to $b$ and a measurable function $f$.[22] Thus one can estimate $\beta$ using any of the sample counterpart of these observations. Depending on how the estimator is going to be used, we may want to use different estimation method but using semiparametric model tends to mask this point because $\beta$ is $\beta$ within a semiparametric model. The second limitation can be overcome by carefully choosing the appropriate estimation method, but he first limitation seems harder to resolve at this point.

### 3.1.2 Changing the parameter

The second approach is to shift the focus of estimation to an aspect of $\theta(\cdot)$ rather than $\theta(\cdot)$ itself. This approach does not restrict the class of functions we consider to a parametric or a semiparametric model. For example Schweder (1975), Ahmad (1976), Hasminskii and Ibragimov (1979) studied estimation of $\int \theta(x)^2 dx$ where $\theta(x)$ is a Lebesgue density of a random variable. This object is of interest in studying rank estimation of a location parameter and also studying optimal density estimation. The parameter can be estimated at the $n^{-1/2}$-rate thus the curse of dimensionality can be avoided.

Stoker (1986) considers average derivative of the form $\int \{\partial\theta(x)/\partial x\} w(x) dx$ where $w(x)$ is a given weight function. Even though $\partial\theta(x)/\partial x$ itself cannot be estimated point-wise at the $n^{-1/2}$-rate Powell, Stock, and Stoker (1989) and Robinson (1989) showed that this type of parameter can be estimated with $n^{-1/2}$-rate regardless of the dimension of $x$. By changing the weighting function $w(x)$ appropriately, the average derivative parameter can inform us about different aspects of

[22] The latter two problems lead to the same solution for $b$ even in a nonparametric setup.

$\partial\theta(x)/\partial x$. Altonji and Ichimura (1998) has studied average derivative estimation when dependent data are observed with censoring. We will discuss average derivative method in some detail in section 5.

As previously discussed, DiNardo, Fortin and Lemieux (1996) studies a density $f(x)$ via conditional density $\theta(x,z)$ and the marginal distribution of $z$, $F_z(z)$:

$$f(x) = \int_Z \theta(x,z)\, dF_z(z).$$

They study various hypothetical wage densities by replacing $F_z(z)$ with hypothetical marginal distributions. In their application $z$ consists of discrete variables. Thus both $f(x)$ and $\theta(x,z)$ are estimated with the same rate. But if $z$ contains a continuous variable, then this is an example in which integration improves the rate of convergence. This is also the case for Heckman, Ichimura, Smith, Todd (1998). In their work $\theta(\cdot)$ is the conditional mean function and $F_z(z)$ is replaced by a distribution which is estimated.

### 3.1.3 Specifying different stochastic assumption within a semiparametric model

Even when the model is restricted to a semiparametric model which has a finite dimensional parameter, such as $\beta$ in the partially linear regression model, it is not always possible to estimate the finite dimensional parameter with the standard $n^{-1/2}$-rate. The role that different stochastic assumptions can play in this regard is clarified in the context of the censored regression model by Powell (1984) and Chamberlain (1986a) and Cosslett (1987). An illustration of the results requires us to fully specify the probability model.

A probability model is specified by a class of conditional or unconditional distribution of a random variable $z$, say $\mathcal{F}$. To distinguish conditional and unconditional models, we write $z = (y,x)$ where $x$ represents conditioning variables. Let $\mathcal{F}_x$ denote a conditional probability model. Sometimes $\mathcal{F}$ is specified indirectly as a known mapping, say $h$, from another parameter space $\Theta$ into a space of distributions, $\mathcal{F} = \{f : f(z) = h(z;\theta), \theta \in \Theta\}$. This is the conventional way the standard parametric model specifies $\mathcal{F}$. When the indirect specification of a probability model can be accomplished based on a finite dimensional space $\Theta$ in some 'smooth' way, the model is called a parametric model.[23]

Consider, for example, the censored linear regression model censored from below at 0, with only an intercept term. In this case the model of the distribution of $y$ is

$$\mathcal{F} = \left\{ f;\ f(y) = h(y-\mu)^{1\{y>0\}} \left[\int_{-\infty}^{-\mu} h(s)ds\right]^{1\{y=0\}}, h \in \Gamma \right\},$$

where $\Gamma$ is a class of densities with certain stochastic properties. The parameter space is $\Theta = \mathbb{R} \times \Gamma$. In the econometric literature in the past it was common to treat the parameter space as $\mathbb{R}$ leaving

[23] Without a smoothness restriction on the mapping from the finite dimensional parameter space to the space of probability distributions, the definition of the parametric model is not meaningful. Without smoothness one can 'encode' an infinite dimensional space into a finite dimensional space.

the nonparametric component $\Gamma$ implicit. Specifying the probability model completely turned out to be an important step towards understanding the convergence rate and efficiency bound of a semiparametric estimator.

As an illustration consider estimating $\mu$ semiparametrically under two alternative stochastic restrictions on $\Gamma$ under random sampling. One model restricts that $h$ has mean 0 and the other model restricts that $h$ has median 0. We will argue that the first stochastic assumption will not allow us to estimate $\mu$ with $n^{-1/2}$ rate but the second assumption will.

To see this, suppose $h$ is known. Then under random sampling, the most efficient estimator is the ML estimator and its asymptotic variance in this case is 1 over

$$\frac{h^2(-\mu)}{H(-\mu)} + \int_{-\mu}^{\infty} \frac{[h'(s)]^2}{h^2(s)} h(s)ds,$$

where $H(t) = \int_{-\infty}^{t} h(s)ds$. Note that the first term can be made arbitrarily small under both models. Under the model with mean 0 restriction, the second term can be made arbitrarily small also because only a small probability needs to be on $[\mu, \infty)$ to satisfy the mean 0 restriction. However, with the median restriction, when $\mu > 0$, and $H(-\mu) < 1$, for example, the second term is strictly positive. To see this, note that the second term divided by $1 - H(-\mu)$ corresponds to the inverse of the asymptotic variance of the ML estimator of the mean when the random variable under consideration is supported on $[-\mu, \infty)$. Since we know that the mean can be estimated with rate $n^{-1/2}$ when the variance is restricted to be finite the infimum of the second term cannot be 0. Thus with the restrictions on $\Gamma$, the infimum over $\Gamma$ of the second term should be strictly positive so that the asymptotic variance is bounded above. Thus whether the conditional mean or the conditional median, or more generally the conditional quantile is restricted to zero makes a fundamental difference.

We intuitively argued that the bound on the asymptotic variance of any estimator of $\mu$ could be obtained by considering the worst case among $\Gamma$ after computing the smallest asymptotic variance of a possible estimators of $\mu$ given a particular function in $\Gamma$. This is the approach of Stein (1956) further developed by various researchers. The work is summarized in Bickel, Klaassen, Ritov, Wellner (1993). Newey (1990) provides a useful survey of the literature as do van der Vaart and Wellner (1996) and van der Vaart (1998). It has been shown that the bound thus computed provides a lower bound of the asymptotic variance of the $n^{1/2}$-consistent "regular" estimators where regularity is defined to exclude super-efficient estimators as well as estimators that use an unknown aspect of the probability model under consideration. When the bound is infinite, then there is no $n^{1/2}$-consistent estimator. A finite found also does not imply that $n^{1/2}$-consistent estimator exists, because it may not be achievable. See Ritov and Bickel (1990) for examples. On the other hand, when there is a regular estimator that achieves the bound, then it is reasonable to call the estimator efficient.[24] For the example considered above, the estimator considered by Powell (1984) gives an example that achieves the $n^{1/2}$-consistency and Newey and Powell (1990) constructs an

[24]For an alternative formulation of an efficiency concept that does not restrict estimators to the regular estimators, see van der Vaart (1998) chapter 8.

asymptotically efficient estimator for the model.

To some extent, these developments partly solve the specification search problem that was described in the introduction. For the censored regression model, for example, the specification search for the error distribution has become completely redundant as the slope parameters can be estimated at the parametric rate without specifying a functional form for the error distribution. However, search problems still remain for the specification of the systematic component of the model. For the average derivative example, the specification search problem reduces to that of fully nonparametric models: the main difficulty being which variables to use and not which functional form to adopt.

In a parametric setting, specification search often makes it difficult to assess the variability of the resulting estimator. In contrast, there are now large classes of semiparametric and nonparametric models for which at least asymptotic assessment of the variability of estimators is possible. Not only has consistency been proved for many estimators, but the explicit form of the asymptotic bias and variance has also been obtained.

# 4 Nonparametric Estimation Methods

While the above discussion of the curse of dimensionality may leave one with an impression that nonparametric methods are useful only for a low dimensional cases, they are nonetheless important to study, if only because they form the building blocks of many semiparametric estimators.

Roughly speaking, there are two types of nonparametric estimation methods: local and global. These two approaches reflect two different ways to reduce the problem of estimating a function into estimation of real numbers. Local approaches consider a real valued function $h(x)$ at a single point $x = x_0$. The problem of estimating a function becomes estimating a real number $h(x_0)$. If we are interested in evaluating the function in the neighborhood of the point $x_0$, we can approximate the function by $h(x_0)$ or, if $h(x)$ is continuously differentiable at $x_0$, then a better approximation might be $h(x_0) + h'(x_0)(x - x_0)$. Thus, the problem of estimating a function at a point may be thought of as estimating two real numbers $h(x_0)$ and $h'(x_0)$, making use of observations in the neighborhood. Either way, if we want to estimate the function over a wider range of $x$ values, the same, point-wise problem can be solved at the different points of evaluation.

Global approaches introduce a coordinate system in a space of functions, which reduces the problem of estimating a function into that of estimating a set of real numbers. Recall that any element $v$ in a $d$-dimensional vector space can be uniquely expressed using a system of independent vectors $\{b_j\}_{j=1}^d$ as $v = \sum_{j=1}^d \theta_j \cdot b_j$, where one can think of $\{b_j\}_{j=1}^d$ as a system of coordinates and $(\theta_1, ..., \theta_d)'$ as the representation of $v$ using the coordinate system. Likewise, using an appropriate set of linearly independent functions $\{\phi_j(x)\}_{j=1}^\infty$ as coordinates any square integrable real valued function can be uniquely expressed by a set of coefficients. That is, given an appropriate set of linearly independent functions $\{\phi_j(x)\}_{j=1}^\infty$, any square integrable function $h(x)$ has unique

coefficients $\{\theta_j\}_{j=1}^{\infty}$ such that

$$h(x) = \sum_{j=1}^{\infty} \theta_j \cdot \phi_j(x).$$

One can think of $\{\phi_j(x)\}_{j=1}^{\infty}$ as a system of coordinates and $(\theta_1, \theta_2, ...)'$ as the representation of $h(x)$ using the coordinate system. This observation allows us to translate the problem of estimating a function into a problem of estimating a sequence of real numbers $\{\theta_j\}_{j=1}^{\infty}$.

Well known bases are polynomial series and Fourier series. These bases are infinitely differentiable everywhere. Other well known bases are polynomial spline bases and wavelet bases. One dimensional linear spline bases are: for a given knot locations $t_j$, $j = 1, \ldots, J$

$$1, x, (x - t_j)1\{x \geq t_j\},$$

quadratic spline bases are:

$$1, x, x^2, (x - t_j)^2 1\{x \geq t_j\},$$

and cubic spline bases are:

$$1, x, x^2, x^3, (x - t_j)^3 1\{x \geq t_j\}.$$

By making the knot locations denser, a larger class of functions can be approximated. A function represented by a linear combination of the linear spline bases is continuous, that represented by the quadratic spline is continuously differentiable, and that represented by the cubic spline is twice continuously differentiable. Higher dimensional functions can be approximated by an appropriate Tensor product of the one dimensional bases. Polynomial spline bases have an unpleasant feature that imposing higher order of smoothness requires more parameters.

Wavelet bases are generated by a single function $\phi$ and written as

$$2^{k/2}\phi(2^k x - \ell)$$

where $k$ is a nonnegative integer and $\ell$ is any integer and $\phi$ satisfies certain conditions so that $\{2^{k/2}\phi(2^k x - \ell)\}_\ell$ is an orthonormal family in $L^2$-space. Now many functions $\phi$, including an infinitely differentiable function, are known to define the orthonormal bases. Since these functions themselves can be infinitely differentiable and yet can approximate any function in $L^2$-space, the bases are useful to examine functions without known degree of smoothness. See Fan and Gijbels (1996) for a concise discussion of the wavelet analysis. For a fuller discussion see Chui (1992) and Daubechies (1992).

Below, we illustrate both local and global approaches to density and conditional mean function estimation. We emphasize commonalities among estimation approaches that on the surface may appear very different. While we believe it is useful to understand the local and global nonparametric approaches, we shall see that even that distinction is not as clear cut as it seems at first.

## 4.1 How do we estimate densities?

As with parametric estimation, nonparametric estimation of a density can be carried out using either likelihood based estimation or a moment based estimation. Here we classify various density estimators, using the maximum likelihood vs method of moment classification in addition to the local vs global classification.

### 4.1.1 Moment based estimators

If there were a standard function $\delta_x(s)$, such that for any continuous function $f$

$$\int_{-\infty}^{+\infty} \delta_x(s) f(s) ds = f(x),$$

then by regarding $f$ as the Lebesgue density function of a random variable $X$, this equality can be used as the moment condition

$$E\{\delta_x(X)\} = f(x)$$

to estimate the density. Unfortunately, it is well known that such function $\delta_x(s)$, called the Dirac-delta function, does not exist as a standard function.[25] However, it can be expressed as a limit of a class of standard functions indexed by a positive real number $h$, say $\delta_x(s,h)$.[26] For example

$$\delta_x(s,h) = \frac{1}{h} K\left(\frac{x-s}{h}\right)$$

where $\int_{-\infty}^{+\infty} K(u)du = 1$ satisfies the requirement for a continuous Lebesgue densities if $\lim_{|u|\to\infty} |u|K(u) = 0$.

Method-of-moment estimation based on this specification for $\delta_x(s,h)$ leads to the so called kernel density estimator of Rosenblatt (1956). See also Parzen (1962).:

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - x_i),$$

where $K_h(s) = h^{-1}K(s/h)$. When the function $K(\cdot)$ is a density function, the estimator itself is a density function. Smoothness on estimated density function can be imposed by choosing a smooth function $K(\cdot)$. See Silverman (1986) for a very useful discussion of the estimator.[27]

Implementing this estimator requires specifying the function $K(\cdot)$, referred to as the kernel function, and the parameter $h$. The parameter $h$ is called the *window-width, bandwidth*, or *smoothing parameter*.

When a symmetric kernel function with a finite variance is used, a calculation using the change of variables formula and the Taylor's series expansion under the assumption that the density is

[25] See for example, Zemanian (1965), p.10.

[26] See Walter and Blum (1976).

[27] Härdle and Linton (1994) summarizes the asymptotic properties of this estimator in chapter 38. See also Scott (1992).

twice continuously differentiable shows that the highest order of the point-wise mean square error of the kernel density estimator is

$$(h^2/2 \int u^2 K(u) ds f''(x))^2 + \frac{1}{nh} \int K^2(u) du f(x).$$

The bandwidth that minimizes the leading terms of the mean squared error is

$$h^* = \left[ \frac{\int K^2(u) du f(x)}{(\int u^2 K(u) ds)^2 [f''(x)]^2} \right]^{1/5} n^{-1/5}.$$

The optimal bandwidth is larger when the density is high, because then the variance is higher; the optimal bandwidth is larger when the second derivative is small, because then the bias is smaller so that wider bandwidth can be tolerated. Because the optimal bandwidth involves unknown density itself and its second derivative, it is not feasible. However, there is a large literature that we review in section 6 that studies methods that use the data to come close to the optimal bandwidth.

With the optimal bandwidth the highest order of the mean square error is

$$\frac{5}{4} (\int K^2(u) du)^{4/5} (\int u^2 K(u) du)^{2/5} f^{4/5}(x) [(f''(x))^2]^{1/5} n^{-4/5}.$$

This shows three things: first, the convergence rate of the kernel density estimator is $n^{-4/5}$, which corresponds to the optimal rate Stone obtained for the estimation of one-dimensional twice continuously differentiable densities. Second, regardless of the unknowns, optimal kernel function can be chosen by minimizing

$$(\int K^2(u) du)(\int u^2 K(u) du)^{1/2},$$

under the restriction that the kernel function is symmetric and the second moment is finite and normalized to 1, Epanechnikov (1969) showed that the optimal kernel function is

$$K(u) = \frac{3}{4 \cdot 5^{3/2}} (5 - u^2) 1\{u^2 \leq 5\}.$$

This kernel function is usually refereed to as the Epanechnikov kernel.[28] The envelope theorem implies that a slight deviation from the optimal kernel function would not affect the asymptotic mean square error very much. In fact, Epanechnikov showed numerically that the efficiency loss by using commonly used kernel functions such as the normal kernel is about 5% and that by the uniform kernel is about 7%. This observation lead subsequent researches to concentrate more on how to choose the bandwidth sequence. Note that the Epanechnikov kernel is not differentiable at the edges of its support. If we impose three times continuous differentiability via the quartic kernel function, sometimes called the biweight kernel,

$$K(u) = \frac{15}{16 \cdot 7^{5/2}} (7 - u^2)^2 1\{u^2 \leq 7\},$$

[28] Sometimes the support is normalized between $-1$ and 1 rather than the variance being normalized to 1. However, this normalization will make the comparison to the kernel function with unbounded support difficult.

the efficiency loss to the first order is less than 1%.[29]

The *histogram estimator* can be viewed as a kernel density estimator which uses a uniform kernel function $K_h(s) = 1(|s| < h)/2$. Although the simplicity of the histogram is appealing and it can be interpreted as an estimator of the cell probability divided by twice the bandwidth for each finite observation, it has two disadvantages; one is that density estimates generated by a histogram are discontinuous at bin end-points, and the other is that there is about 7% efficiency loss discussed above. Figure 2 compares the density of earnings estimated by a histogram to that estimated using a kernel density estimator.

{*Figure 2: Comparison of earnings density estimated by a histogram and by a kernel density estimator*}

Another density estimator which can be viewed as a kernel density estimator is the nearest neighbor estimator. The estimator is based on the equality $\int_{x_0-R_n}^{x_0+R_n} f(s)ds = \Pr\{|X - x_0| \leq R_n\}$, where $f$ is the Lebesgue density of random variable $X$. Because the left hand side is approximately $2R_n f(x_0)$ and the right hand side can be estimated by the fraction of observations which fall within the $R_n$ distance from $x_0$, by using the distance $R_n$ to the nearest $k_n$ observations from $x_0$, the density at $x = x_0$ can be estimated by equating $2R_n f(x_0)$ and $k_n/n$; i.e. by $k_n/(2R_n n)$. This can be written as

$$n^{-1} \sum_{i=1}^{n} K_{R_n}(x - x_i)$$

where the kernel function is the uniform kernel function.[30] Thus the nearest neighbor estimator can be viewed as a histogram estimator for a particular way of choosing the bandwidth. Note that the way bandwidth is selected does not consider the second derivative of the density at the point of estimation, so when the density is twice continuously differentiable the nearest neighbor estimator cannot be a optimal.

The estimators discussed so far are all local estimators. We next show that method of moment based global estimators can be viewed also as a local estimator. As discussed earlier, let $\{\phi_j(x)\}_{j=1}^{\infty}$ be an orthonormal bases in the space of square integrable functions and consider the class of Lebesgue densities in the same space. Then one can write

$$f(x) = \sum_{j=1}^{\infty} c_j \phi_j(x)$$

for some sequence $\{c_j\}_{j=1}^{\infty}$. The coefficients can be computed by

$$\int f(x)\phi_k(x)dx = \sum_{j=1}^{\infty} c_j \int \phi_j(x)\phi_k(x)dx = c_k,$$

where the last equality follows from the orthonormality of $\{\phi_j(x)\}_{j=1}^{\infty}$.

Thus a global method to estimate the Lebesgue density in $L_2$ is to use the first $J$ elements of the series just discussed and estimating $c_j$ by the sample average of $\phi_j(X)$ where $X$ has the Lebesgue

[29] See for example, Scott (1992), Table 6.2.

[30] See Moore and Yackel (1977a, b).

density $f(x)$. In this case the estimator of $c_j$ is $\hat{c}_j = n^{-1}\sum_{i=1}^{n}\phi_j(x_i)$ so that the estimator of $f(x)$ is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}\phi_j(x_i)\phi_j(x).$$

This results in another example of $\delta_x(s,h)$. Denoting the number of series used by $J = 1/h$:

$$\delta_x(s,h) = \sum_{j=1}^{1/h}\phi_j(s)\,\phi_j(x).$$

This form of an approximation to the delta function is known as a reproducing kernel. See Weinert (1982), Saito (1989), Wahba (1990), and Berlinet and Thomas-Agnan (2003). For example, when we consider densities supported on $[-\pi,\pi]$ and 0 at the boundaries, we can use $1/(2\pi), \cos(x)/\pi, \sin(x)/\pi, \cos(2x)/\pi, \sin(2x)/\pi, ...$ as the orthonormal bases. In this case one can show that $\delta_x(s, 1/J)$ is the Dirichlet kernel:

$$\delta_x(s,1/J) = \frac{1}{2\pi}\frac{\sin\frac{2J+1}{2}(s-x)}{\sin\frac{s-x}{2}}.$$

Figure 3 plots this function.

{*Figure 3: implicit kernel function for the Fourier series density estimator*}

We are not advocating using the series estimator as discussed above. In fact this simple version of the implementation has been shown to have undesirable features that have been improved. For a discussion see Scott (1992).

A notable difference between the kernel density and series expansion estimators is that kernel functions that correspond to orthogonal expansion methods have support independent from the number of terms in the expansion, whereas standard kernel functions have a support that depends on the bandwidth choice if the kernel function is supported on a finite interval.

For the general series estimators, the highest order of the bias and the variance have not been characterized although the rate of convergence have been characterized. For the wavelet based bases, however, the highest order of the bias and the variance are computed by Hall and Patil (1995). See also Huang (1999) and Ochiai and Naito (2003).

We have seen that moment based density estimators can be regarded as reflecting different ways to approximate the delta function. A single parameter $h$ in the approximation $\delta_x(s,h)$ is used to construct a model of densities. If $\int \delta_x(s,h)dx = 1$ and $\delta_x(s) \geq 0$, then an estimator itself is a valid density. As discussed, Epanechnikov (1969) argued that among the kernel density estimators, the choice of bandwidth is more important than the choice of kernel function. For the same reason, the above discussion may indicate that among method-of-moment based methods the more important issue is how to choose the smoothing parameter rather than which method of moment estimator to use. This remains to be seen.

### 4.1.2 Likelihood-based approaches

Another natural way of estimating densities is a maximum likelihood (ML) approach; however, a straightforward application of the likelihood method fails in nonparametric density estimation. To see why consider the ML estimator

$$\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(x_i)$$

where $\mathcal{F}$ is an *a priori* specified class of densities. If $\mathcal{F}$ is not restricted, then for each $n$ one can choose an $f$ with spikes at $x_i$ and yet $f$ can be a density. Thus the likelihood can be made as high as desired regardless of the underlying density and the method leads to an inconsistent estimator.

Many modifications have been proposed to resolve this failure by restricting $\mathcal{F}$ in some way.[31] Imposing smoothness alone does not correct the situation. To see this, first observe that the likelihood value is only affected by values of $f(x)$ on the data points $x_1$, ...,$x_n$. As one can construct a polynomial function that passes through any given finite points that are a subset of the graph of $\log f(x)$, the likelihood value can be made arbitrarily large. Stronger restrictions are needed.

As discussed below, some restrictions are needed regardless of whether one takes a global or a local approach. The global method, such as that of Stone et. al. (1997), restricts the rate at which more complex functions are included in $\mathcal{F}$ as the sample size increases. The local method attempts to approximate the density locally holding the complexity of the functions fixed. The approach taken by Hjort and Jones (1996) and Loader (1996) is to approximate a density locally by a parametric density.

**Global likelihood estimation** The global likelihood-based approach restricts the rate at which complex functions are included in $\mathcal{F}$ as the sample size increases. Here, we describe a density estimation implementation of Stone's extended linear modeling, as exposited in Stone et al. (1997). Their starting point is to observe that the log-density function can be written in the form

$$l(h, X) = h(X) - \log \int_{\mathcal{X}} \exp h(x)\, dx$$

for any function $h(x) \in H$, where $H$ is a linear space of real-valued functions on $\mathcal{X}$. The second term on the right hand side ensures that $\exp[l(h, X)]$ is a proper density.

Stone et. al. (1997) define the estimator of the log-density as the maximizer of the log-likelihood function

$$\sum_{i=1}^{n} h(X_i) - n \log \int_{\mathcal{X}} \exp[h(x)]\, dx$$

over $h$ in a finite dimensional linear subset of $H$, denoted $G$. With no restriction on $H$ to a smaller subset $G$, the problem pointed out earlier in relation to inconsistency of the unrestricted ML estimator also arises here. By choosing $h$ to have spikes at observation points we can make $\sum_{i=1}^{n} h(X_i)$ as large as we wish, while keeping the contribution to $n \log \int_{\mathcal{X}} \exp[h(x)]\, dx$ small. Also, for any constant value $C$, $h(x)$ and $h(x) + C$ give rise to the same log-likelihood value so we

[31] See Prakasa-Rao (1983), Silverman (1984), and Scott (1992).

need a normalization. Stone et. al. (1997) use the normalization $E[h(X)] = 0$, which guarantees a unique optimizer in $G$ since the log likelihood function is strictly concave. The implementation of the method depends crucially on how $G$ is chosen. The choice of $G$ represents the finite dimensional model used to approximate the unknown density. In their formulation of $d$-dimensional functions, the first stage is the additive separable model. The second stage includes two dimensional function etc. In this way, the additive separable model could be embedded in a series of less restrictive models.

**Local likelihood estimation** Loader (1996) and Hjort and Jones (1996) propose a localized likelihood based estimator. The local likelihood is defined as

$$\mathcal{L}(f, x) = \sum_{i=1}^{n} K_h(x - X_i) \log f(X_i) - n \int_{\mathcal{X}} K_h(x - X_i) f(u)\, du.$$

Because the data are localized through the use of kernel weighting, we need only to approximate the log-density locally. Loader considers polynomial approximation of the log density, which is equivalent to using exponential models. Hjort and Jones consider approximation by general parametric models. If we do not restrict the class of models to a small subset like the ones considered in these papers, then the optimization problem does not have a well defined solution.

To gain insight into the form of the above objective function, we show that one can view the objective function as an approximation to a likelihood for observing data only in an area within $h$ of point $x$. When the density is $f$, the likelihood contribution if the data falls within the interval is $f(X_i)$ but if not, then it also contributes by computing the probability of not observing in the interval. Thus we can write the likelihood as

$$\sum_{i=1}^{n} I_i \log f(X_i) + (1 - I_i) \log \left(1 - \int_{\mathcal{X}} I\{|x - u| \leq h\} f(u)\, du\right).$$

Using the approximation $\log\left(1 - \int_{\mathcal{X}} I\{|x-u| \leq h\} f(u)\, du\right) \approx -\int_{\mathcal{X}} I\{|x-u| \leq h\} f(u)\, du$ gives

$$\begin{aligned} &\sum_{i=1}^{n} I_i \log f(X_i) - n \int_{\mathcal{X}} I\{|x - u| \leq h\} f(u)\, du \\ &- \sum_{i=1}^{n} I_i \int_{\mathcal{X}} I\{|x - u| \leq h\} f(u)\, du, \end{aligned}$$

where the leading two terms are of higher order. Approximating the indicator function by the kernel function $K_h(x - X_i)$ gives the objective function

$$\sum_{i=1}^{n} K_h(x - X_i) \log f(X_i) - n \int_{\mathcal{X}} K_h(x - u) f(u)\, du,$$

which is the objective function studied by Loader (1996) and Hjort and Jones (1996).[32]

[32]The local likelihood estimator is available as a supplement to the Splus statistical software package. In section 6, we present some Monte Carlo results on the performance of these estimators.

We have grouped density estimation methods into moment-based and likelihood-based methods. Recent developments in empirical likelihood literature suggest a link between the method of moment estimators and likelihood estimators, which still needs to be clarified in this context.

### 4.2 How do we estimate conditional mean functions?

As with density estimation, there are both local and global approaches to estimating the conditional mean function. Because the conditional mean function does not characterize the conditional distribution, most of the methods analyzed extensively in the literature are based on the method-of-moments approach rather than the likelihood approach. Let $\mathcal{M}$ denote a class of functions in which the conditional mean function $m(x) = E(Y|X = x)$ lies. We can characterize the conditional mean function in two ways: as the solution to

$$\inf_{g(\cdot)\in\mathcal{M}} E\{[Y - g(X)]^2\}$$

or as the solution to

$$\inf_{g(\cdot)\in\mathcal{M}} E\{(Y - g(X))^2 | X = x\}.$$

The global method is based on the first characterization and the local method on the second. Analogous to the ML-based density estimation, both global and local approaches to estimating conditional mean functions require that the space $\mathcal{M}$ be restricted to avoid over-fitting.

Below we discuss nonparametric estimators of the conditional mean function. Estimators of the conditional quantile function can be constructed by replacing the quadratic loss function with that of Koenker and Bassett (1978). Also, see Tsybakov (1982), Härdle and Gasser (1984), and Chaudhuri (1991a, b).

**The global approach** As described earlier, the global approach to nonparametric estimation constructs a sequence of parametric models $M_n$ such that approximation error of $m(\cdot)$ by an element of $M_n$ eventually goes down to zero as $n \to \infty$. A well known sequence is a sequence of polynomial functions, a sequence of spline functions,[33] or a sequence of wavelets as discussed above. All sequences specify for each $n$ some set of functions $\{\phi_{nj}(x)\}_{j=1}^{J_n}$, and use them to define the sequence of models by

$$M_n = \{f;\ f(x) = \theta_1\phi_{n1}(x) + \cdots + \theta_{J_n}\phi_{nJ_n}(x) \text{ for some } \theta_1, ..., \theta_{J_n} \in \mathbb{R}\}.$$

Then, for each $n$ the global estimator can be defined as $\hat{m}(x, J_n) = \hat{\theta}_1\phi_{n1}(x) + \cdots + \hat{\theta}_{J_n}\phi_{nJ_n}(x)$, where $\hat{\theta}_1$,...,$\hat{\theta}_{J_n}$ are obtained by the least squares minimization problem of the following objective function:

$$\sum_{i=1}^{n} \left[y_i - \left(\theta_1\phi_{n1}(x_i) + \cdots + \theta_{J_n}\phi_{nJ_n}(x_i)\right)\right]^2.$$

As discussed before, for more than 1 regressor cases, appropriate Tensor product of a one dimensional bases are used to construct the base functions.

[33] See Schoenberg (1964) and also Eubank (1999) and Green and Silverman (1994).

Different global methods can be viewed as different combinations of decisions about how the class $\mathcal{M}$ is restricted and how the data are used in choosing the class $\mathcal{M}$. Clearly the properties of this estimator crucially depends on how we choose the base functions $\left\{\phi_{nj}(x)\right\}_{j=1}^{J_n}$ and $J_n$. Typically the order in which different base functions are brought in is given and the literature discusses how to choose $J_n$ using a model selection criterion. For example when the polynomial series are used base functions are ordered in terms of the degree of polynomials. In the wavelet literature, there is an attempt to endogenize the choice of the bases themselves so as to estimate the degree of smoothness.

Global approaches can be a convenient way of imposing global properties of underlying functions such as monotonicity, concavity, and additive separability. It is also easy to restrict a class of functions so that any function in the class goes through a certain point.

For global estimation methods, there has been less progress in analyzing the form of the first order bias in comparison to local methods. Although the rate of convergence is known, the exact expression for the highest order term is known only for limited cases. See Newey (1997) for the convergence rate results and see Zhou, Shen and Wolfe (1998) and Huang (2003) for some results about the first order bias computations.

**The local approach** Let $f(y,x)$ and $f(x)$ denote the joint density of $(Y,X)$ and the marginal density of $X$, respectively. Using the Dirac-delta function, $\delta_x(s)$, as used previously in setting up the moment condition for the density estimation (Section 4.1), we can write

$$\begin{aligned}\int\int \left[y-g(s)\right]^2 f(y,s)\,\delta_x(s)\,dsdy &= \int \left[y-g(x)\right]^2 f(y,x)\,dy \\ &= E\{(Y-g(X))^2\,|X=x\}f(x).\end{aligned}$$

As the last term is proportional to $E\{(Y-g(X))^2|X=x\}$, the solution to the infimum problem is the same if $f(x)>0$. Following the same logic as for the density estimation case, one can construct a sample analog objective function using some approximation to the Dirac-delta function.

If we do not restrict the class of functions ($\mathcal{M}$) over which infimum is taken, then the optimization problem does not have a well defined solution. Different local estimation methods can be viewed as different combinations of decisions about (1) how to approximate the Dirac-delta function (2) how to restrict the class $\mathcal{M}$ and (3) how to use the data in choosing the approximation and the class $\mathcal{M}$.

For example, if we approximate the Dirac-delta function by

$$\frac{1}{h}K\left(\frac{x-s}{h}\right)$$

as we did in the density estimation case, and restrict $\mathcal{M}$ to the class of constant functions, the left-hand side of the above expression has the sample analog:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\beta)^2\,K_h(x_i-x).$$

Minimizing this with respect to $\beta$ we get the kernel regression estimator:[34]

$$\hat{m}_K(x) = \frac{\sum_{i=1}^n y_i K_h(x_i - x)}{\sum_{i=1}^n K_h(x_i - x)},$$

whenever the denominator is not zero. Writing

$$w_{ni}^K(x) = \frac{K_h(x_i - x)}{\sum_{i=1}^n K_h(x_i - x)}$$

we see that $\hat{m}_K(x) = \sum_{i=1}^n y_i w_{ni}^K(x)$ and $\sum_{i=1}^n w_{ni}^K(x) = 1$.

If function $K(s)$ takes the form $1(|s| \leq 1)$ where $|s|$ denotes a norm of $s$ and the smoothing parameter $h$ is chosen to be the distance between $x$ and the $k$th closest observation in $\{x_i\}_{i=1}^n$, then the estimator is the $k$th-nearest neighbor estimator.

For the same Dirac-delta approximation, when $\mathcal{M}$ is replaced by a class of a finite dimensional polynomial function, we get the local polynomial regression estimator of E(Y|X=x) at $x = x_0$.[35] It is defined as the solution corresponding to $\beta_0$ of the following minimization problem:

$$\min_{\beta^{(0)},\beta^{(1)},\ldots,\beta^{(p)}} \frac{1}{n}\sum_{i=1}^n \left[ y_i - \beta^{(0)} - \sum_{\nu=1}^p \sum_{j_1+\cdots+j_d=\nu} \frac{1}{j_1!\cdots j_d!}\beta_{\nu,j_1,\ldots,j_d}(x_{i1}-x_{01})^{j_1}\cdots(x_{id}-x_{0d})^{j_d} \right]^2 K_h(x_i - x_0),$$

where for $\nu = 1, \ldots, p$ the length of vector $\beta^{(\nu)}$ is $(\nu+d-1)!/((d-1)!\nu!)$ and its elements are denoted by $\beta_{\nu,j_1,\ldots,j_d}$ where $j_1+\cdots+j_d = \nu$ and $j_1, \ldots, j_d$ are non-negative integers. For concreteness and for later purpose we order $j = (j_1, \ldots, j_d)$ lexicographically putting highest order to the first element, the next to the second element, etc.

To gain an understanding of the objective function, observe that

$$Y = m(X) + \epsilon = m(x_0) + [m(X) - m(x_0)] + \epsilon.$$

Consider the one-dimensional case and assume that $K(\cdot)$ is a symmetric, unimodal density function supported on the interval $[-1, 1]$. In that case, observations whose $X = x_i$ are close to $x_0$ receive more weight than others and if an observation's $X = x_i$ is more than $h$ apart from $x_0$, it receives 0 weight. If function $m(\cdot)$ is continuous at $X = x_0$, then when the approximation error $[m(X) - m(x_0)]$ is not very big so long as we restrict attention to observations whose $x_i$ is close to $x_0$. Thus ignoring the approximation error, minimizing the objective function for the kernel regression estimator is justified.

To motivate the objective function of the higher order polynomial estimator, consider a one dimensional case and let $m^{(\nu)}(x)$ denote the $\nu$th order derivative of function $m(\cdot)$. Observe that

$$\begin{aligned} Y &= m(X) + \epsilon \\ &= m(x_0) + m^{(1)}(x_0)(X - x_0) + \cdots + m^{(p)}(x_0)(X - x_0)^p/p! \\ &+ \{m(X) - [m(x_0) + m^{(1)}(x_0)(X - x_0) + \cdots + m^{(p)}(x_0)(X - x_0)^p/p!]\} + \epsilon, \end{aligned}$$

[34] See Nadaraya (1964) and Watson (1964) and Härdle (1990).

[35] See Stone (1977) and Fan and Gijbels (1996) p.105–106.

where $\{m(X) - [m(x_0) + m^{(1)}(x_0)(X - x_0) + \cdots + m^{(p)}(x_0)(X - x_0)^p/p!]\}$ constitutes the approximation error. The objective function is the weighted least squares objective function ignoring the approximation error where the observations whose $x_i$ are closer to $x_0$ receive higher weights. Clearly, the solution corresponding to the constant term is the estimator of the conditional mean function evaluated at $x_0$ and the solution corresponding to the coefficient of $(x_i - x_0)^\nu/\nu!$ is the estimator of the $\nu$th order derivative of the conditional mean function evaluated at $x_0$. For higher dimensional problems, we interpret $m^{(\nu)}(x_0)$, as a vector of partial derivatives of order $\nu$ and $(X - x_0)^{(\nu)}/\nu!$ as a vector of elements $(X_1 - x_{01})^{j_1} \cdots (X_d - x_{0d})^{j_d}/(j_1! \cdots j_d!)$, where $j_1 + \cdots + j_d = \nu$. For concreteness, we assume both are ordered in the lexicographical way as above.

Fan (1992) clarifies the theoretical reasons why we may prefer to use the local polynomial regression estimator with $p \geq 1$ instead of the kernel regression estimator ($p = 0$). The advantage is the ability of the estimator to control the bias in *finite sample*. As we have seen above, in finite sample the kernel regression estimator ignores $[m(X) - m(x_0)]$, which is of order $h$ in the neighborhood of $x_0$ when the underlying function is twice differentiable with bounded second derivative. If the local linear estimator is used, then under the same condition, the approximation error ignored is of order $h^2$ in the neighborhood of $x_0$. If $p$th order polynomial is used and the underlying function is at least $r$-times differentiable with bounded $r$th derivative where $r \geq p + 1$, the approximation error is of order $h^{p+1}$ in *finite sample*. This leads to practical and theoretical advantages.

For the kernel regression estimator evaluated at the interior point of the support of regressors, when the underlying function is twice differentiable and the second derivative is bounded, the first order asymptotic analysis shows that the asymptotic bias is of order $h^2$, which is the same order with the local linear estimator. However, note that this is an asymptotic result and applicable to interior points. For the local linear estimator, the bias is of order $h^2$ in finite sample whenever the estimator is well defined. For the local polynomial regression estimator of order $p$, so long as the estimator is defined and the underlying function is sufficiently smooth, the bias is of order $h^{p+1}$ in finite sample. This is the practical advantage.

When a nonparametric estimator is used to construct a semiparametric estimator and the asymptotic properties of the resulting semiparametric estimator is examined, as we shall see later, typically the uniform convergence needs to be established with a certain convergence rate. Since the same convergence rate can be achieved without the boundary consideration, the theoretical development simplifies. This is the theoretical advantage.

We clarify above points in some detail as the results will be useful to understand the asymptotic results discussed below and the bandwidth selection methods discussed later. Let $j = (j_1, \ldots, j_d)$ and denote $|j| = j_1 + \cdots + j_d$. Also let $\beta = (\beta^{(0)}, \beta^{(1)\prime}, \ldots, \beta^{(p)\prime})'$, $N_u = (u + d - 1)!/((d - 1)!u!)$, $N = \sum_{u=0}^{p} N_u$, and $X^{(u)}$ be an $n \times N_u$ matrix with the $i$th row being $(x_i - x_0)^j/j!$ for $|j| = u$, interpreted as specified above, $\iota_n$ be the vector of $n$ ones, $X = (\iota_n X^{(1)} \cdots X^{(p)})$ (an $n \times N$ matrix), $y = (y_1, \ldots, y_n)'$ and $W$ be an $n \times n$ diagonal matrix with $i$th diagonal element being $K_h(x_i - x_0)$.

With these notations, the local polynomial objective function can be written as

$$(y - X\beta)'W(y - X\beta)$$

so that the local polynomial estimator is, when it exists, $\hat{\beta} = (X'WX)^{-1}X'Wy$. The local polynomial estimator of the conditional mean function is the first element of $\hat{\beta}$ so that it can be written as $\sum_{i=1}^{n} w_{ni}^{L}(x_0)y_i$ where

$$(w_{n1}^{L}(x_0), \ldots, w_{nn}^{L}(x_0)) = e_N'(X'WX)^{-1}X'W,$$

where $e_N$ is a vector of length $N$ with first element being one and the rest of the elements are zero. Observe that

$$(w_{n1}^{L}(x_0), \ldots, w_{nn}^{L}(x_0))X = e_N'(X'WX)^{-1}X'WX = e_N'I_N.$$

Reading off the row, we observe that $\sum_{i=1}^{n} w_{ni}^{L}(x_0) = 1$, $\sum_{i=1}^{n} w_{ni}^{L}(x_0)(x_i - x_0) = 0$, and generally $\sum_{i=1}^{n} w_{ni}^{L}(x_0)(x_i - x_0)^{(j)}/j! = 0$ for any $j$ with $1 \leq |j| \leq p$. As we shall see below, these orthogonality properties of the weight function are the source that enables the estimator to control bias in finite sample.

The weights for the kernel regression estimator satisfies $\sum_{i=1}^{n} w_{ni}^{K}(x_0) = 1$, but satisfies

$$\sum_{i=1}^{n} w_{ni}^{K}(x_0)(x_i - x_0) \to 0$$

only asymptotically when $x_0$ is at the interior point of the support of $x_i$ and this does not hold asymptotically if $x_0$ is on the boundary of the support of $x_i$.

One might think that the limitation of the kernel regression estimator at the boundary points is not so important practically, because there are many more interior points than boundary points. However, two points need to be taken into account. First, the comparable performance of the kernel regression estimator in interior points is obtained asymptotically, not in the finite sample as for the local polynomial estimator. Second, in finite sample, it is entirely plausible that the data are unevenly distributed, so that there are many more data points lying on one side of the point of evaluation ($x_0$) than the other. This is even more likely to occur in higher dimensions. In these cases, the asymptotic properties of the kernel regression estimator may not capture well the finite sample behavior. In some sense, in finite sample, there are likely many points at which the boundary behavior of the estimator may better represent its performance.

To see these points more clearly, define $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$, $\beta_0^{(\nu)} = m^{(\nu)}(x_0)$ for $\nu = 0, \ldots, p$ and $\beta_0$ to be the vector of length $N$ constructed by stacking these sub-vectors. We can write

$$y = X\beta_0 + r + \epsilon$$

where $r = (r_1, \ldots, r_n)' = m - X\beta_0$ with $m = (m(x_1), \ldots, m(x_n))'$. Thus

$$\hat{\beta} = \beta_0 + (X'WX)^{-1}X'Wr + (X'WX)^{-1}X'W\epsilon.$$

The second term on the right-hand side is the bias term and the third, the variance term. We examine the bias and the variance terms in turn.

**Bias** Let $H$ be the diagonal matrix with $N_u$ diagonal elements of $1/h^u$ for $u = 0, \ldots, p$, in this order. Then

$$\begin{aligned}\hat{\beta}^{(0)} &= \beta_0^{(0)} + e_N' H(HX'WXH)^{-1}HX'Wr + e_N' H(HX'WXH)^{-1}HX'W\epsilon \\ &= \beta_0^{(0)} + e_N'(HX'WXH)^{-1}HX'Wr + e_N'(HX'WXH)^{-1}HX'W\epsilon.\end{aligned}$$

One can show that $HX'WXH/(nh^d)$ converges in probability to an invertible matrix, under general conditions specified later. To see this, note that the typical element of the matrix is, for vectors of non-negative integers $j$ and $j'$,

$$\frac{1}{nh^d}\sum_{i=1}^{n}((x_i - x_0)/h)^{(j)}((x_i - x_0)/h)^{(j')}K((x_i - x_0)/h)/(j!j'!).$$

Applying the Taylor series expansion we obtain

$$r_i = m^{(p+1)}(\bar{x}_i)(x_i - x_0)^{(p+1)}/(p+1)!,$$

where $m^{(p+1)}(\bar{x}_i)$ is a row vector of length $N_{p+1}$, consisting of $m^{(j)}(\bar{x}_i)$ with $|j| = p+1$ and $\bar{x}_i$ lies on a line connecting $x_i$ and $x_0$. Using this result, a typical element of $HX'Wr/(nh^d)$ can be written as, using the same $j$ and $j'$ as above,

$$\frac{1}{nh^d}\sum_{i=1}^{n}((x_i - x_0)/h)^{(j)}(x_i - x_0)^{(j')}m^{(j')}(\bar{x}_i)K((x_i - x_0)/h)/(j!j'!),$$

where here, $|j'| = p+1$. Since $\|(x_i - x_0)^{(j')}\| = O(h^{p+1})$ when the kernel function used has a bounded support, if the $p+1$st order derivative of $m(x)$ at $x = x_0$ is bounded, then the bias term is of order $h^{p+1}$.

Note that when the $p+1$st derivative is Lipschitz continuous at $x = x_0$, the leading term of the bias can be expressed as

$$e_N'(HX'WXH)^{-1}HX'WX^{(p+1)}m^{(p+1)}(x_0).$$

**Variance** Conditional variance of $e_N'(X'WX)^{-1}X'W\epsilon$ is

$$e_N'(X'WX)^{-1}X'W\Sigma WX(X'WX)^{-1}e_N,$$

where $\Sigma$ is an $n \times n$ diagonal matrix with the $i$th diagonal element $\sigma^2(x_i) = E(\epsilon_i^2|x_i)$. This can be rewritten as

$$\begin{aligned}& e_N' H(HX'WXH)^{-1}HX'W\Sigma WXH(HX'WXH)^{-1}He_N \\ = \;& e_N'(HX'WXH)^{-1}HX'W\Sigma WXH(HX'WXH)^{-1}e_N.\end{aligned}$$

1 Combining the earlier calculation about $HX'WXH/(nh^d)$ with the observation that the typical element of $HX'W\Sigma WXH/(nh^d)$ can be written as

$$\frac{1}{nh^d}\sum_{i=1}^{n}((x_i - x_0)/h)^{(j)}((x_i - x_0)/h)^{(j')}\sigma^2(x_i)K^2((x_i - x_0)/h)/(j!j'!),$$

we see that the variance is of order $1/(nh^d)$. Note that when the conditional variance function is Lipschitz continuous at $x_0$, the highest order term of the conditional variance can be expressed as

$$e_N'(HX'WXH)^{-1}HX'W^2XH(HX'WXH)^{-1}e_N\sigma^2(x_0).$$

These finite sample expressions of the bias term and the conditional variance term will later be used to approximate the mean squared error, which can be used to optimally choose the bandwidth $h$.

The asymptotic properties of the local polynomial estimator has been developed by many authors, but the following results due to Masry (1996a, 1996b) seem to be the most comprehensive. We assume stationarity of $\{(X_t, Y_t)\}$, and define the local polynomial regression estimator of $E(Y_{t+s}|X_t = x) = m(x)$ and its derivatives at $x = x_0$ where $x_0 \in \mathbb{R}^d$.

Let $f(x)$ denote the Lebesgue density of $X_t$, $f(x, x', \ell)$ denote the joint Lebesgue density of $X_t$ and $X_{t+\ell}$, $j = (j_1, \ldots, j_d)$,

$$D^j m(x) = \frac{\partial^{|j|} m(x)}{\partial^{j_1} x_1 \cdots \partial^{j_d} x_d},$$

its local polynomial estimator of order $p$ by $\hat{\beta}_{|j|,j}(x)$, and define for $(0, \ldots, 0) \in \mathbb{R}^d$, $\hat{\beta}_{0,(0,\ldots,0)}(x) = \hat{\beta}^{(0)}$. Masry (1996b) establishes the conditions under which local polynomial estimator converges uniformly over a compact set.

**Theorem 1.** *Let $D$ be a compact subset of $\mathbb{R}^d$. If*

1. *the kernel function $K(\cdot)$ is bounded with compact support (There exists $A > 0$ such that $K(u) = 0$ for $\|u\| > A$.) and there exists $C > 0$ such that for any $(j_1, \ldots, j_d)$ such that $0 \le j_1 + \cdots j_d \le 2p + 1$*

$$|u_1^{j_1} \cdots u_d^{j_d} K(u) - v_1^{j_1} \cdots v_d^{j_d} K(v)| \le C\|u - v\|,$$

2. *the stationary process $\{(X_t, Y_t)\}$ is strongly mixing with the mixing coefficient $\alpha(k)$ satisfying*

$$\sum_{j=1}^{\infty} j^a \alpha(j)^{1-2/\nu} < \infty$$

   *for some $\nu > 2$ and $a > 1 - 2/\nu$,*

3. *there exists $C > 0$ such that $f(x) < C$, $f(x)$ is uniformly continuous on $\mathbb{R}^d$, and $\inf_{x \in D} f(x) > 0$,*

4. *there exists $C > 0$ such that $f(u, v, \ell) < C$,*

5. *the conditional density $f_{X_0|Y_s}(x|y)$ of $X_0$ given $Y_s$ exists and is uniformly bounded,*

6. *the conditional density $f_{(X_0,X_\ell)|(Y_s,Y_{s+\ell})}(x, x'|y, y')$ of $(X_0, X_\ell)$ given $(Y_s, Y_{s+\ell})$ exists and is uniformly bounded for all $\ell \ge 1$,*

7. *the $p+1$st order of derivative of $m(x)$ is uniformly bounded and the $p+1$st order derivative is Lipschitz continuous, and*

8. *$E(|Y|^\sigma) < \infty$ for some $\sigma > \nu$,*

*then,*

$$\sup_{x \in D} |\hat{\beta}_{|j|,j}(x) - D^j m(x)| = O((\ln n/(nh^{d+2|j|}))^{1/2}) + O(h^{p-|j|+1}).$$

Point-wise variance goes down with rate $1/(nh^d)$ as discussed above when $|j| = 0$. The $\ln n$ factor is the penalty we need to pay for uniform convergence.

Masry (1996a) establishes the asymptotic normality of the local polynomial estimator in an interior point of the support of $X_t$.[36] Let $M$ and $\Gamma$ be $N \times N$ matrices with $N_u \times N_v$, sub-matrices $M_{u,v}$ and $\Gamma_{u,v}$ for $u, v = 0, \ldots, p$, respectively, where the typical elements of $M_{u,v}$ is $\int x_1^{j_1+j_1'} \cdots x_d^{j_d+j_d'} K(x)dx/(j!j'!)$ with $|j| = u$ and $|j'| = v$ and the typical element of $\Gamma_{u,v}$ is $\int x_1^{j_1+j_1'} \cdots x_d^{j_d+j_d'} K^2(x)dx/(j!j'!)$with $|j| = u$ and $|j'| = v$. Analogously define $M_{u,p+1}$ for $u = 0, \ldots p$ and define the $N \times N_{p+1}$ matrix $B$ as

$$\begin{pmatrix} M_{0,p+1} \\ M_{1,p+1} \\ \vdots \\ M_{p,p+1} \end{pmatrix}$$

and recall that we write $m^{(p+1)}(x)$ to denote a vector of $D^j m(x)$ with $|j| = p+1$ in the lexicographic order discussed above.[37]

Note that matrices $M$, $\Gamma$, and $B$ are the probability limits of $HX'WXH/(nh^d)$, $HX'W^2XH/(nh_n^d)$, and $HX'WX^{(p+1)}/(nh^d)$, respectively, when $x_0$ is an interior point of the support of $X_t$.

**Theorem 2.** *Suppose $x_0$ is an interior point of the support of $X_t$. Let $h = O(n^{-1/(d+2p+2)})$ as $n \to \infty$. If the conditional distribution and the conditional variance of $Y_s$ given $X_0 = x$ are continuous at $x = x_0$, $f(x)$ is continuous at $x_0$, $f(x_0) > 0$ and if*

1. *the kernel function $K(\cdot)$ is bounded with compact support,*

2. *the stationary process $\{(X_t, Y_t)\}$ is strongly mixing with the mixing coefficient $\alpha(k)$ satisfying*

$$\sum_{j=1}^{\infty} j^a \alpha(j)^{1-2/\nu} < \infty$$

*for some $\nu > 2$ and $a > 1-2/\nu$, and there exists $\nu_n = o((nh^d)^{1/2})$ such that $(n/h^d)^{1/2}\alpha(\nu_n) \to 0$ as $n \to \infty$,*

[36] The following results imposes comparable conditions as those above, although Masry (1996a) establishes results under somewhat weaker conditions on the kernel functions and results include cases under $\rho$-mixing as well.

[37] Our definition of $M$ is different from that in the Masry's paper by $j!j'!$ for each element of $M$ and thus the asymptotic bias and variance expressions differ as well reflecting only the difference in the notations.

3. *there exists* $C > 0$ *such that* $f(x) < C$,

4. *there exists* $C > 0$ *such that* $f(u, v, \ell) < C$,

5. *the conditional density* $f_{(X_0,X_\ell)|(Y_s,Y_{s+\ell})}(x, x'|y, y')$ *of* $(X_0, X_\ell)$ *given* $(Y_s, Y_{s+\ell})$ *exists and is uniformly bounded for all* $\ell \geq 1$,

6. *the* $p+1$*st order of derivative of* $m(x)$ *is uniformly bounded and the* $p+1$*st order derivative is Lipschitz continuous, and*

7. $E(|Y|^\nu) < \infty$ *for* $\nu$ *defined above,*

*then,*

$$(nh^{d+2|j|})^{1/2}\left([\hat{\beta}_{|j|,j}(x_0) - D^j m(x_0)] - (M^{-1}Bm^{(p+1)}(x_0))_i h^{p+1-|j|}\right)$$

*converges in distribution to the zero mean random variable with variance*

$$\frac{\sigma^2(x_0)}{f(x_0)}(M^{-1}\Gamma M^{-1})_{i,i}$$

*where* $i$ *denotes the order in which* $j$ *appear in matrix* $M$.

The convergence rate coincides with the optimal rate computed by Stone (1982). The theorem specifies the rate at which $h$ should converge to 0, but does not specify how to choose $h$. Section six discusses how to choose the smoothing parameter.

Note that the first order bias term

$$(M^{-1}Bm^{(p+1)}(x_0))_i h^{p+1-|j|}$$

depends on the $p+1$st order derivatives but does *not*, in general, depend on the distribution of the conditioning variable, other than the fact that $f(x_0) > 0$ has been used in arriving at the formula. When the kernel function used is symmetric and $p - |j|$ is even, then the bias term is of order $h^{p+2-|j|}$ and involves the derivative of regressor density.[38] This corresponds to the case of the kernel regression estimator.[39]

The order of the variance depends on the dimension of the function being estimated and the order of the derivative of the target function, but does not depend on the degree of the polynomial used in estimation. However, the constant term in the variance expression does depend on the degree of the polynomial used. It has been observed that the constant term does not change when $p$ moves from a lower even number to the next odd number, for example from 0 to 1. It does go up when moving up from an odd number to the next even number, for example from 1 to 2.[40] Thus, moving up by one from an even number to an odd number reduces the bias, but does not add to the variance. So when there is a choice, we should choose $p$ to be an odd number. In particular, it is better to use a local linear estimator to estimate the conditional mean function

[38] See Fan and Gijbels (1996), Theorem 3.1 for a discussion of a univariate case.

[39] See Härdle and Linton (1994).

[40] See Ruppert and Wand (1994) and Fan and Gijbels (1996) section 3.3.

rather than a kernel regression estimator. Note that this is a result at interior points and also when the underlying function is at least $p+1$-times continuously differentiable.

Another point to note about the form of the first order variance is that it is the same regardless of whether the errors are allowed to be correlated or not. This is a standard but an unpleasant result in nonparametric asymptotic analysis as pointed out by Robinson (1983) for the case of kernel density estimation. It is unpleasant, because, for any finite number of observations, the observations that fall in the fixed neighborhood of $x_0$ would be correlated especially in high frequency data analysis. See Conley, Hansen, Liu (1997) for a bootstrap approach to assess the variability.[41]

Here, we have discussed local polynomial estimation of the conditional mean function. For a discussion of locally linear estimation of the conditional quantile function, see Chaudhuri (1991a, b) and Yu and Jones (1998).

# 5 Semiparametric Estimation

We review some semiparametric estimation methods used in econometrics. As discussed in section two, the curse-of-dimensionality problem associated with nonparametric density and conditional mean function estimators makes the methods impractical in applications with many regressors and modest size samples. Semiparametric modeling approaches offer a middle ground between fully nonparametric and fully parametric approaches. They achieve faster rates of convergence for conditional mean functions or other parameters of interest by employing one of the three approaches discussed earlier: by imposing some parametric restrictions, by changing the target parameters, or by imposing quantile restrictions in the case of limited dependent variable models. The nonparametric density and conditional mean function estimators described in the last section form the building blocks of a variety of semiparametric estimators.

In section two, we considered one semiparametric model–the partially linear model–and described its application to the problems of estimating consumer demand functions and to controlling for sample selection. Here we consider that model in greater detail as well as other classes of semiparametric models for conditional mean function estimation, including additive separable models, index models, and average derivative models with and without index restrictions. We also review censored LAD estimator of Powell (1984) and the Maximum Score estimator of Manski (1975, 1985) for the limited dependent variable models as examples of exploiting quantile restrictions. These methods embody distinct ideas that are applicable in other contexts. A detailed discussion of techniques for deriving the distribution theory is left for section seven.

## 5.1 Conditional mean function estimation with an additive structure

Suppose the relationship of interest is $E(Y|X=x)=g(x)$, where $X$ is a random vector of length $d$ and $g$ is an unknown function from $\mathbb{R}^d$ into $\mathbb{R}$. As described earlier, we face the curse of di-

[41] Another approach may be to compute the finite sample variance formula and estimate it analogously to the Newey-West approach.

mensionality if fully nonparametric estimator of $g(x)$ were to be used. Another problem is that nonparametrically estimated $g$ functions become difficult to interpret when the estimated surface can no longer be visualized and the effect of any regressor on the dependent variable depends on the values of all the other regressors.

We consider three classes of semiparametric estimators for $g(x)$ that impose different kinds of modeling restrictions designed to overcome the curse-of-dimensionality problem and to make estimates easier to interpret. The first class, *additively separable models*, restricts $g(x)$ to lie in the space of functions that can be written as an additively separable function of the regressors. The second class, *single index models*, assumes that $X$ affects $Y$ only through an index $X'\beta$. That is, $g(x) = g(x'\beta)$. *Multiple index models* allow the conditional mean of $Y$ to depend on multiple indices. The third class, *partially linear models*, assumes that the function $g(x)$ can be decomposed into a linear component and a nonparametric component, thereby extending the traditional linear modeling framework to include a nonparametric term. Partially linear restrictions are often imposed in connection with index model restrictions, giving rise to partially linear, single or multiple index models.

### 5.1.1 Additively separable models

An additively separable model restricts $g(x)$ to be additively separable in the components of the vector $X$:

$$E(Y|X) = \alpha + g_1(X_1) + g_2(X_2) + g_3(X_3) + \cdots + g_d(X_d),$$

where the $g_i(x_i), i = 1..d$, are assumed to be unknown and are nonparametrically estimated. A key advantage of imposing additive separability is that the nonparametric estimators of the $g_i(x_i)$ functions as well as of the conditional mean function $E(Y|X = x)$ can be made to converge at the univariate nonparametric rate. Another advantage is interpretive: the model allows for graphical depiction of the effect of $x_j$ on $y$ holding other regressors constant. The separability assumption is also not as restrictive as it may seem, because some regressors could be interactions of other regressors (e.g. $x_3 = x_1x_2$). However, for $g_i(x_i)$ to be nonparametrically identified, it is necessary to rule out general forms of collinearity between the regressors. That is, we could not allow $x_1 = \psi(x_k)$ for some $\psi$ function, for example, and still separately identify $g_1(x), \ldots, g_d(x)$.[42]

#### Estimation Methods

**Back-fitting algorithms** As described in Hastie and Tibshirani (1990), additively separable models can be solved through an algorithm called *back-fitting*.

The algorithm involves three steps:

**(i)** Choose initial starting values for $\alpha$ and for $g_j$. A good starting value might set $\alpha^0 = average(Y)$ and $g_j^0$ equal to the values predicted by a linear in $x$ least squares regression of $Y$ on a constant and all the regressors.

[42] See the discussion of concurvity in Hastie and Tibshirani (1990).

**(ii)** For each $j = 1..d$, define $g_j = \hat{E}(y - \alpha - \sum_{k \neq j} g_k^0(x_k) \mid x_j)$, where $g_k^0$ is the most recent estimate of $g_k(x_k)$ (the starting value at the first iteration). The conditional expectation is estimated by a smoothing method, such as kernel or local linear regression, or series expansion or spline regression. At this stage, if it is desired that a functional form restriction be imposed on the shape of one or more of the $g_j$ functions, then the restriction can be imposed by setting, for example, $\hat{E}(y - \alpha - \sum_{k \neq j} g_k^0(x_k) \mid x_j) = x_j \beta_j$.

**(iii)** Repeat step (ii) until convergence is reached (when the estimated $g_j(x_j)$ functions no longer change).[43]

Back-fitting can require many iterations to reach convergence, but it is relatively easy to implement and is available in the software package Splus. Disadvantages of the method are that consistency has not been shown when nonparametric smoothing methods are used in step (ii) and there is as of yet no general distribution theory available that can be used to evaluate the variation of the estimators.

**An estimator based on integration** An alternative approach to estimating the additively separable model, which is studied by Newey (1994), Härdle and Linton (1996), Linton, Chen, Wang and Härdle (1997) and others. Although it is more difficult to implement than the back-fitting procedure, because it requires a pilot estimator of the nonparametric model $g(x)$, the integration approach has the advantage of having a distribution theory available.

For notational simplicity, consider the additively separable model with two regressors $Y = \alpha + g_1(X_1) + g_2(X_2) + \varepsilon$. Define the integrated parameter

$$\tilde{g}_1(x_1) = \int g(x_1, x_2) dF_{x_2}.$$

Note that this is generally *not* equal to $E(Y|X_1 = x_1)$ which would be

$$E(Y|X_1 = x_1) = \int g(x_1, x_2) dF_{x_2|X_1 = x_1}.$$

If $X_1$ and $X_2$ are independent, then the two parameters coincide. The integration estimator is given by

$$\hat{g}_1(x_1) = n^{-1} \sum_{i=1}^{n} \hat{g}(x_1, x_{2i}).$$

If the model is additive, then $\hat{g}_1(x_1)$ estimates $g_1(x_1)$ up to an additive constant. Reversing the roles of $x_1$and $x_2$ obtains an estimator for $g_2(x_2)$, again up to scale.

In general, we do not really believe that the underlying function $g(x_1, x_2)$ is additively separable but that we use the model as a convenient way to summarize data. From this perspective, the integration estimator proposes to examine the effect of one variable $X_1$ on the dependent variable

[43] Also see Hastie and Tibshirani (1990) for discussion of a modified back-fitting algorithm that, in some circumstances, converges in fewer iterations.

after integrating out the rest of the variables $X_2, \ldots X_d$ using the marginal distribution of $X_2, \ldots X_d$, which would be exactly the correct procedure if the underlying function $g$ is indeed additively separable between $X_1$ and $X_2, \ldots X_d$.

The back fitting algorithm seems to be an attempt to obtain the solution to the least squares problem within the class of additively separable functions. Although these two sets of functions should coincide, up to an additive constant terms, if underlying function $g$ is additively separable, if not, the two estimates in general would converge to different functions.

Newey (1994) shows that the estimator $\hat{g}_1(x_1)$ converges at a one-dimensional nonparametric rate because of the averaging. As we have seen, the convergence rate decreased because the rate at which we obtain data decreased if we needed to condition on a point in a higher dimension space. Since there is no need to condition on $X_2, \ldots X_d$ for examining $g_1(x_1)$, the convergence rate corresponds to that for one-dimensional cases.

As noted above, an advantage of estimating additive models through integration is that the distribution theory for the estimators has been developed.[44] A disadvantage of the integration estimator is that it requires that the higher dimensional estimate of the $g(x)$ be calculated prior to averaging, and existing distribution theory for the estimator requires that negative kernel functions be used for bias reduction.

**Generalized additive models** The additive modeling framework has been generalized to allow for known or unknown transformations of the dependent variable, $Y$. That is, estimators are available for models of the form

$$\theta(Y) = \alpha + g_1(X_1) + g_2(X_2) + \cdots + g_d(X_d) + \varepsilon,$$

where the link function $\theta$ may be a known transformation (such as the Box-Cox transformation) or may be assumed to be unknown and nonparametrically estimated along with the $g_j$ functions. Hastie and Tibshirani (1990) describe how to modify back-fitting procedures to accommodate binary response data and survival data, when the link function is known. For the case of an unknown $\theta$ function, Breiman and Friedman (1985) propose an estimation procedure called ACE (Alternating Conditional Expectation).[45] Linton, Chen, Wang and Härdle (1997) describe an instrumental variables procedure for estimating the $\theta$ function, which is based on the identifying assumption that the model is only additively separable for the correct transformation so that misspecification in $\theta$ shows up as a correlation between the error terms and the instruments. We are not aware of empirical applications of these methods in economics, although generalized additive models (GAMs) and ACE seem potentially very useful ways for empirical researchers to gain some flexibility in modeling the conditional mean function while at the same time avoiding the curse-of-dimensionality.

[44] See, for example, Härdle and Linton (1996).

[45] ACE is also discussed in Hastie and Tibshirani (1990). The ACE algorithm is available in the software package Splus.

### 5.1.2 Single Index Model

The single index model restricts the function $g(x)$ under consideration to be

$$g(x) = \phi(x'\beta_0)$$

where $\phi$ is an unknown function. An estimator of the slope coefficients $\beta_0$ in the single index model that allows for discrete regressors and regressors which may be functionally related is studied by Ichimura (1993).

Consider the single index model in the conditional mean function:

$$Y_i = \phi\left(X_i'\beta_0\right) + \varepsilon_i$$
$$E\left(\varepsilon_i | X_i\right) = 0.$$

This model arises naturally in a variety of limited dependent variable models in which the observed dependent variable $Y_i$ is modeled as a transformation of $X_i'\beta_0$ and an unobserved variables which are independent of $X_i$. See Heckman and Robb (1985) and Stoker (1986). Also, this model can be viewed simply as a generalization of the regression function.

Observe that

$$\begin{aligned} m_W(b) &\equiv E\left\{\left[Y - E\left(Y|X'b\right)\right]^2 W(X)\right\} \\ &= E\left\{\varepsilon^2 W(X)\right\} + E\left\{\left[\phi\left(X'\beta_0\right) - E\left(Y|X'b\right)\right]^2 W(X)\right\} \end{aligned}$$

The computation makes clear that, for any function $W(x)$, the variation in $Y$ has two sources: the variation in $X'\beta_0$ and that in $\varepsilon$ and that if we choose $b$ to be proportional to $\beta_0$, then contribution to the variation due to the variation in $X'\beta_0$ becomes zero in function $m_W(b)$ as $E(Y|X'b) = \phi(X'\beta_0)$ in that case. This observation lead to defining an estimator as

$$\min_b \frac{1}{n}\sum_{i=1}^{n}\left[y_i - E\left(y_i|x_i'b\right)\right]^2 W(x_i)$$

if we knew the conditional mean function $E(Y_i|X_i'b)$. As we do not know it, we need to replace it with its estimate. But since the conditional mean function cannot be estimated at points where the density of $X_i'b$ is low, we need to introduce trimming as other estimators we examined earlier.

The trimming function in this case has a further complication. Even if the density of $X$ is bounded away from zero, the density of $X'b$ is not, in general. This can be understood by considering two variables that has the uniform distribution on the unit square and considering the density corresponding to the sum.

A simple way around this problem is to define the trimming function as follows:

$$I_i = 1\{x_i \in \mathcal{X}\},$$

where $\mathcal{X}$ denotes a fixed interior points of the support of $X_i$ by at least certain distance. Note that over this set $\mathcal{X}$, by construction the density of $x$ is bounded away from zero and that the density of $X'b$ is also bounded away from zero.

Another point to note is that for any constant value $c \neq 0$, $E(Y|X'b = x'b) = E(Y|X'(cb) = x'(cb))$ so that we cannot identify the length of $\beta_0$. Thus we define the estimator to be the minimizer of the following objective function after replacing $E(Y_i|X_i'b)$ with a nonparametric estimator of it:

$$\min_{b \in \{b:b'b=1\}} \frac{1}{n} \sum_{i=1}^{n} \left[ y_i - \hat{E}\left(y_i|x_i'b\right) \right]^2 W(x_i) I_i.$$

In implementation, two forms of normalization are used; in some cases $\beta'\beta = 1$ is imposed and in other cases one of the coefficient is set to 1.[46] In either case, the Var-Cov matrix of the estimator is $V^- \Omega V^-$, where

$$\begin{aligned} V &= E\left\{ \left[\varphi'\left(x'\beta_0\right)\right]^2 \left[\tilde{x} - E\left(\tilde{x}|x'\beta_0\right)\right] \left[\tilde{x} - E\left(\tilde{x}|x'\beta_0\right)\right]' \right\}, \\ \Omega &= E\left\{ \sigma^2(x) \left[\varphi'\left(x'\beta_0\right)\right]^2 \left[\tilde{x} - E\left(\tilde{x}|x'\beta_0\right)\right] \left[\tilde{x} - E\left(\tilde{x}|x'\beta_0\right)\right]' \right\}, \\ \sigma^2(x) &= V(y|x) \end{aligned}$$

and all of the expectations are taken over a given set $\mathcal{X}$ over which the density of $x'\beta_0$ is assumed to be bounded away from 0. When $\beta'\beta = 1$, $\tilde{x} = x$ and when one of the coefficients is set to 1, $\tilde{x}$ is the original regressors except the regressor whose coefficient is set to 1. For the first normalization, note that $\Omega\beta_0 = 0$ and $V\beta_0 = 0$ hold so that $V$ and $\Omega$ are not invertible.

There are two sources of efficiency loss. One is that the variation in $\tilde{x} - E(\tilde{x}|x'\beta_0)$ is used rather than the variation in $\tilde{x}$. The other is that heteroskedasticity is not accounted for in the estimation. While the first problem arises as $\phi$ is unknown, and hence is genuine to the formulation of the problem, the second problem can be resolved by weighting if the model is truly single index. Oftentimes, however, we use the single index model as a convenient approximation to a more general function. Ichimura and Lee (2006) shows that if the single index model is used when the underlying model is not single index, the SLS estimator still is consistent to a vector which best approximates the conditional mean function within the single index model, and it is asymptotically normal but its asymptotic variance contains an additional term. They discuss how to estimate the asymptotic variance term including this additional term and hence how to make the estimator robust to misspecification. Here the discussion used the linear single index, but the same idea applies to the nonlinear index model and also to the case of multiple indices. See Ichimura and Lee (1991).

When the dependent variable is discrete, the more natural objective function is likelihood based. Klein and Spady (1993) examines the case of binary choice models and shows that the estimator is efficient among semiparametric estimators.

Blundell and Powell (2003) considers the single index model with an endogenous regressor and Ichimura and Lee (2006) considers the estimation of the conditional quantile function when the conditional quantile function is modeled as a single index function.

[46] We consider $W(x) = 1$ for simplicity below. See Ichimura (1993) for the weighted case. In general we need to modify the standard estimation of $E(Y|X'b)$ to achieve efficiency by weighting.

### 5.1.3 Partially Linear Regression Model

The partially linear regression model extends the linear regression model to include a nonparametric component and specifies:

$$Y = X'\beta_0 + \varphi(Z) + \varepsilon$$

where $X \in R^p$ and $Z \in R^q$ do not have common variables. If they do, then the common variables would be regarded as a part of $Z$ but not $X$ because the coefficients that correspond to the common variables would be not identifiable. If there is no cross terms of $z$ among $X$'s, then the model presumes additive separability of $\varphi(Z)$ and $X$, which may be too restrictive in some applications.

This framework is convenient for a model with many regressors, where fully nonparametric estimation is often impractical. It is also a good choice for a model that contains discrete regressors along with a few continuous ones. As discussed in section two, this model has been broadly applied in economics, mainly to the problem of estimating Engel curves and to the problem of controlling for sample selection bias. Estimators for the partially linear model are studied in Heckman (1980, 1990), Shiller (1984), Stock (1991), Wahba (1984), Engle, Grander, Rice and Weiss (1986), Chamberlain (1986b), Powell (1987), Newey (1988), Robinson (1988), Ichimura and Lee (1991), Andrews (1991), Cosslett (1991), Choi (1992), Ahn and Powell (1993), Honoré and Powell (1994), Yatchew (1997), Heckman, Ichimura, Smith, and Todd (1998a), Heckman, Ichimura, and Todd (1998b) and others.

As we saw, the nonparametric convergence rate would depend on the number of continuous regressors in $(X, Z)$. In the partially linear regression framework, the convergence rate of the estimator of $\varphi$ depends only on the number of continuous regressors among $z$ and that the $n^{1/2}$-consistent estimation of $\beta$ can be carried out regardless of the number of continuous regressors in $(X, Z)$ provided there is enough smoothness in underlying functions as shown by Robinson (1988).

To consider the estimator Robinson studied, observe that

$$E(Y|Z=z) = E\left(X'|Z=z\right)\beta_0 + \varphi(z)$$

so that

$$Y - E(Y|Z=z) = (X - E(X|Z=z))'\beta_0 + \varepsilon.$$

If we knew $E(Y|Z=z)$ and $E(X|Z=z)$ then one could estimate $\beta_0$ by the ordinary least squares method of $Y - E(Y|Z=z)$ on $X - E(X|Z=z)$. Since we do not know them, we can estimate them by some nonparametric method, call them $\hat{E}(Y|Z=z)$ and $\hat{E}(X|Z=z)$, and estimate $\beta_0$ by

$$\left(\sum_{i=1}^{N}\left[x_i - \hat{E}(x_i|z_i)\right]\left[x_i - \hat{E}(x_i|z_i)\right]'\right)^{-1}\sum_{i=1}^{N}\left[x_i - \hat{E}(x_i|z_i)\right]\left[y_i - \hat{E}(y_i|z_i)\right].$$

Since the conditional mean functions will not be estimated well where the density of $Z$ is low, Robinson makes use of a trimming function $\hat{I}_i = 1\left\{\hat{f}(z_i) > b_n\right\}$, where $\hat{f}(z)$ is a kernel density

estimator, for a given sequence of numbers $\{b_n\}$.[47] The estimator is defined as

$$\hat{\beta} = \left( \sum_{i=1}^{N} \left[ x_i - \hat{E}\left( x_i | z_i \right) \right] \left[ x_i - \hat{E}\left( x_i | z_i \right) \right]' \hat{I}_i \right)^{-1} \sum_{i=1}^{N} \left[ x_i - \hat{E}\left( x_i | z_i \right) \right] \left[ y_i - \hat{E}\left( y_i | z_i \right) \right] \hat{I}_i.$$

The estimation method is reminiscent of an interpretation of OLS estimator: consider the OLS estimation of

$$Y = X'\beta_0 + Z'\gamma + \varepsilon.$$

Then as it is well known the OLS estimator of $\beta_0$ is the OLS estimator of $u_y$ on $u_x$ where $u_y$ is the OLS residual of running $Y$ on $Z$ and $u_x$ is the OLS residual of running $X$ on $Z$.[48] Here, the first stage is replaced by nonparametric regressions.

Let $\alpha$ and $\mu$ be nonnegative real numbers and $m$ be the integer such that $m - 1 \leq \mu \leq m$. For such $\mu > 0$, $\Im_{\mu}^{\alpha}$ is the class of functions $g : R^q \to R$ satisfying: $g$ is $(m-1)$-times partially differentiable for all $z$; for some $\rho > 0$, $\sup_{y \in \{y; |y-z| < \rho\}} |g(y) - g(z) - Q_{m-1}(y, z)| / |y - z|^{\mu} \leq h(z)$, where $Q_0 = 0$ and for $m \geq 2$, $Q_{m-1}(y, z)$ is the $(m-1)$th-degree homogeneous polynomial in $y - z$ with coefficients the partial derivatives of $g$ at $z$ of order 1 through $m - 1$; and $g(z)$, its partial derivatives of order $(m-1)$ and less, and $h(z)$, all have $\alpha$th moments.

Robinson uses kernel regression estimator with independent kernel functions. He introduces the following notation: $K_l$, $l \geq 1$ is the class of even functions $k : R \to R$ satisfying

$$\int_{-\infty}^{\infty} u^i k(u)\, du = \begin{cases} 1 \text{ if } i = 0 \\ 0 \text{ if } i = 1, ...l - 1 \end{cases}$$

$$k(u) = O\left( \left( 1 + |u|^{l+1+\delta} \right)^{-1} \right), \text{ for some } \delta > 0.$$

In the statement below, $k$ is the kernel function, $a$ is the bandwidth for estimating regression function and density, and $b$ is the trimming value, $q$ is the dimension of $z$. Both $a$ and $b$ depend on $N$ although the notation does not explicitly express it.

**Theorem 3.** *(Robinson) Let the following conditions hold: (i) $(X_i, Y_i, Z_i)$, $i = 1, 2...$, are independent and distributed as $(X, Y, Z)$; (ii) the model specification is correct; (iii) $\varepsilon$ is independent of $(X, Z)$; (iv) $E\left(\varepsilon^2\right) = \sigma^2 < \infty$; (v) $E\left(|X|^4\right) < \infty$; (vi) $Z$ admits a pdf $f$ such that $f \in \Im_{\lambda}^{\infty}$, for some $\lambda > 0$; (vii) $E(X|Z = z) \in \Im_{\mu}^{2}$, for some $\mu > 0$; (viii) $\varphi(z) \in \Im_{\nu}^{4}$, for some $\nu > 0$; (ix) as $N \to \infty$, $Na^{2q}b^4 \to \infty$, $na^{2\min(\lambda+1,\mu)+2\min(\lambda+1,\nu)}b^{-4} \to 0$, $a^{\min(\lambda+1,2\lambda,\mu,\nu)}b^{-2} \to 0$, $b \to 0$; (x) $k \in K_{\max(l+m-1,l+n-1)}$, for the integers $l$, $m$, $n$ such that $l - 1 < \lambda \leq l$, $m - 1 < \mu \leq m$, and $n - 1 < \nu \leq n$. Then the condition*

$$\Phi \equiv E\left\{ [x - E(x|z)] [x - E(x|z)]' \right\} \text{ is positive definite}$$

*is necessary and sufficient for $\sqrt{N}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2 \Phi^{-1}\right)$ and*

$$\hat{\sigma}^2 \left( N^{-1} \sum_{i=1}^{N} \left[ x_i - \hat{E}\left( x_i | z_i \right) \right] \left[ x_i - \hat{E}\left( x_i | z_i \right) \right]' \hat{I}_i \right)^{-1} \xrightarrow{p} \sigma^2 \Phi^{-1},$$

[47] This trimming is used by Bickel (1982).

[48] Frisch-Waugh double residual regression. See Goldberger (1968) and Malinvaud (1970.)

*where*

$$\hat{\sigma}^2 = N^{-1} \sum_{i=1}^{N} \left[ y_i - \hat{E}\left(y_i|z_i\right) - \left(x_i - \hat{E}\left(x_i|z_i\right)\right)' \hat{\beta} \right]^2.$$

As stated earlier, the convergence rate of $\hat{\beta}$ is $\sqrt{N}$, which does not depend on the dimension of $Z$, despite the presence of $\varphi$. The theorem is stated for the kernel regression estimator, but the result should hold for other nonparametric estimators as discussed in section 7.

If $\hat{E}$ is a linear in dependent variable estimator, then $\hat{\sigma}^2$ can be rewritten as

$$N^{-1} \sum_{i=1}^{N} \left[ y_i - x_i'\hat{\beta} - \hat{E}\left(y_i - x_i'\hat{\beta}|z_i\right) \right]^2,$$

which is a natural estimator of $\sigma^2$.

Compared to the OLS estimation without $\varphi$ under homoskedasticity variance is higher because

$$Var\left(x\right) = \Phi + Var\left(E\left(x|z\right)\right).$$

When there is heteroskedasticity so that (iii) does not hold, under analogous conditions

$$\sqrt{N}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, \Phi^{-1}\Omega\Phi^{-1}\right),$$

where

$$\Omega = E\left\{\varepsilon^2 \left[x - E\left(x|z\right)\right]\left[x - E\left(x|z\right)\right]'\right\}.$$

The partially linear regression model also resembles the conditional mean function in the sample selection models. If the outcome equation is specified as $Y = X'\beta + u$ and the selection equation is specified by the latent model of the form $1z'\theta + v > 0$, where $(u, v)$ and $(X, Z)$ are independent, then without specifying the joint distribution of $(u, v)$, the following relationship holds:

$$Y = X'\beta_0 + \varphi\left(Z'\theta\right) + \varepsilon,$$
$$E\left(\varepsilon|X, Z\right) = 0.$$

Note that in this case, there is more structure in $\varphi$ function and that $\theta$ (up to a scalar) can be estimated from the data about selection. Without this structure, as discussed above, the partially linear regression model only identifies coefficients of $X$ variables that are not in the $Z$ variables.

Powell (1987) made use of this observation, modified Robinson's estimator so that there is no need for trimming, and discussed estimation of $\beta_0$. Ahn and Powell (1993) extended this approach further based on the observation that in the sample selection model one can write the conditional mean function as

$$Y = X'\beta_0 + \varphi\left(P\left(Z\right)\right) + \varepsilon,$$
$$E\left(\varepsilon|X, Z\right) = 0,$$

where $P(z)$ is the probability of being selected into samples, which can be estimated from the data about selection.[49] Ichimura and Lee (1991) propose a way of simultaneously estimating $\beta$ and $\theta$ with truncated data. Yatchew (1997) proposes to examine the differencing idea of Powell (1987) to a finite number. Heckman, Ichimura, Smith, and Todd (1998a), Heckman, Ichimura, and Todd (1998b) study estimation of $\beta$ and $\varphi(P(z))$, allowing for parametrically estimated $P(z)$ and data-dependent bandwidths. The estimator they study is basically the same with the estimator studied by Robinson but they use local polynomial estimator instead of the kernel regression estimator, instead of $Z$, they have a parametric form $P(z'\theta)$ where $\theta$ is estimated by $\hat{\theta}$ from the data on selection, use trimming based on the estimated low percentile (usually 1 or 2%) of $P(z_i'\hat{\theta})$, denoted as $\hat{q}_n$ so that the trimming function is written as $\hat{I}_i = 1(\hat{f}(\hat{P}_i) > \hat{q}_n)$ where $\hat{f}(\cdot)$ is the kernel density estimator of the density of $P(z'\theta)$, and smoothing parameter can be data dependent. Estimation of $\varphi$ is done using the estimated $\beta$ to purge $Y$ of its dependence on $X$, we can estimate $\varphi(p_0)$ by a local linear regression of $Y_i - X_i'\hat{\beta}$ on $\hat{P}_i$ evaluated at $p_0$, which we denote it by $\hat{\varphi(p_0)}$

The following theorem summarize the results by Heckman, Ichimura, Todd (1998b). $D_i$ denote the indicator whether the $i$th observation is in the sample or not.

**Theorem 4.** *Assume that*

*(i) Data* $\{(X_i, Y_i, Z_i, D_i)\}$ *are i.i.d.,* $E\{||x_i||^{2+\varepsilon} + ||z_i||^{2+\varepsilon}\} < \infty$ *for some* $\varepsilon > 0$, *and* $E\{|y_i|^3\} < \infty$,

*(ii)* $\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2}\sum_{i=1}^n \psi(z_i, d_i) + o_p(1)$, *where* $n^{1/2}\sum_{i=1}^n \psi(z_i)$ *converges in distribution to a normal random vector,*

*(iii) the kernel function* $K(\cdot)$ *is supported on* $[-1, 1]$ *and it is twice continuously differentiable,*

*(iv)* $P(z_i'\theta)$ *is twice continuously differentiable with respect to* $\theta$ *and both derivatives have second moments,*

*(v)* $E(X|P)$, $E\{\varphi(P)\}$ *are twice continuously differentiable with respect to* $\theta$,

*(vi)* $H_1 = E\{[X - E(X|P)][X - E(X|P)]'I\}$ *evaluated at the true* $\theta = \theta_0$ *is nonsingular.*

*(vii) The density of* $P(Z'\theta)$, $f_\theta$, *is uniformly bounded and uniformly continuous in the neighborhood of* $\theta_0$ *and for any* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that if* $||\theta - \theta_0|| < \delta$ *then* $\sup_{0 \le s \le 1} |f_\theta(s) - f_{\theta_0}(s)| < \varepsilon$.

*(viii)* $na_n^3/\log n \to \infty$ *and* $na_n^8 \to 0$.

*Then*

$$n^{1/2}(\hat{\beta} - \beta_0) = n^{-1/2}\sum_{i=1}^n H_1^{-1}\{[X_i - E(X|P_i)]\varepsilon_i I_i + H_2\psi(Z_i, D_i)\} + o_p(1)$$

[49] Establishing asymptotic distribution theory for an estimator that involves trimming which uses estimated $\theta$ or estimated $P(z)$ would be a non-trivial task. Powell (1987) and Ahn and Powell (1993) avoided the need for trimming by a clever re-weighting scheme. This approach have been developed to be applicable to broader models by Honoré and Powell (1994), Honoré and Powell (2005), and Aradillas-Lopez, Honoré, and Powell (2005).

*where* $H_2 = E\{[X - E(X|P)]P(Z'\theta_0)[Z - E(Z|P)]'I\}$.

*If in addition to the assumptions above, the following assumptions hold:*

*(ix)* $\varphi$ *is twice continuously differentiable,*

*(x)* $f_{\theta_0}(p_0) > 0$*,*

*(xi) the bandwidth sequence satisfies* $\hat{a}_n = \hat{\alpha}_n n^{-1/5}$, plim $\hat{\alpha}_n = \alpha_0 > 0$*,*

*(xii)* $\sigma^2(p_0) = E[|Y - X'\beta|^2 \,|P = p_0]$ *is finite and continuous at* $p_0$*,*

*then,*

$$n^{2/5}(\hat{\varphi}(p_0) - \varphi(p_0)) \sim N(B, V)$$

*where*

$$B = \frac{1}{2}\varphi''(p_0)\left[\int s^2 K(s)ds\right]\alpha_0^2$$

$$V = \frac{\text{Var }(Y - X'\beta|P = p_0)}{f_{\theta_0}(p_0)\alpha_0}\int K^2(s)ds,$$

*where* $\varphi''(p_0)$ *is the second derivative of the regression function.*

### 5.2 Improving the convergence rate by changing the parameter

The prototypical way to improving the convergence rate is by averaging. If we give up estimating a function at a point and instead average the point estimates over a region, we can, under some conditions, improve the convergence rate. This point is clear enough for the case of the conditional mean function $m(x) = E(Y|X = x)$. We saw that the convergence rate of the estimation of the conditional mean function depends on the number of continuous conditioning variables and the underlying smoothness of the conditional mean function with respect to these variables. Let $X = (X_1, X_2)$. Instead of estimating $m(x_1, x_2)$, one can estimate $m(x_1, A) = E(Y|X_1 = x_1, X_2 \in A)$ for some region $A$. In this case, since it is equivalent to having less continuous regressors, the convergence would only depend on the number of continuous regressors among $X_1$.

An analogous result holds for the estimation of the average of a nonparametric estimator of the derivative of a function. The average derivative estimator is examined by Stoker (1986) and its asymptotic distribution theory, in modified forms, is established by Powell, Stock, and Stoker (1989), Robinson (1989), and Härdle and Stoker (1989). Newey and Stoker (1993) discusses efficiency issues.

Note that when $E(Y|X) = g(X)$, the solution to

$$\min_b E\left\{[Y - X'b]^2\right\} = \min_b E\left\{[g(X) - X'b]^2\right\} + E[Var(Y|X)]$$

corresponds to the OLS estimator and $b^* = E(XX')^{-1}E[XY]$ which can be interpreted as the best predictor of the form $X'b^*$ as observed by White (1980).

Since we are also interested in measuring the marginal effect of a change in regressors to dependent variables, $\partial g/\partial x$, we may also want to estimate $\delta_k$ that solves

$$\min_{\delta_k} E\left\{(\partial g/\partial x - \delta_k)^2\right\}$$

for each $k = 1, ..., d$. Clearly the solution is $\delta_k^* = E\{\partial g/\partial x\}$. Stoker (1986) proposed estimation of $\delta_k^*$.

Another case $\delta_k^*$ is of interest is when $g(x) = \phi(x'\beta_0)$. Stoker observed that many limited dependent variable models have this property. In this case

$$\partial g/\partial x = \phi'\left(x'\beta_0\right) \cdot \beta_0.$$

Thus $E(\partial g/\partial x) = c \cdot \beta_0$ for some constant $c$: estimation of the average derivative corresponds to $\beta_0$ parameter up to a constant term.

However, the interpretation of the average derivative as $\beta_0$ parameter up to a constant term depends on the assumption that (i) there is no discrete regressors among regressors and (ii) there is no functional relationship among regressors. These two assumptions may make the direct application of the average derivative method unsuitable for many limited dependent variable models. This issue is not relevant if we interpret the average derivative in a nonparametric context.

As we discussed earlier, $g(x)$ and its derivatives can be estimated consistently by non-parametric estimators. But as we noted there, convergence rate is very slow especially when $K$ is large and/or when we estimate higher order derivatives. It turns out that $\delta_k^*$ can be estimated $1/\sqrt{n}$-consistently, the typical rate at which parametric estimators converge.

Let $\hat{\Delta}(x)$ be a nonparametric estimator of $\partial g/\partial x$ at a point $x$. Then a natural estimator of $\delta^* = (\delta_1^*, \delta_2^*, ..., \delta_d^*)'$ is

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}(x_i).$$

Stoker does not examine this estimator but instead bases his estimator on the integration by parts argument. We present his argument for one dimension case but the same argument goes through for a higher dimension: with an appropriate boundary conditions as made explicit in the computation below

$$\begin{aligned} E\left(g'\right) &= \int_{-\infty}^{\infty} g'(x) f(x)\, dx \\ &= g(x) f(x)\,|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} g(x) f'(x)\, dx \\ &= -\int_{-\infty}^{\infty} g(x) \frac{f'(x)}{f(x)} f(x)\, dx = -E\left(Y \frac{f'}{f}\right) \end{aligned}$$

Thus by making use of a nonparametric estimator of $f(x)$ and its derivative, one can estimate the average derivative. As the ratio $f'/f$ won't be estimable where $f$ is low, the estimator is defined

making use of a trimming function $\hat{I}_i = 1\left\{\hat{f}(x_i) > b_n\right\}$ for a given sequence of numbers $\{b_n\}$.

$$\begin{aligned}\hat{\delta} &= \frac{1}{n}\sum_{i=1}^{n}\frac{-\partial\hat{f}(x_i)/\partial x}{\hat{f}(x_i)}y_i\hat{I}_i, \text{ where}\\ \hat{f}(x) &= \frac{1}{n}\sum_{j=1}^{n}\frac{1}{a_n^K}K\left(\frac{x-x_j}{a_n}\right).\end{aligned}$$

The estimator can be obtained directly without any optimization. Härdle and Stoker (1989) shows:

**Theorem 5.** *Consider $y_i = g(x_i) + \varepsilon_i$ with $E(\varepsilon_i|x_i) = 0$ under iid sampling. If*

1. *The regressors have density $f(x)$ where the support of $f$ is a convex subset of $R^K$,*
2. *$f(x) = 0$ at the boundary of the support,*
3. *$g(x)$ is continuously differentiable almost everywhere,*
4. *$E\left\{y^2(\partial\log f(x)/\partial x)(\partial\log f(x)/\partial x)'\right\}$ and $E\left\{(\partial g/\partial x)(\partial g/\partial x)'\right\}$ are finite and $E\left(y^2|x\right)$ is continuous,*
5. *$f(x)$ is differentiable up to $p \geq K+2$,*
6. *$f(x)$ and $g(x)$ obey local Lipschitz conditions, i.e. for $v$ in neighborhood of 0, there exist functions $\omega_f$, $\omega_{f'}$, $\omega_{g'}$, and $\omega_{\ell g}$ such that*

$$\begin{aligned}|f(x+\nu) - f(x)| &\leq \omega_f(x)|\nu|,\\ \left|f'(x+\nu) - f'(x)\right| &\leq \omega_{f'}(x)|\nu|,\\ \left|g'(x+\nu) - g'(x)\right| &\leq \omega_{g'}(x)|\nu|\\ \left|\frac{\partial\log f(x+\nu)}{\partial x}g(x+\nu) - \frac{\partial\log f(x)}{\partial x}g(x)\right| &\leq \omega_{\ell g}(x)|\nu|\end{aligned}$$

   *where second moments of $\omega_f$, $\omega_{f'}$, $\omega_{g'}$, and $\omega_{\ell g}$ are all finite,*

7. *Let $A_n = \{x|f(x) > b_n\}$. As $n \to \infty$,*

$$\int_{A_n^c} g(x)\frac{\partial f(x)}{\partial x}dx = o\left(n^{-1/2}\right),$$

8. *Let $f^{(p)}$ denote the pth order derivative of $f$. $f^{(p)}$ is locally Hölder continuous: there exists $\gamma > 0$ and $c(x)$ such that*

$$\left|f^{(p)}(x+\nu) - f^{(p)}(x)\right| \leq c(x)|\nu|^{\gamma},$$

   *where second moments of $f^{(p)}$ and $c(x)$ are finite,*

9. *The kernel function* $K(u)$, $u \in R^K$ *has finite support, is symmetric, has* $p+\gamma$*-absolute moments, and* $K(u)=0$ *at the boundary points, and* $K(u)$ *is of order* $p$*, i.e.* $\int_{R^K} K(u)\,du = 1$,

$$\begin{aligned}\int_{R^K} u_1^{\ell_1} u_2^{\ell_2} \cdots u_\rho^{\ell_\rho} K(u_1, u_2, ..., u_K)\,du &= 0 \text{ where } \ell_1 + \cdots + \ell_\rho < p, \text{ for all } \rho \leq K, \text{ and} \\ \int_{R^K} u_1^{\ell_1} u_2^{\ell_2} \cdots u_\rho^{\ell_\rho} K(u_1, u_2, ..., u_K)\,du &\neq 0 \text{ where } \ell_1 + \cdots + \ell_\rho = p, \text{ for all } \rho \leq K,\end{aligned}$$

10. *As* $n \to \infty$, $a_n \to 0$, $b_n \to 0$, $a_n/b_n \to 0$ *and for some* $\varepsilon > 0$, $n^{1-\varepsilon} a_n^{2K-2} b_n^4 \to \infty$, *and* $n a_n^{2p-2} \to 0$,

*Then*

$$\sqrt{n}\left(\hat{\delta} - \delta\right) \xrightarrow{d} N(0, \Sigma),$$

*where*

$$\Sigma = E\left\{\left[\frac{\partial g}{\partial x} - E\left(\frac{\partial g}{\partial x}\right)\right]\left[\frac{\partial g}{\partial x} - E\left(\frac{\partial g}{\partial x}\right)\right]'\right\} + E\left\{\sigma_\varepsilon^2 \frac{\partial \log f(x)}{\partial x}\frac{\partial \log f(x)}{\partial x'}\right\}.$$

Although $b_n$ has to converge to zero, there is no restriction at the speed at which that convergence has to happen in this condition. The speed requirement comes from assumption 7. As $n a_n^{2p-2} \to 0$, the parameter $a_n$ does need to converge to zero sufficiently fast. In order for these bandwidth requirements to be mutually consistent, the density $f$ needs to approach 0 sufficiently smoothly.

As observed above, the estimator is based on some boundary conditions. When the boundary conditions do not hold, then direct estimation of the average of a nonparametric estimator of the derivative would be preferable. Also, in deriving the theoretical properties of the estimator, negative kernel functions are used to "kill" the bias term asymptotically. Additionally, $E(\varepsilon|x) = 0$ is needed, so that models with endogenous regressors can not be treated with this estimator. Lastly, if some of the regressors are discrete, the derivative is clearly not defined. Even in this case, however, if one restricts taking derivative with respect to the continuous regressors, then the arguments would go through without a modification. See Härdle and Stoker (1989) for estimation of the asymptotic variance-covariance matrix. Newey and Stoker (1993) showed that the estimator has the variance and covariance matrix that coincides with the smallest variance-covariance matrix within nonparametric estimators that are $1/\sqrt{n}$-consistent to $\delta^*$.

### 5.3 Usage of different stochastic assumptions

As we discussed in the context of the censored regression model, a quantile restriction leads to $n^{1/2}$-consistent estimator even in the presence of an infinite dimensional nuisance parameter. This important result was shown by Powell(1984). A conditional mean restriction is not sufficient. The same idea applied to the binary response model does not lead to $n^{1/2}$-consistent estimator. We will see why via a discussion of Manski's (1975, 1985) maximum score estimator.

### 5.3.1 Censored Regression Model

The model we study is

$$\begin{aligned} y_t^* &= x_t'\beta_0 - \varepsilon_t \\ y_t &= \begin{cases} y_t^* \text{ if } y_t^* > 0 \\ 0 \text{ if } y_t^* \leq 0. \end{cases} \end{aligned}$$

where the conditional median of $\varepsilon$ is assumed to be 0. In econometric literature, Powell (1984) is the first to explicitly recognize essentially the parametric nature of the conditional quantile function under the censored regression model even though the conditional distribution of $\varepsilon$ is restricted to have the conditional median to be 0.

There are two observations that lead to Powell's estimator. First, when $x_t'\beta_0 > 0$ the median of observed dependent variable is still $x_t'\beta_0$ and that when $x_t'\beta_0 < 0$ the median of observed dependent variable is 0 so that the median of the observed dependent variable is known to have the following parametric form:

$$\max\left\{0, x_t'\beta_0\right\}.$$

Second, the minimizer of $\sum_{t=1}^{T}|y_t - a|$ over $a$ estimates the median consistently. Thus the estimator is defined as the minimizer of

$$\inf_b \sum_{t=1}^{T} \left|y_t - \max\left\{0, x_t'b\right\}\right|.$$

Powell (1984) showed that the estimator is $n^{1/2}$-consistent and asymptotically normal:

$$\sqrt{T}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \lim_{T\to\infty} C_T^{-1} M_T C_T^{-1}\right)$$

where

$$\begin{aligned} C_T &= E\left\{T^{-1}\sum_{t=1}^{T} 2f_t\left(0|x_t\right)\cdot 1\left(x_t'\beta_0 > 0\right) x_t x_t'\right\} \text{ and} \\ M_T &= E\left\{T^{-1}\sum_{t=1}^{T} 1\left(x_t'\beta_0 > 0\right) x_t x_t'\right\}. \end{aligned}$$

When $f_t\left(0|x_t\right) = f\left(0\right)$, $C_T = 2f\left(0\right)$ and thus

$$\sqrt{T}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \lim_{T\to\infty} \frac{1}{4f\left(0\right)} M_T^{-1}\right).$$

Under this assumption Powell provides consistent estimator of $f\left(0\right)$ and $\lim_{T\to\infty} M_T$. When we have i.i.d. sampling $f\left(0|x\right)$ can be estimated consistently and thus $C_T$ also, under some regularity conditions.

### 5.3.2 Binary response model

For the case of binary response, the model is:

$$y_i = 1\left(x_i'\beta_0 - \varepsilon_i > 0\right).$$

Observe that

$$E\left(y_i|x_i\right) = F_\varepsilon\left(x_i'\beta_0|x_i\right),$$

where $F_\varepsilon$ is the cumulative distribution function of $\varepsilon$. If the median of $\varepsilon_i$ given $x_i$ is 0, that is, if

$$F_\varepsilon\left(s|x_i\right) = 1/2 \text{ if and only if } s = 0,$$

then the medial of $y_i$ given $x_i$ is 1 if $x_i'\beta_0 > 0$ and 0 if $x_i'\beta_0 < 0$. That is the conditional median function of $y_i$ is known to be parametric and the form is $1\left(x_i'\beta_0 > 0\right)$. Thus, based on the quantile regression idea, a natural estimator is to find the minimizer of the following objective function

$$\sum_{i=1}^{n}\left|y_i - 1\left(x_i'\beta > 0\right)\right|,$$

as in the censored LAD estimator. As Manski (1985) discusses, minimizing this objective function is equivalent to maximizing the maximum score objective function of Manski (1985):

$$\sum_{i=1}^{n}(2y_i - 1)sign(x_i'\beta),$$

where $sign(s)$ equals 1 if $s > 0$ and $-1$ if $s < 0$ and equals 0 if $s = 0$. Unlike the objective function of the censored LAD estimator, this objective function changes the value around the points $x_i'\beta = 0$. As the observations corresponding to this line is measure zero when there is a continuous regressor is present, the convergence rate is not $n^{-1/2}$. Kim and Pollard (1990) showed that in fact the estimator converges with rate $n^{-1/3}$. Note that this convergence rate corresponds to that of nonparametric estimators which do not exploit smoothness. Horowitz (1992) showed how to exploit the smoothness of the underlying conditional CDF and improve the convergence rate when indeed the underlying CDF is smooth. His estimator replaces the unsmooth sign function by a smooth function.[50]

# 6 Smoothing parameter choice and trimming

The flexible estimators described in sections 4 and 5 are specified up to some choice of smoothing parameter. For local estimators, the smoothing parameter choice corresponds to choosing the bandwidth parameter. For global estimators, the smoothing parameter choice corresponds to choosing the bases functions to include in the expansion. For semiparametric estimators, in addition to choosing the smoothing parameter, implementation of the estimators also requires choosing

[50]Horowitz (1992) implementation does not exactly correspond to a smoothed version of the Manski's objective function as Horowitz replaces the sign function with a smooth CDF function.

a method of trimming the data, as discussed in section 5. In this section, we discuss the problem of smoothing parameter choice in the context of density and conditional mean function estimation and also in the context of semiparametric estimation. We also discuss trimming methods.

One way of choosing smoothing parameters is to use graphical diagnostics, which reveal how an estimated surface changes in response to varying the smoothing parameters. For a simple problem, some argue that this can be a reasonable way of selecting smoothing parameters. But this procedure is subjective and hence the choice would be hard to justify formally or communicate to others. In addition, even at the subjective level, it is questionable if we can visualize something corresponding to bias and variance of the estimator. Moreover, for a higher dimensional problems or for cases where nonparametric estimators are being used as input into a semiparametric estimation problem, an implicit criteria the graphical approach uses is not necessarily appropriate and is too user-intensive to be practical. A more automatic bandwidth selection method is needed. For nonparametric density and regression estimation, the importance of developing data-based methods to guide researchers in selecting bandwidths is well recognized and a variety of bandwidth selectors have been proposed in the statistics and econometrics literatures. All the methods select the bandwidth to minimize error in estimation with respect to a certain criteria. They differ in the criteria used for measuring estimation error.

We summarize results in the literature as well as our own Monte Carlo studies evaluating the performance of different smoothing parameter selection methods. Our discussion is limited to the bandwidth selection methods for the kernel density estimator and local polynomial estimators.

There are two types of smoothing parameters: constant, or sometimes refer to as global, and variable. A global smoothing parameter is held fixed for the entire domain of the function being estimated and a variable smoothing parameter is allowed to vary at each point of the domain.[51]

For the density estimation, we discuss global bandwidth choice for the kernel density estimator. The advantage of a variable bandwidth is that it adapts better to the design of the data. A disadvantage, in the case of the kernel density estimation, is that once the bandwidth is allowed to depend on the data, the resulting estimator is no longer guaranteed to be a density. For regression estimation, this problem does not exist so we will consider both global and local bandwidth selectors.

## 6.1 Methods for selecting smoothing parameters in the kernel density estimation

As was shown in Table 1, the efficiency of the kernel density estimator

$$f_n(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

depends more on the choice of bandwidth $h$ than on the choice of kernel function $K(\cdot)$ within a class of commonly used kernels. Therefore, in the following discussion we take the choice of kernel function as given and focus on the question of how to choose the smoothing parameter.

[51] Fan and Gijbels (1992) studies a bandwidth selection method which differ for each data point and refers to the method as a "global variable" method.

The three bandwidth selection methods we discuss are the rule of thumb (ROT) method, the least square cross validation (LSCV), and the smoothed bootstrap (SB) method by Taylor (1989).

The ROT method is chosen for its simplicity in implementation. The other two methods are chosen for their theoretical coherence as well as reliable performance in Monte Carlo studies in the literature and our own.

The loss function underlying all three methods of selecting the bandwidth of the kernel density estimator is the highest order of the integrated mean squared error:

$$\int E\{(f_n(x;h) - f(x))^2\}dx = \int [Var((f_n(x;h))) + Bias^2(x)]dx,$$

where $Bias(x) = E[f_n(x;h)] - f(x)$ and here and below, the integration is taken over the whole real line.

Three methods differ in ways to approximate this objective function. If we wish to choose the bandwidth local to a particular point $x$, then clearly we should examine $E\{(f_n(x;h) - f(x))^2\}$ at the point $x$ rather than examining the overall measure such as above.

**Rule of Thumb** Under suitable regularity conditions IMSE can be approximated by the sum of two terms:

$$AIMSE(h) = \frac{c_{2K}}{nh} + \frac{\sigma_K^2}{4}h^4 \int [f''(x)]^2 dx,$$

where $c_{2K} = \int K^2(s)ds$ and $\sigma_K^2 = \int s^2 K(s)ds$. The first term represents the variance and the second term represents the bias term.[52]

The $h$ that minimizes the AIMSE is

$$h_{AIMSE} = \left[\frac{c_{2K}}{\sigma_K^2} \frac{1}{[\int [f''(x)]^2 dx]}\right]^{1/5} n^{-1/5}. \tag{6.1}$$

The optimal bandwidth decreases with the size of the sample and increases when the effect of bias on the AMISE is greater; i.e. when $\int [f''(x)]^2 dx$ is larger.

From equation (6.1) we see that estimating the global optimal plug-in bandwidth that minimizes the *AIMSE* requires obtaining an estimate of $\int [f''(x)]^2 dx$.

ROT estimates the unknown quantity by assuming a value based on a parametric family, usually the $N(\mu, \sigma^2)$ distribution. Under normality,

$$\int f''(x)^2 dx = \sigma^{-5} \frac{3}{8\sqrt{\pi}} \approx 0.212\sigma^{-5}.$$

If in addition a normal kernel is used, the ROT bandwidth is approximately equal to $1.06\sigma n^{-1/5}$.

[52] The highest order approximation to the MSE at point $x$ is

$$\frac{c_{2K}}{nh} f(x) + \frac{1}{4}h^4 \sigma_K^2 [f''(x)]^2.$$

There is a different trade-off between variance bias at each point reflecting different values of $f(x)$ and $f''(x)$. Thus it seems more desirable to choose a point-wise bandwidth.

Because the scale parameter $\sigma$ is potentially sensitive to outliers, Silverman (1986) suggests using a more robust *rule-of-thumb* estimator, where the interquartile range of the data replaces the sample standard deviation as a scale parameter. It is given by $h_{ROT} = 1.06 \min(\hat{\sigma}, \hat{R}/1.34)n^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation and $\hat{R}$ the estimated interquartile range. (for Gaussian data, $\hat{R} \approx 1.34\hat{\sigma}$).[53]

Clearly, when the underlying density is not normal, the ROT method does not consistently estimate the $h_{AIMSE}$ and hence suboptimal. However, because it converges to 0 with an appropriate rate, it does yield a consistent and asymptotically normal kernel density estimator when the underlying density is twice continuously differentiable.

**Least Square Cross validation** The least square cross validation (LSCV) discussed by Stone (1974) chooses the bandwidth that minimizes the estimated integrated squared error (ISE):

$$ISE = \int [f_n(x;h) - f(x)]^2 dx.$$

Hall (1982) showed that under regularity conditions

$$ISE = IMSE + o_p(h^4 + (nh)^{-1})$$

so that minimizing the ISE and minimizing the IMSE is equivalent to the first order under some regularity conditions.

Note that

$$ISE = \int [f_n(x;h)]^2 dx - 2\int f_n(x;h)f(x)dx + \int [f(x)]^2 dx$$

and that the last term does not depend on $h$ so minimizing the sum of the first two terms is equivalent to minimizing the ISE. Although the second term is not computable because $f(x)$ is not known, its unbiased estimator can be constructed by

$$-2\frac{1}{n}\sum_{i=1}^{n} f_{ni}(X_i;h),$$

where $f_{ni}(x;h) = (n-1)^{-1}\sum_{j\neq i} K((x-X_i)/h)/h$.

Thus the LSCV chooses the bandwidth that minimizes

$$AISE(h) = \int [f_n(x;h)]^2 dx - 2\frac{1}{n}\sum_{i=1}^{n} f_{ni}(X_i;h).$$

Note that if we use $f_n(x;h)$ in place of $f_{ni}(x;h)$, then the LSCV yields an inconsistent method. To see this observe that

$$\int [f_n(x;h)]^2 dx = \frac{\int K^2(s)ds}{nh} + \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j\neq i} K * K((X_i - X_j)/h),$$

[53]The *rule-of-thumb* method can of course be tailored to a particular application. For example, if a researcher strongly suspected bimodality in the density, he/she may want to use a bimodal parametric density for the plug-in estimator.

where $K * K(u) = \int K(u-s)K(s)ds$. So if there is no duplication in the observations $\{X_i\}_{i=1}^n$ and

$$\int K^2(s)ds < 2K(0)$$

and $\lim_{|s|\to\infty} |s|K(s) = 0$ as well as $\lim_{|s|\to\infty} |s|K * K(s) = 0$, then choosing $h$ small will make the objective function small. Since this holds regardless of $f(x)$, the LSCV yields an inconsistent method. Note that $\int K^2(s)ds < 2K(0)$ holds for most kernel functions such as those densities that has a single peak at 0.[54]

When there is no duplication of observations, on the other hand, the "delete one" modification fixes the problem as defined above. However, the same issue which was avoided by the "delete one" modification arises when there are duplication of observations. Since the duplication of observations arises naturally if there is discretization, one needs to be aware of this potential problem when applying the LSCV.

Hall (1983) and Stone (1984) justified LSCV as a data dependent method to choose the optimal bandwidth. In particular, Stone (1984) showed that, only assuming boundedness of $f(x)$ (and its marginals, for multivariate case),

$$\frac{ISE(h_{LSCV})}{ISE(h_{opt})} \to 1$$

as $n \to \infty$ with probability 1, where $h_{opt}$ minimizes $ISE(h)$.

**Smoothed Bootstrap** The smoothed bootstrap method of Taylor (1989) is motivated by the formula obtained when estimating $\int E\{[f_n(x;h) - f(x)]^2\}dx$ by a bootstrap sample generated from $f_n(x;h)$. That is, writing $X_i^*$ to be sampled from distribution $f_n(x;h)$, one can estimate $\int E\{[f_n(x;h) - f(x)]^2\}dx$ by

$$E^*\left\{\frac{1}{nh}\sum_{i=1}^n K((x-X_i^*)/h) - \frac{1}{nh}\sum_{i=1}^n K((x-X_i)/h)\right\}^2.$$

Taylor (1989) observes that this can be explicitly computed when Gaussian kernel is used and its integration over $x$ is:

$$\frac{1}{2n^2h(2\pi)^{1/2}}\left[\sum_{i=1}^n\sum_{j=1}^n \exp\left\{-\frac{X_j-X_i}{8h^2}\right\} - \frac{4}{3^{1/2}}\sum_{i=1}^n\sum_{j=1}^n \exp\left\{-\frac{X_j-X_i}{6h^2}\right\} + 2^{1/2}\sum_{i=1}^n\sum_{j=1}^n \exp\left\{-\frac{X_j-X_i}{4h^2}\right\} + n2^{1/2}\right].$$

He modifies the above formula to sum over $i \neq j$. The modified objective function, $B^*(h)$, say, is then minimized to define the data dependent bandwidth.

[54] For these functions $\int K^2(x)dx$ can be regarded as the mean of $K(x)$ and it has to be lower than its maximum $K(0)$.

Taylor (1989) shows that

$$Var\{B^*(h)\} = \frac{0.026}{8n^2h\pi^{1/2}} \int [f(x)]^2 dx + O(h/n^2).$$

It is an order of magnitude less than the corresponding object for the LSCV, $Var(AISE(h))$ computed by Scott and Terrell (1987):

$$Var(AISE(h)) = \frac{4}{n}\left[\int [f(x)]^3 dx - \left\{\int [f(x)]^2 dx\right\}\right] + O(1/(n^2h) + h^4/n).$$

**A brief discussion of other methods** Other methods which perform well in Monte Carlo studies is the method of Sheather and Jones (1991) and its modification by Jones, Marron, and Sheather (1996). We did not discuss this method here as the method seems theoretically incoherent. Like the ROT method, their approach targets the optimal bandwidth when the underlying density is twice continuously differentiable. But the method presumes that the density has higher order derivatives so that the target is not necessarily an interesting object from a theoretical point of view.

From a statistical perspective, the least square based objective functions we have discussed above may seem ad hoc. Indeed the literature has considered likelihood based methods to selecting the bandwidth as well. However, Schuster and Gregory (1981) showed that when the tail of the target density is thicker than exponential decay, then choosing the bandwidth by the likelihood based cross validation leads to an inconsistent density estimator.

**Empirical performance** Several published studies examine bandwidth performance in real data examples and in Monte Carlo settings. They include Jones, Marron and Sheather (1992), Cao, Cuevas, and Gonzales-Maniega (1994), Park and Turlach (1992), Park and Marron (1990), Härdle (1991), Cleveland and Loader (1996) and Loader (1995). Below we summarize commonalities and disparities in findings across studies and then present some findings from our own Monte Carlo study. More empirical evidence needs to be accumulated to better understand how different methods compare under data designs that commonly arise in economics.

In their evaluation of rule-of-thumb (ROT) methods, Silverman (1986), JMS (1992) and Härdle (1991) conclude that a ROT estimator with a normal reference density has a tendency to over-smooth, or choose too large a bandwidth, particularly when the data is highly skewed or is multi-modal. In two separate examples, Jones, Marron, and Sheather (1996) (Hereafter JMS) and Härdle find that the ROT estimator is unable to detect a simple case of bimodality.[55]

The LSCV estimator tends to suffer from the opposite problem: under-smoothing. JMS conclude that because of under-smoothing, the LSCV procedure leads to high variability and overall unreliability in choosing the optimal bandwidth. Hall and Marron (1991) partly explain the under-smoothing tendency by showing that LSCV frequently gives local minima and the tendency to

[55]This drawback could possibly be overcome by using a more flexible parametric family as a reference in constructing the plug-in estimate of $\int [f''(x)]^2 dx$. For example, a mixture of normals could be used.

under-smooth likely comes from not finding the global minimum. Park and Marron (1990) and Loader (1995) point out that LSCV is nonetheless the method of choice for cases where the researcher is only willing to maintain a limited degree of smoothness on the true density. Most other bandwidth selection methods require smoothness assumptions on higher order derivatives. In Loader's simulations, the LSCV approach performs well. This was also the finding in our own simulations.

The SB selector has only been studied in a few papers. JMS find its performance to be close to that of the Sheather and Jones' method. Faraway and Juhn (1990) compare the SB and LSCV procedures and find that SB performs better, which they attribute mostly to its lower variability. For further evidence on relative performance of bandwidth selectors, see Hall, Sheather, Jones and Marron (1991), and Park, Kim, and Marron (1994), and Loader (1995).

## 6.2 Methods for selecting smoothing parameters in the local polynomial estimator of a regression function

Here, we consider the problem of choosing the smoothing parameter for a local polynomial estimator of a fixed degree; typically equal to one (i.e. local linear regression). In particular we discuss a rule of thumb method by Fan and Gijbels (1996), the least square cross validation, and Fan and Gijbels's method (1995) of residual square criteria (RSC). These methods do not require an initial bandwidth selection. We also discuss Fan and Gijbels's (1995) finite sample approximation method as a prototype of an attempt to improving on these methods.

These methods are the standard bandwidth selection methods, but the limitation of these methods are discussed in view of the alternatives proposed by Fan, Hall, Martin, and Patil (1996), Doksum, Peterson, and Samarov (2000), and Prewitt and Lohr (2006).

### 6.2.1 A general discussion

All the methods we discuss estimate in some ways asymptotic mean square error (AMSE) of estimating $D^j m(x_0)$ ($j = (j_1, \ldots, j_d)$) at a point for the case of the local bandwidth or its integral with some weights for the case of the global bandwidth. For the local polynomial estimator of order $p$ when the underlying one dimensional regression function is at least $p+1$ times continuously differentiable function the AMSE at a point can be obtained by inspecting Theorem 2:

$$AMSE(x_0) = [(M^{-1}Bm^{(p+1)}(x_0))_\ell]^2 h^{2(p+1-|j|)} + \frac{\sigma^2(x_0)/f(x_0)}{nh^{d+2|j|}}(M^{-1}\Gamma M^{-1})_{\ell,\ell}$$

where $\ell$ is the order in which $j$ appear. Note that in using this formula, we assume that $p - |j|$ is odd so that the bias term does not vanish.

Thus the asymptotically optimum point-wise bandwidth is

$$h_{opt,p,j}(x_0) = \left[\frac{[(d+2|j|)(M^{-1}\Gamma M^{-1})_{\ell,\ell}\sigma^2(x_0)/f(x_0)]}{2(p+1-|j|)[(M^{-1}Bm^{(p+1)}(x_0))_\ell]^2 n}\right]^{1/(2p+d+2)}.$$

The optimum bandwidth depends on three factors: the conditional variance, the density of regressors, and the $p+1$st derivative of the underlying function. The $p+1$st derivative enters because

we consider the local polynomial estimator of order $p$ and the size of the $p+1$st derivative captures a local deviation from the $p$th order model used. When there is a larger variance (high $\sigma^2(x_0)$), less data (low $f(x_0)$), or less deviation from the model (high $\|m^{(p+1)}(x_0)\|$) then we want to use a wider bandwidth.

Sometimes, statistical packages choose a fixed proportion of the data nearest to the point of evaluation ($x_0$) by default. This approach will effectively choose wider bandwidth at a lower density region. In view of the result above, this may be appropriate when the variance and the model approximation is roughly constant. However, generally the approach cannot be an optimal way to choose the bandwidth as it does not have a way to accommodate the two other factors affecting the optimal bandwidth. In addition, the method does not give us an idea what the appropriate fixed proportion may be.

### 6.2.2 One step methods

**Rule of Thumb** Fan and Gijbels (1996) proposes a ROT method for choosing a global bandwidth. Optimum global bandwidth is obtained by minimizing the integrated version of the AMSE(x) using some weight function, say $w(x)$ over $x$:

$$AMSE = \int [(M^{-1}Bm^{(p+1)}(x))_\ell]^2 w(x)dx h^{2(p+1-|j|)} + \frac{\int[\sigma^2(x)/f(x)]w(x)dx}{nh^{d+2|j|}}(M^{-1}\Gamma M^{-1})_{\ell,\ell}$$

Thus the optimum global bandwidth is expressed exactly as the local one except that each of the functions in the expression above are replaced by the integrated versions:

$$h_{opt,global,p,j} = \left[\frac{[(d+2|j|)(M^{-1}\Gamma M^{-1})_{\ell,\ell}\int[\sigma^2(x)/f(x)]w(x)dx]}{2(p+1-|j|)\int[(M^{-1}Bm^{(p+1)}(x))_\ell]^2 w(x)dxn}\right]^{1/(2p+d+2)}.$$

They propose to use $w(x) = f(x)w_0(x)$ for a given $w_0(x)$, estimate $m(x)$ by a global polynomial of order $p+3$, $\hat{m}_{p+3}(x)$ so that the $p+1$st derivative $\hat{m}^{(p+1)}_{p+3}(x)$ has enough flexibility, and use the residuals $y_i - \hat{m}_{p+3}(x_i)$ from the global polynomial regression to estimate the global residual variance, say $\hat{\sigma}^2$ and defined the ROT bandwidth:

$$h_{ROT,p,j} = \left[\frac{[(d+2|j|)(M^{-1}\Gamma M^{-1})_{i,i}\hat{\sigma}^2\int w_0(x)dx}{2(p+1-|j|)\sum_{i=1}^n[(M^{-1}B\hat{m}^{(p+1)}_{p+3}(x_i))_\ell]^2 w_0(x_i)}\right]^{1/(2p+d+2)}.$$

Effectively, the method presumes homoskedasticity. Note that $\int[(M^{-1}Bm^{(p+1)}(x))_\ell]^2 w(x)dxn$ is replaced by its consistent estimator

$$\sum_{i=1}^n [(M^{-1}B\hat{m}^{(p+1)}_{p+3}(x_i))_\ell]^2 w_0(x_i).$$

In implementation, they used a constant function on the support of the regressors as $w_0$.

**Least Square Cross Validation** The LSCV bandwidth is a method for obtaining the optimum bandwidth for estimating the conditional mean function. A global bandwidth is chosen to minimize a weighted sum of the squared prediction errors:

$$h_{LSCV} = \arg\min_h \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{i,h}(x_i))^2 w_0(x_i).$$

where $\hat{m}_{i,h}(x_i)$ is the local polynomial regression function estimator computed without using the $i$th observation but evaluated at $x_i$. The $i$th observation has to be omitted, because if we use all observations to estimate the conditional mean function, by choosing the bandwidth very small, one can always make the objective function 0. [56]

Another consideration in carrying out LSCV is that the local linear estimator in one dimension is defined only when there are at least two data points within the support of the kernel weight function. This effectively places a lower bound on the values of bandwidths that can be considered.[57]

Note the importance of using $w_0(x)$ in the objective function. Without the weight function LSCV chooses a global bandwidth with $w(x) = f(x)$ thus unless the regressor distribution is bounded, the objective function may not converge to a meaningful object when the conditional variance is bounded away from zero, for example.

**Residual Squares Criterion** Fan and Gijbels (1995) proposes an objective function for choosing the bandwidth appropriate for estimating the conditional mean function and its derivatives by the local polynomial estimator of order $p$ in one dimensional problems. Note that in one dimensional problems the $AMSE(x_0)$ simplifies to

$$AMSE(x_0) = [(M^{-1}B)_\ell m^{(p+1)}(x_0)]^2 h^{2(p+1-|j|)} + \frac{\sigma^2(x_0)/f(x_0)}{nh^{d+2|j|}}(M^{-1}\Gamma M^{-1})_{\ell,\ell}$$

because $m^{(p+1)}(x_0)$ is a scalar. Thus the bandwidth optimal for estimating the regression function can be adjusted by a known factor to produce the optimum bandwidth suitable for estimating the derivatives of the regression function. Thus they study

$$RSC(x_0) = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 K_h(x_i - x_0)}{trace\{W - WX(X'WX)^{-1}X'W\}}(1 + (p+1)\hat{V}),$$

where $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)' = X\hat{\beta}$ is the local polynomial fit using all the estimated coefficients, $\hat{V}$ is the 1-1 element of $(X'WX)^{-1}(X'W^2X)(X'WX)^{-1}$. The minimizer of this objective function multiplied by a known factor is the local RSC bandwidth. The multiplying factor depends on $p$ and the order of the derivative being estimated. They show that this method selects the locally optimum bandwidth asymptotically.[58]

[56] When there are duplicate observations in the sense that the $(y_i, x_i)$ pair is the same for multiple observations, then the "leave-one-out" $\hat{m}_{i,h}(x_i)$ estimator needs to be modified to also exclude duplicate observations. Otherwise the problem the leave-one-out approach aims to avoid would not be avoided.

[57] By restricting the range of the bandwidth to be above a certain smallest value, we may not need to use the delete-one-method, which is computationally costly.

[58] See Fan and Gijbels (1995), Table 1 for the adjustment factors.

To understand the objective function we examine each term of the expression separately. Note that since $\hat{\beta} = (X'WX)^{-1}X'Wy$, the denominator, ignoring the $1 + (p+1)V$ term can be written as

$$y'(I - WX(X'WX)^{-1}X')W(I - X(X'WX)^{-1}X'W)y = y'(W - WX(X'WX)^{-1}X'W)y$$

Recall that $y = X\beta_0 + r + \epsilon$. Since the term related to $X\beta_0$ vanishes and ignoring the cross terms of $r$ and $\epsilon$ as they are smaller order, the leading two terms are

$$r'(W - WX(X'WX)^{-1}X'W)r \qquad \text{and} \qquad \epsilon'(W - WX(X'WX)^{-1}X'W)\epsilon.$$

For the local linear case the first term divided by the trace in the denominator of the definition of RSC converges to $[m^{(p+1)}(x_0)]^2 h^{2(p+1)}$ times a constant (say $C$) and the second term divided by the same trace converges to $\sigma^2(x_0)$.

As we saw, the $\hat{V}$ is approximately constant (say $C'$) divided by $(nh)f(x_0)$. Thus

$$\begin{aligned} & (C[m^{(p+1)}(x_0)]^2 h^{2(p+1)} + \sigma^2(x_0)) \left(1 + \frac{(1+p)C'}{nhf(x_0)}\right) \\ = \quad & \sigma^2(x_0) + C[m^{(p+1)}(x_0)]^2 h^{2(p+1)} + \frac{(1+p)C'\sigma^2(x_0)}{nhf(x_0)} + o(h^4 + 1/(nh)). \end{aligned}$$

The minimizer is proportional to the optimum bandwidth by a known factor as desired.

They advocate using the integrated version of the $RSC(x_0)$ over an interval to select a global bandwidth. In fact even for the local bandwidth, they advocate using locally integrated version of $RSC(x_0)$ objective function. Clearly the adjustment term does not change.

### 6.2.3 Two step methods

The methods discussed above do not require that an initial bandwidth be specified. As discussed, other methods proposed in the literature attempt to improve on these procedures by using the first stage estimates as inputs into a second stage.

The methods estimate the bias and variance terms. Note that to estimate the bias term, which involves the $p+1$st order derivative, we need to assume that the function is smoother than required for estimating the regression function itself. For example, when a twice continuously differentiable function is being estimated by the local linear regression estimator, the bias term depends on the second order derivative. To compute the optimum bandwidth for estimating the second order derivative, the underlying function is assumed to be at least $p+1$-times continuously differentiable, or in this case at least three times continuously differentiable. But for a function with that degree of smoothness, the local linear estimator does not achieve the optimum rate of convergence. Thus the bandwidth computed does not have an overall optimality property. In this case, we will be estimating the optimal bandwidth optimum given that the local linear estimator is used in estimation.

**Fan and Gijbels's finite sample method** Fan and Gijbels (1995) proposes to use their RSC bandwidths to construct a "refined" bandwidth. Instead of using the asymptotic formula, they propose to use the finite sample counter-part discussed in section 4. The bias is

$$(X'WX)^{-1}X'Wr$$

and the variance is

$$(X'WX)^{-1}X'WXW^2X(X'WX)^{-1}\sigma^2(x_0).$$

Because $r = m - X\beta_0$, the bias is not known. But it can be approximated by

$$(X'WX)^{-1}X'W\tau,$$

where the $\tau$ is a vector of length $n$ with the $i$th element to be

$$(x_i - x_0)^{(p+1)}/(p+1)!\beta^{(p+1)} + \cdots + (x_i - x_0)^{(p+a)}/(p+a)!\beta^{(p+a)}.$$

They advocate using $a = 2$ or 3 as a target bias expression. Writing $S_n = X'WX$ and $S_{n,s,t} = [X^{(s)\prime}/s!]W[X^{(t)}/t!$, we can write $(X'WX)^{-1}X'W\tau$ as

$$S_n^{-1}\begin{pmatrix} S_{n,0,p+1}\beta^{(p+1)} + \cdots + S_{n,0,p+a}\beta^{(p+a)} \\ \vdots \\ S_{n,p,p+1}\beta^{(p+1)} + \cdots + S_{n,p,p+a}\beta^{(p+a)} \end{pmatrix}.$$

The unknown terms $\beta^{(p+1)}, \ldots, \beta^{(p+a)}$ can be estimated using the local polynomial estimator of degree $p+a$. For this step, RSC method is being advocated. They also note that the finite sample performance was better when the terms corresponding to $S_{n,s,p+t}$ where $s+t > p+a$ are set to 0. These terms are smaller order terms than the target bias expression.

The conditional variance is estimated by the same expression corresponding to the first expression of the RSC objective function:

$$\frac{\sum_{i=1}^n (y_i - \hat{y_{p+a_i}})^2 K_h(x_i - x_0)}{trace\{W - WX_{p+a}(X'_{p+a}WX_{p+a})^{-1}X'_{p+a}W\}},$$

where $X_{p+a}$ corresponds to regressors of the $p+a$th degree local polynomial estimator and $\hat{y}_{p+a} = X_{p+a}\hat{\beta}_{p+a}$. Because the higher degree local polynomial estimator is used, the bias contribution is of order $h^{p+a+1}$ and thus can be ignored. The estimated bias and variance terms are then used to form the estimated mean square error used to choose the bandwidth.

**Other methods** Ruppert (1997) proposes instead to estimate the bias term by running the OLS regression of

$$\hat{m}_h^{(j)}(x_0) = c_0(x_0) + c_{p+1-|j|}(x_0)h^{p+1-|j|} + \cdots + c_{p+a-|j|}h^{p+a-|j|}$$

using different $h$ values as regressors and the corresponding $\hat{m}_h^{(j)}(x_0)$ values as the dependent variable. This formulation is motivated by the asymptotic bias calculation. The estimated terms

after the first one are used to estimate the bias. Rupert (1997) replaces Fan and Gijbels's bias estimator in the finite sample method with this estimator in approximating the asymptotic mean square error.

Note that the point-wise optimum bandwidth becomes infinite when $m^{(p+1)}(x_0) = 0$, even though this may hold only at $x_0$, so that the $p$th order approximation does not hold globally. This is a limitation of considering the optimum bandwidth point-wise. Fan, Hall, Martin, and Patil (1996) considers modeling the local bandwidth globally using the LSCV objective function. While they describe the method for the kernel regression estimator, the method is clearly applicable to local polynomial estimator. Their objective function is

$$\sum_{i=1}^{n}[y_i - m_i(x_i, h(x_i))]^2$$

where in their case

$$m_i(x, h(x)) = \frac{\sum_{j\neq i} y_j K((x - x_j)/h(x))}{\sum_{j\neq i} K((x - x_j)/h(x))}$$

and $h(x) = h_0 g(x)$ for some $g(x)$ to be in a prespecified class of functions. This approach avoids approaching the problem point-wise and also makes the global LSCV method a local method.

Doksum, Petersen, and Samarov (2000) argues that the asymptotic formula used to construct approximation to the asymptotic mean square error is valid only for small bandwidths. They show that for larger bandwidths, a finite differencing gives a better approximation.

Prewitt and Lohr (2006) develops a way to eliminate a too small bandwidth from being considered, using the ratio of the largest to the smallest eigenvalues of the matrix $X'WX/(nh^d)$, drawing an analogy between local polynomial methods and regular linear regression analysis. This approach could be applied to prevent to guard against too small a bandwidth being chosen by any of the above methods.

## 6.3 How to choose smoothing parameters in semiparametric models

Relatively few papers have examined the problem of how to choose smoothing parameters in implementing semiparametric models.[59] Here we provide a brief account of some of the developments in this area of research.

### 6.3.1 Optimal bandwidth choice in average derivative estimation

The problem of choosing the optimal bandwidth in average derivative estimation is considered in Powell and Stoker (1996), Härdle, Hart, Marron, and Tsybakov (1992), Härdle and Tsybakov (1993), and Nishiyama and Robinson (2000, 2001). Härdle et. al. (1992) study bandwidth choice for the estimation of univariate unweighted average derivatives. Härdle and Tsybakov (1993) and

[59] See Härdle, Hall, and Ichimura (1993), Härdle, Hart, Marron, and Tsybakov (1992), Härdle and Tsybakov (1993), Hall and Horowitz (1990), Hall and Marron (1987), Horowitz (1992), Ichimura and Linton (2005), Linton (1995, 1996), Nishiyama and Robinson (2000, 2001), Powell and Stoker (1996), Stoker (1996), and Robinson (1991).

Powell and Stoker (1996) study a variety of weighted average derivative estimators for higher dimensions under a variety of weighting schemes using asymptotic mean square error as a criterion. Nishiyama and Robinson (2000, 2001) proposes to use an approximation to the asymptotic normality as a criterion.

Here, we describe the approach taken in Powell and Stoker (1996) as a prototype analysis of an optimal plug-in bandwidth selection to minimize the leading terms of the asymptotic mean-squared error of a semiparametric estimators. Recall from section 5 of the chapter that an indirect density weighted average derivative estimator takes the form

$$\hat{\delta}_{WIAD} = -\frac{2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x)}{\partial x}y_{i.}$$

As shown in Powell and Stoker (1989), this estimator can alternatively be written as

$$\begin{pmatrix} N \\ 2 \end{pmatrix}^{-1}\sum_{i<j} p(z_i, z_j, h),$$

where $z_i = (x_i, y_i)$, for a $d$-dimensional vector $x_i$ and a scalar $y_i$, $p(z_i, z_j, h) = -h^{-d-1}K'\left(\frac{x_i - x_k}{h}\right)(y_i - y_j)$ and $K(\cdot)$ is a kernel function satisfying $K(u) = K(-u)$, $K'(\cdot)$ denote the $d$-dimensional vector of partial derivatives of $K(\cdot)$, $\int K(u)du = 1, \int K(u)u^l du = 0$ for $l < \alpha, \int K(u)u^\alpha du \neq 0$ (for commonly used kernel functions, $\alpha = 2$). A requirement for asymptotic normality of the estimator is $2\alpha > d + 2$,. Define

$$\hat{r}(z_i, h) = \frac{1}{N-1}\sum_{j\neq i} p(z_i, z_j, h),$$

$r_0(z) = \lim_{h\to 0} E[\hat{r}(z, h)]$. Note that

$$E[\hat{r}(z, h)] - r_0(z) = s(z)h^\alpha + o(h^\alpha)$$

for some $s(z)$ under the assumption on the kernel function, among others and

$$E(\|p(z, z_j, h)\|^2) = q(z)h^{-\gamma} + o(h^{-\gamma}).$$

For the average derivative case, $\gamma = d + 2$ and

$$q(z) = [(y - E(y|x))^2 + Var(y|x)]f(x)\sum_{j=1}^{k}\int K_j^2(s)ds$$

where $K_j$ denote the $j$th element of $K'$. As shown in Powell and Stoker (1996), the leading terms of the mean-squared error of $\hat{\delta}_{WIAD}$ are

$$[E(s(z_i))]^2 h^{2\alpha} + 4n^{-1}Var[r_0(z_i)] + 2n^{-1}C_0 h^\alpha + 2n^{-2}E[q(z_i)]h^{-(d+2)}$$
$$+o(h^{2\alpha}) + o(h^\alpha/n) + o(1/(n^2 h^{d+2})).$$

Minimizing over $h$ (noting that the variance term does not depend on the bandwidth) and keeping only the leading terms gives the optimal plug-in bandwidth selector[60]:

$$h_{opt} = \left[\frac{(d+2)E[q(z_i)]}{\alpha[E(s(x))]^2}\right]^{1/(2\alpha+d+2)} \left[\frac{1}{n}\right]^{2/(2\alpha+d+2)}.$$

The method calls for using a high order kernel so that $2\alpha > d+2$. However, a simulation study conducted by Horowitz and Härdle (1996) found that using a second order kernel produced a more stable results.

Robinson (1995) showed that the normal approximation to the asymptotic distribution of the density weighted averaged derivative estimator could be worse than for the standard parametric case, depending on the bandwidth. In particular, he showed that, under some regularity conditions, the approximation error is of order

$$n^{-1/2} + n^{-1}h^{-d-2} + n^{1/2}h^{\alpha} + h^{M-1},$$

where $M$ denotes the order of differentiability of the conditional mean function of $y$ given $x$. Thus, the bandwidth required to make the order of approximation comparable to the parametric case of $n^{-1/2}$ is, for some $C \in (1, \infty)$ when $M - 1 \geq \alpha/2$,

$$(Cn^{1/(2d+4)})^{-1} \leq h \leq Cn^{-\alpha}.$$

The optimum bandwidth proposed by Powell and Stoker (1996) does not satisfy the second inequality, (the bias contribution to the normal approximation dominates) so using the bandwidth will make the normal approximation to be suboptimal. Nishiyama and Robinson (2000, 2001) derived the optimum bandwidth when approximation to the normality is the criterion.

Given that the normal approximation is worse than the parametric cases, Nishiyama and Robinson (2005) examines the bootstrap approximation and provides sufficient conditions under which the bootstrap approximates the asymptotic distribution to a higher order. While the work is carried out in detail for the particular case of the average derivative estimator, no doubt the technologies developed would be useful for investigating properties of other estimators.

### 6.3.2 Other works

Härdle, Hall, and Ichimura (1993) studies the semiparametric least squares estimation of the single index model and proposes to optimize over the bandwidth as well as the unknown coefficient. They propose a way of choosing the bandwidth that is asymptotically optimal for estimating the conditional mean function. It is not in general optimal for estimating the unknown coefficient, although the asymptotic distribution theory will still be valid with that choice of bandwidth.

Hall and Horowitz (1990), Horowitz (1992), Ichimura and Linton (2005) and Linton (1995) study optimum bandwidth selection for estimation of censored regression models, binary choice models, program evaluation models and the partially linear regression models, respectively. All

[60] See Proposition 4.1 in Powell and Stoker (1996).

these papers use the leading terms of the asymptotic mean square error terms as the criterion in choosing the optimum bandwidth.

Compared to the literature in the nonparametric estimation, the literature in selecting the smoothing parameter for estimators of semiparametric model parameters is sparse. Much more research needs to be done in this direction. Without specifying ways of choosing the bandwidth parameter, the estimators not well defined.

## 6.4 Trimming

### 6.4.1 What is trimming?

In the context of computing a statistics, trimming refers to a practice to systematically discarding the contribution of estimated function values to the statistics when some properties hold at the points the function is being evaluated. Usually the term "trimming function" refers to an indicator function indicating which points to include, rather than which point to discard.

### 6.4.2 Three reasons for trimming

There are three reasons for trimming. First, a parameter studied may not make sense without trimming. Second, a statistic may not make sense without trimming, or third, the statistics may not have desirable properties asymptotically without trimming.

As an example for the first case, consider estimating the conditional mean function $m(x)$. Recall that this function is defined at any point in the support, $S$, of the conditioning random vector so more precisely we should write it as $m(x) \cdot 1(x \in S)$. If we are to estimate the conditional mean function at observed data points, the indicator function is always 1, so that we can ignore the trimming function, but otherwise, the definition of the parameter calls for it. Parameters examined in section 2 provide some other examples where trimming is needed. We saw there that the identifiable parameter under the matching assumption needed to satisfy the common support condition. Therefore, the definition of the average treatment on treated parameter, for example, incorporated the trimming function as in

$$\frac{E\{(Y_1 - Y_0) 1(X \in S) | D = 1\}}{E\{1(X \in S) | D = 1\}}.$$

where $S$ denotes the common support of regressors $X$.

As an example for the second reason for trimming, recall the definition of the kernel regression estimator using the Epanechnikov kernel with optimal bandwidth. With this estimator, there is a positive probability that the denominator is zero, so that the estimator is not necessarily well defined. The estimator is well defined only if there is a data point in the appropriate neighborhood.

There are at least two distinct technical reasons for trimming in order to establish desirable properties of the statistics under consideration. First, to secure local data and second, to avoid the boundary value problem. Consider the same estimator and assume we want to show that the estimator converges with a rate uniformly over a given domain. Then at any point over the domain,

the density of the conditioning vector needs to be bounded away from 0 by the amount dictated by the convergence rate of the estimator we wish to obtain. For one thing, if the density is too low, then we cannot hope to obtain the local observation comparable to other regions. From a theoretical point of view, we can assume that the density is bounded away from 0, but of course in application, the condition does not necessarily hold and hence we have to introduce trimming. We also need to ensure that the function is not evaluated at points too close to the boundary value.

The third case for trimming often arises in examining semiparametric estimators which use nonparametric estimators in their construction. In establishing asymptotic properties of the semiparametric estimator, a uniform convergence rate of the nonparametric estimator is used.

The need for trimming for all cases is uncontroversial. But we have heard some claims for ignoring trimming "in practice" as "it does not matter very much." While we also think it would be nice if it were true, we emphasize that at this point we know of no systematic empirical or theoretical study which substantiates the claim.

### 6.4.3 How trimming is done

Sometimes trimming is specified using a priori chosen set over which some desirable properties hold, such as density be bounded away from zero. There is no provision for how we should choose such a set given a finite amount of observations.

Bickel (1982) introduced the trimming function that does not depend on a priori knowledge of the shape of the support in the context of adaptive estimation. In carrying out trimming of certain data points with low density, he used estimated density. A deterministic sequence which converges to zero is used to decide which points correspond to "low" density points.

While theoretically this procedure can be carried out without knowing anything about the density, in finite sample, the procedure might inadvertently trim out a high fraction of observations. To avoid this problem, Heckman, Ichimura, Smith, and Todd (1998) proposed defining a trimming function using a quantile of the estimated density.

An additional complication arises for the case of the index model. Consider for concreteness the linear index model. In this case we need to find low points of density corresponding to any index defined by a linear combination of the regressors. It may seem enough to trim observations based on the joint density of the regressors but that is not the case. To see this consider two independent regressors both distributed uniformly over unit intervals. On the support the density of the regressors are bounded away from zero. But any linear combination of the two regressors will not be bounded away from zero at the minimum and the maximum points when indeed two regressors are involved in the linear combination. This is because the density is low when the length of the line segment that leads to the same value for the linear combination is short. At the points that have the minimum and the maximum values of the linear combination, the corresponding length of the line segments are zero.

In addition to the density being bounded away from zero, trimming in this case needs to guarantee that the points of estimation are interior points of the support, so that the length of

the line segments will be away from zero. Clearly, one can presume a priori knowledge about the support and define trimming function using the knowledge.

One way to define the trimming function empirically is to use the estimated density as previously described. In this case, we need to keep points only if the density values are above certain value and that in a neighborhood there is no points with density values below the prespecified value. The prespecified value can be defined using the quantiles of the estimated density as in the previous case.

Given that the index models are used when we do not have enough observations to use fully nonparametric models, the above trimming approach may be unattractive because it uses fully nonparametric density estimator. An alternative approach which only involves one dimensional density estimation is to search over the lowest one dimensional density estimate at each point. We only keep a point if the point does not correspond to a low density point for any linear combination of regressors. Clearly this approach is computationally intensive. A practical alternative to this approach may be to try out the density estimation of the index defined by the bases of the space of the coefficients and keep all points which are above the prespecified low density values.

# 7 Asymptotic distribution of semiparametric estimators

In this section, we gather some basic asymptotic results that are useful in deriving the asymptotic distribution for semiparametric estimators. The structure underlying the asymptotic distribution of semiparametric estimators has been clarified greatly through the works of Aït-Sahalia (1992), Andrews (1994), Newey (1994), Sherman (1994), Ai and Chen (2003), Chen, Linton, and Keilegom (2003), and Ichimura and Lee (2006). Using these results, the asymptotic variance-covariance matrix of most of the semiparametric estimators can be easily computed. Chen (in this handbook) describes this development for the semiparametric GMM estimators, so we will describe the developments with regard to semiparametric M-estimators, summarizing the results obtained by Ichimura and Lee (2006).

Let $Z$ denote the random variable of dimension $\mathbf{R}^{d_z}$ with the support $\mathcal{S}$. Also, let $\theta_0$ be an element of a finite-dimensional parameter space $\Theta \subset \mathbf{R}^{d_\theta}$ that minimizes $E[m(Z, \theta, f_0(\cdot, \theta)]$, for an unknown, $d_f$-vector-valued function $f_0 \in \mathcal{F}$, where $\mathcal{F}$ is a Banach space of $d_f$-vector-valued function of $Z$ on the domain $\mathcal{U}$ with the supremum norm. We assume that for each $\theta \in \Theta$, $f(\cdot, \theta) \in \mathcal{F}$. Note that function $f(\cdot, \theta)$ is a function of $Z$, but the $\cdot$ argument may be different from $Z$. This is the reason for introducing the notation of $\mathcal{U}$. We will discuss this again with an example.

We denote the Euclidean norm by $\|\cdot\|$, $\|f\|_{\mathcal{F}} = \sup_{\theta \in \Theta} \sup_{z \in \mathcal{S}} \|f(z, \theta)\|$ for any $f(\cdot, \theta) \in \mathcal{F}$, and $\|(\theta, f)\|_{\Theta \times \mathcal{F}} = \|\theta\| + \|f\|_{\mathcal{F}}$. When $f$ depends on $\theta$, $\|f(\cdot, \theta)\|_\infty$ is understood to be the supremum norm with $\theta$ fixed.

Let the function $m(Z, \theta, f)$ denote a known, real-valued function that may depend on the data $Z$ and parameter $\theta$ directly and also possibly indirectly through $f$, for example, if $f$ depends on $\theta$. The function $m$ can depend on $f$ only via a particular value $Z$, in which case $m$ is a regular function with respect to $f(Z, \theta)$, or it can depend on an entire function $f(\cdot, \theta)$, in which case $m$ is

a functional with respect to $f$ for each $Z$ and $\theta$. In any case, we assume that $m(z, \theta, f)$ is defined over $\mathbf{S} \times \Theta \times \mathcal{F}$.

Assume that for each $\theta$, a nonparametric estimator $\hat{f}_n(\cdot, \theta)$ of $f_0(\cdot, \theta)$ is available. We define an M-estimator of $\theta_0$ as the minimizer of

$$\hat{S}_n(\theta) \equiv n^{-1} \sum_{i=1}^{n} m(Z_i, \theta, \hat{f}_n(\cdot, \theta)),$$

under the assumption that the observed data $\{Z_i : i = 1, \ldots, n\}$ are a random sample of $Z$. Let $\hat{\theta}_n$ denote the resulting estimator of $\theta_0$.

Examples that fit within this framework include the estimators studied by Robinson (1988), Powell, Stock, and Stoker (1989), Ichimura (1993), and Klein and Spady (1993) among many others, but the framework is also general enough to include the single-index quantile regression estimator, as discussed in Ichimura and Lee (2006).

Here, we will use the semiparametric least squares (SLS) estimator of Ichimura (1993) as a working example to illustrate how the assumptions and theorems can be applied to derive the distribution theory. In the SLS case, $Z = (Y, X)$ and

$$m(Z, \theta, f(\cdot, \theta)) = (Y - f(X'\theta, \theta))^2 1(X \in \mathcal{X})/2.$$

We assume $E(Y|X) = \phi(X'\theta_0)$. In this example, $\theta$ enters $m$ only via $f$ and $m$ depends on $f$ only via its value at $X$. Note that the $\cdot$ argument in this case is one-dimensional, although $X$ is in general a vector. In this example, $\mathcal{U}$ is the support of $X'\theta$.

To state the assumptions and results of Ichimura and Lee (2006), we need to introduce some more notation. For any $\delta_1 > 0$ and $\delta_2 > 0$, define $\Theta_{\delta_1} = \{\theta \in \Theta : \|\theta - \theta_0\| < \delta_0\}$ and $\mathcal{F}_{\delta_1, \delta_2} = \{f \in \mathcal{F} : \sup_{\theta \in \Theta_{\delta_1}} \|f(\cdot, \theta) - f_0(\cdot, \theta_0)\|_\infty < \delta_2\}$.

## 7.1 Assumptions

The function $m$ is not required to be differentiable, but is assumed to satisfy the following conditions.

**Assumption 7.1.** *For any $(\theta_1, f_1)$ and $(\theta_2, f_2)$ in $\Theta_{\delta_1} \times \mathcal{F}_{\delta_1, \delta_2}$, there exist linear operators $\Delta_1(z) \cdot (\theta_1 - \theta_2)$ and $\Delta_2(z, f_1(\cdot) - f_2(\cdot))$ and a function $\dot{m}(z, \delta_1, \delta_2)$ satisfying*

$$(a) \quad |m(z, \theta_1, f_1(\cdot)) - m(z, \theta_2, f_2(\cdot)) - \Delta_1(z)(\theta_1 - \theta_2) - \Delta_2(z, f_1(\cdot) - f_2(\cdot))| \\ \leq [\|\theta_1 - \theta_2\| + \|f_1(\cdot) - f_2(\cdot)\|_\infty] \, \dot{m}(z, \delta_1, \delta_2),$$

*and*

$$(b) \quad E[\dot{m}^2(Z, \delta_1, \delta_2)]^{1/2} \leq C \, (\delta_1^{\alpha_1} + \delta_2^{\alpha_2})$$

*for some constants $C < \infty$, $\alpha_1 > 0$, and $\alpha_2 > 0$.*[61]

[61] Here, $\Delta_1$, $\Delta_2$, and $\dot{m}$ may depend on $(\theta_2, f_2(\cdot))$. However, we suppress the dependence on $(\theta_2, f_2(\cdot))$ for the sake of simplicity in notation.

Ichimura and Lee (2006) verifies the condition for the single-index semiparametric quantile regression estimator. The condition is easier to verify for differentiable cases. Note that $\Delta_1(z)$ and $\Delta_2(z)$ corresponds to the "derivatives" of $m$ with respect to $\theta$ and $f$, respectively. Because $m$ is generally a functional in $f$, the first "derivative" with respect to $f$ is a linear operator, whereas the "derivative" with respect to $\theta$ can be expressed as a finite dimensional vector.

For SLS, the function $m$ depends on $f$ only via $f(X'\theta, \theta)$, so that both "derivatives" correspond to a finite dimensional vector. One can guess the forms of $\Delta_1(z)$ and $\Delta_2(z)$ by taking derivatives and evaluating them at the true values. Because the function $m$ does not depend on $\theta$ directly, $\Delta_1(z) = 0$ and $\Delta_2(z) = -(Y - f_0(X'\theta_0, \theta_0))$. One can verify that with these functions, the assumption holds with $\dot{m}(z, \delta_1, \delta_2) = \delta_2$, so that $\alpha_2 = 1$.

While the function $m$ is allowed to be non-differentiable, its expected value is assumed to be differentiable with respect to $\theta$ and $f$ (as assumed in Pollard (1985)). Denote the expected value by $m^*(\theta, f) = E[m(Z, \theta, f)]$.

**Assumption 7.2.** *$m^*(\theta, f)$ is twice continuously Fréchet differentiable in an open, convex neighborhood of $(\theta_0, f_0(\cdot, \theta_0))$ with respect to a norm $\|(\theta, f)\|_{\Theta \times \mathcal{F}}$.*

For the SLS example, the Fréchet derivative with respect to $\theta$ is zero and hence the cross derivative is also. The Freć het derivative with respect to $f$ is $D_f m^*(\theta, f)(h) = -E[(Y - f(X))h(X)1\{X \in \mathcal{X}\}]$ and the second Freć het derivative with respect to $f$ is $D_{f,f} m^*(\theta, f)(h_1, h_2) = E[h_1(X)h_2(X)1\{X \in \mathcal{X}\}]$.

The class of functions $\mathcal{F}$ needs to be restricted as well. To characterize the nature of the restriction, we first introduce a few additional notations. Let $\underline{\alpha}$ denote the greatest integer strictly smaller than $\alpha$, $j = (j_1, \ldots, j_d)$, and let

$$\|g\|_\alpha = \max_{|j| \leq \underline{\alpha}} \sup_x |D^j g(x)| + \max_{|j| = \underline{\alpha}} \sup_{x,y} \frac{|D^j g(x) - D^j g(y)|}{\|x - y\|^{\alpha - \underline{\alpha}}},$$

where the suprema are taken over all $x, y$ in the interior of $\mathcal{U}$ with $x \neq y$. Then $\mathcal{C}_M^\alpha(\mathcal{U})$ is defined as the set of all continuous functions $g : \mathcal{U} \subset \mathbf{R}^d \mapsto \mathbf{R}$ with $\|g\|_\alpha \leq M$.

**Assumption 7.3.** *$f_0(\cdot, \theta)$ is twice continuously differentiable on $\Theta_{\delta_1}$ with bounded derivatives on $\mathcal{U}$ and $\mathcal{F}$ is a subset of $C_M^\alpha(\mathcal{U})$, where $\mathcal{U}$ is a finite union of bounded convex subsets of $\mathbf{R}^{d_u}$ with non-empty interior where $\alpha > d_u/2$.*

For SLS, $f_0(u, \theta) = E(Y | X'\theta = u)$. The assumption requires that $f_0$ is twice continuously differentiable with respect to $\theta$. In the SLS case, $d_u = 1$, so we do not require differentiability with respect to $u$.

The next set of assumptions are restrictions on the estimator of $f_0$.

**Assumption 7.4.** *(a) For any $\theta \in \Theta_{\delta_1}$, $\hat{f}_n(\cdot, \theta) \in \mathcal{C}_M^\alpha(\mathcal{X})$ with probability approaching one.*

*(b) $\sup_{\theta \in \Theta_{\delta_1}} \left\| \hat{f}_n(\cdot, \theta) - f_0(\cdot, \theta) \right\|_\infty = O_p(\tilde{\delta}_2)$ for $\tilde{\delta}_2$ satisfying $n^{1/2} \tilde{\delta}_2^{1+\alpha_2} \to 0$.*

*(c) For any $\varepsilon > 0$ and $\delta > 0$, independent of $\theta$, there exists $n_0$ such that for all $n \geq n_0$, the following holds:*

$$Pr\left\{\left\|[\hat{f}_n(\cdot,\theta) - \hat{f}_n(\cdot,\theta_0)] - [f_0(\cdot,\theta) - f_0(\cdot,\theta_0)]\right\|_\infty \leq \delta \left\|\theta - \theta_0\right\|\right\} \geq 1 - \varepsilon.$$

Condition (b) requires that $\hat{f}_n(\cdot,\theta)$ converge uniformly in probability. If $\alpha_2 = 1$ (smooth $m$), then $\tilde{\delta}_2 = o(n^{-1/4})$; when $\alpha_2 = 0.5$ (non-smooth $m$), then $\tilde{\delta}_2 = o(n^{-1/3})$. In general, $\hat{f}_n(\cdot,\theta)$ needs to converge at a faster rate when $m$ is less smooth.

Condition (c) is satisfied if $\hat{f}_n(\cdot,\theta)$ is differentiable with respect to $\theta$ and the derivative converges uniformly to $\partial f_0(\cdot,\theta)/\partial\theta$ over both arguments. This is shown by Ichimura (1993) for the SLS example. Ichimura and Lee's (1991) results in the appendix is useful in proving analogous results in other kernel based estimators.

The next set of assumptions are joint conditions on the second Fréchet derivative of $m^*(\theta, f)$ with respect to $f$ and the estimator of $f_0$. Write $D_{f,f}m^*(\theta, f) = \int w(\theta, f(\cdot,\theta))h_1(\cdot)h_2(\cdot)dP$, where $P$ is the measure of $Z$.

**Assumption 7.5.** *One of the following three conditions holds:*

*(i) $w(\theta, f(\cdot,\theta))$ does not depend on $\theta$ or $f(\cdot,\theta)$ and is bounded.*

*(ii) $\|w(\theta, f(\cdot,\theta)) - w(\theta_0, f_0(\cdot,\theta_0))\| \leq C_w \|\theta - \theta_0\|$ for some finite constant $C_w$ and $\sup_{\theta\in\Theta_{\delta_1}} \left\|\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)\right\|_\infty = o_p(n^{-1/4})$.*

*(iii) $\|w(\theta, f(\cdot,\theta)) - w(\theta_0, f_0(\cdot,\theta_0))\| \leq C_w [\|\theta - \theta_0\| + \|f(\cdot,\theta) - f_0(\cdot,\theta_0)\|_\infty]$ for some finite constant $C_w$*

We saw that for SLS, case $(i)$ applies.

The following assumption is made first to accommodate cases where estimation of $f_0$ has an effect on the asymptotic distribution of the estimator of $\theta_0$. Later, sufficient conditions for this higher level assumption are discussed.

**Assumption 7.6.** *(a) As a function of $\theta$, $D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]$ is twice continuously differentiable on $\Theta_{\delta_1}$ with probability approaching one.*

*(b) There exists a $d_\theta$-row-vector-valued $\Gamma_1(z)$ such that $E[\Gamma_1(Z)] = 0$, $E[\Gamma_1(Z)\Gamma_1^T(Z)] < \infty$ and nonsingular,*

$$\frac{d}{d\theta^T}\left(D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]\right)\Big|_{\theta=\theta_0} = n^{-1}\sum_{i=1}^n \Gamma_1(Z_i) + o_p(n^{-1/2}). \tag{7.1}$$

In (b), $\Gamma_1(z)$ captures the effects of the first stage estimation of $f_0$. Two cases where the derivative is easy to compute are: when $f_0$ does not depend on $\theta$ and when $D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]$ is identically zero. For SLS estimator, $D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]$ is identically zero so that there is no first order effect of estimating $f_0$.

The following proposition proved in Ichimura and Lee (2006) provides a set of sufficient conditions for computing the adjustment term that appears in Assumption 7.6.

**Proposition 7.1.** *Assume that*

*(a)*

$$D_f m^*(\theta, f_0(\cdot,\theta))[h(\cdot)] = \int h(\cdot) g(\cdot,\theta) dP, \tag{7.2}$$

*(b) $g(\cdot,\theta)$ is twice continuously differentiable with respect to $\theta$ with probability one,*

*(c) $\hat{f}_n(\cdot,\theta)$ has an asymptotic linear form: for any $\theta \in \Theta_{\delta_1}$,*

$$\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta) = n^{-1}\sum_{j=1}^{n} \varphi_{nj}(\cdot,\theta) + b_n(\cdot,\theta) + R_n(\cdot,\theta), \tag{7.3}$$

*where $\varphi_{nj}(\cdot,\theta)$ is a stochastic term that has expectation zero (with respect to the $j$-th observation), $b_n(\cdot,\theta)$ is a bias term satisfying $\sup_{z,\theta} \|b_n(z,\theta)\| = o(n^{-1/2})$ , and $R_n(\cdot,\theta)$ is a remainder term satisfying $\sup_{z,\theta} \|R_n(z,\theta)\| = o_p(n^{-1/2})$.*

*(d) $\hat{f}_n(\cdot,\theta)$ is twice continuously differentiable with respect to $\theta$ with probability approaching one and $\partial \hat{f}_n(\cdot,\theta)\partial\theta$ also has an asymptotic linear form:*

$$\frac{\partial \hat{f}_n(\cdot,\theta)}{\partial\theta} - \frac{\partial f_0(\cdot,\theta)}{\partial\theta} = n^{-1}\sum_{j=1}^{n} \tilde{\varphi}_{nj}(\cdot,\theta) + o_p(n^{-1/2}), \tag{7.4}$$

*uniformly over $(z,\theta)$, where $\tilde{\varphi}_{nj}(\cdot,\theta)$ is a stochastic term that has expectation zero (with respect to the $j$-th observation), and*

*(e) there exists a $d_\theta$-row-vector-valued $\Gamma_1(z)$ such that $E[\Gamma_1(Z)] = 0$ and*

$$\max_{1\le i\le n} \|\Gamma_{n1}(Z_i) - \Gamma_1(Z_i)\| = o_p\left(n^{-1/2}\right),$$

*where*

$$\Gamma_{n1}(Z_i) = \int \tilde{\varphi}_{ni}(\cdot,\theta_0) g(\cdot,\theta_0)\, dP + \int \varphi_{ni}(\cdot,\theta_0) \frac{\partial g(\cdot,\theta_0)}{\partial\theta}\, dP. \tag{7.5}$$

*Then Assumption 7.6 is satisfied.*

## 7.2 Main results on asymptotic distribution

First some notation. Let $\Delta_{10}(z)$ and $\Delta_{20}(z,h)$ denote $\Delta_1(z)$ and $\Delta_2(z,h)$ in Assumption 7.1 with $(\theta_1, f_1) = (\theta, f)$ and $(\theta_2, f_2) = (\theta_0, f_0(\cdot,\theta_0))$. Thus, $\Delta_{10}(z)(\theta-\theta_0) + \Delta_{20}(z, f(\cdot,\theta) - f_0(\cdot,\theta_0))$ is a linear approximation of $m(z,\theta,f(\cdot,\theta)) - m(z,\theta_0,f_0(\cdot,\theta_0))$. Define $\Delta^*_{20}[h] = E[\Delta_{20}(Z,h)]$ for fixed $h$. Also define a $d_\theta$-row-vector-valued function $\Gamma_0(z)$ such that

$$\Gamma_0(z) = \Delta_{10}(z) - E[\Delta_{10}(Z)] + \Delta_{20}\left[z, \frac{\partial f_0(\cdot,\theta_0)}{\partial\theta^T}\right] - \Delta^*_{20}\left[\frac{\partial f_0(\cdot,\theta_0)}{\partial\theta^T}\right] + \Gamma_1(z),$$

$\Omega_0 = E[\Gamma_0(Z)^T\Gamma_0(Z)]$, and

$$V_0 = \frac{d^2\, m^*(\theta, f_0(\cdot,\theta))}{d\theta\, d\theta^T}\Big|_{\theta=\theta_0}.$$

Notice that $V_0$ is the Hessian matrix of $m^*(\theta, f_0(\cdot,\theta))$ with respect to $\theta$, evaluated at $\theta = \theta_0$.

The following theorem gives the asymptotic distribution of $\hat{\theta}_n$.

**Theorem 7.2.** *Assume that $\theta_0$ is an interior point of $\Theta$, $\theta_0$ is a unique minimizer of $m^*(\theta, f_0(\cdot,\theta)$, and $\hat{\theta}_n$ is a consistent estimator of $\theta_0$. Moreover, assume that $\{Z_i : i = 1,\ldots,n\}$ are a random sample of $Z$. Let Assumptions 7.1-7.6 hold. Assume that there exists $C(z)$ satisfying $\|\Delta_{20}[z, h(\cdot,\theta)]\| \leq C(z)\,\|h(\cdot,\theta)\|_\infty$ for any $\theta$ and $\|C(Z)\|_{L^2(P)} < \infty$. Also, assume that $\Omega_0$ exists and $V_0$ is a positive definite matrix. Then*

$$n^{1/2}(\hat{\theta}_n - \theta_0) \to_d \mathbf{N}(0, V_0^{-1}\Omega_0 V_0^{-1}).$$

Let $\partial_1 m^*(\theta, f)$ denote a vector of the usual partial derivatives of $m^*(\theta, f)$ with respect to the first argument $\theta$. In this notation, $\partial_1 m^*(\theta, f(\cdot,\theta))$ denotes the partial derivative of $m^*(\theta, f)$ with respect to the first argument $\theta$, evaluated at $(\theta, f) = (\theta, f(\cdot,\theta))$. Similarly, let $\partial_1^2 m^*(\theta, f)$ denote the usual Hessian matrix of $m^*(\theta, f)$ with respect to $\theta$, holding $f$ constant. Using this notation, note that by the chain rule, the expression of $V_0$ can be written as[62]

$$\begin{aligned} V_0 &= \frac{d^2\, m^*(\theta, f_0(\cdot,\theta))}{d\theta\, d\theta^T}\Big|_{\theta=\theta_0} \\ &= \partial_1^2 m^*(\theta_0, f_0(\cdot,\theta_0)) + D_{ff} m^*(\theta_0, f_0(\cdot,\theta_0))\left[\frac{\partial f_0(\cdot,\theta_0)}{\partial\theta}, \frac{\partial f_0(\cdot,\theta_0)}{\partial\theta^T}\right] \\ &\quad + 2\left\{D_f\left[\partial_1 m^*(\theta_0, f_0(\cdot,\theta_0))^T\right]\left[\frac{\partial f_0(\cdot,\theta_0)}{\partial\theta}\right]\right\} + D_f m^*(\theta_0, f_0(\cdot,\theta_0))\left[\frac{\partial^2 f_0(\cdot,\theta_0)}{\partial\theta\partial\theta^T}\right]. \end{aligned}$$

For the SLS case, note that $\partial_1 m^*(\theta, f_0(\cdot,\theta)) = 0$ and that $D_f m^*(\theta_0, f_0(\cdot,\theta_0))(h) = 0$ so that

$$V_0 = D_{ff} m^*(\theta_0, f_0(\cdot,\theta_0))\left[\frac{\partial f_0(\cdot,\theta_0)}{\partial\theta}, \frac{\partial f_0(\cdot,\theta_0)}{\partial\theta^T}\right].$$

Because $\partial f_0(X'\theta,\theta)/\partial\theta$ evaluated at $\theta_0$ is $\phi'(X'\theta_0)[\tilde{X} - E(\tilde{X}|X'\theta_0)]$, where $\tilde{X}$ is all $X$ except for the variable whose associated coefficient is set to 1 (required for normalization). Thus,

$$V_0 = E\{[\phi'(X'\theta_0)]^2 1\{X \in \mathcal{X}\}[\tilde{X} - E(\tilde{X}|X'\theta_0)][\tilde{X} - E(\tilde{X}|X'\theta_0)]'\}$$

and

$$\Omega_0 = E\{[\phi'(X'\theta_0)]^2\epsilon^2 1\{X \in \mathcal{X}\}[\tilde{X} - E(\tilde{X}|X'\theta_0)][\tilde{X} - E(\tilde{X}|X'\theta_0)]'\}$$

where $\epsilon = Y - \phi(X'\theta_0)$.

As indicated above, when $f_0$ does not depend on $\theta$, one can easily compute the adjustment term $\Gamma_1(z)$. It turns out that one can relax the smoothness condition on function $m$ with respect to $f$ as well. The following assumptions are invoked in the theorem below, which gives the asymptotic distribution of $\hat{\theta}_n$ when the first-stage nonparametric estimator $\hat{f}_n(\cdot,\theta)$ does not depend on $\theta$.

[62] See Ichimura and Lee (2006) Appendix for the expression of $V_0$ when $d_f > 1$.

**Assumption 7.7.** *For any $(\theta_1, f)$ and $(\theta_2, f)$ in $\Theta_{\delta_1} \times \mathcal{F}_{\delta_2}$, there exist a $d_\theta$-row-vector-valued function $\Delta_1(z, \theta_2, f)$ and a function $\dot{m}(z, \delta_1)$ satisfying*

$$(a) \quad |m(z, \theta_1, f(\cdot)) - m(z, \theta_2, f(\cdot)) - \Delta_1(z, \theta_2, f)(\theta_1 - \theta_2)| \leq \|\theta_1 - \theta_2\| \, \dot{m}(z, \delta_1),$$

$$(b) \quad \|\dot{m}(Z, \delta_1)\|_{L^2(P)} \leq C\delta_1^{\alpha_1} \quad \text{for some constants } C < \infty \text{ and } \alpha_1 > 0,$$

*and*

$$(c) \quad \sup_{f \in \mathcal{F}_{\delta_2}} \left\| n^{-1} \sum_{i=1}^{n} \{\Delta_1(Z_i, \theta_0, f) - E[\Delta_1(Z, \theta_0, f)]\} - \{\Delta_1(Z_i, \theta_0, f_0) - E[\Delta_1(Z, \theta_0, f_0)]\} \right\| = o_p\left(n^{-1/2}\right) \quad \text{for any } \delta_2 \to 0.$$

**Assumption 7.8.** *(a) $f_0(\cdot)$ is an element of $\mathcal{C}_M^\alpha(\mathcal{X})$ for some $\alpha > d_1/2$, where $d_1$ is the dimension of the argument of $f_0(\cdot)$ and $\mathcal{X}$ is a finite union of bounded, convex subset of $\mathbf{R}^{d_1}$ with nonempty interior.*

*(b) $\hat{f}_n(\cdot) \in \mathcal{C}_M^\alpha(\mathcal{X})$ with probability approaching one.*

*(c) $\left\| \hat{f}_n(\cdot) - f_0(\cdot) \right\|_\infty = o_p(1)$.*

We next state the theorem providing the asymptotic distribution of $\hat{\theta}_n$ when the first-stage nonparametric estimator $\hat{f}_n(\cdot, \theta)$ does not depend on $\theta$.

**Theorem 6.** *Assume that $\theta_0$ is an interior point of $\Theta$, $\theta_0$ is a unique minimizer of $m^*(\theta, f_0(\cdot))$, and $\hat{\theta}_n$ is a consistent estimator of $\theta_0$. Moreover, assume that $\{Z_i : i = 1, \ldots, n\}$ are a random sample of $Z$. Let Assumptions 7.2, 7.5, 7.6, and 7.8 hold. Assume that either Assumption 7.1 or Assumption 7.7 holds. Also, assume that $\Omega_0 = E[\Gamma_0(Z)^T \Gamma_0(Z)^T]$ exists and $V_0$ is a positive definite matrix, where*

$$\Gamma_0(z) = \Delta_1(z, \theta_0, f_0) - E[\Delta_1(Z, \theta_0, f_0)] + \Gamma_1(z)$$

*and*

$$V_0 = \frac{\partial^2 m^*(\theta_0, f_0(\cdot))}{\partial \theta \partial \theta^T}.$$

*Then*

$$n^{1/2}(\hat{\theta}_n - \theta_0) \to_d \mathbf{N}(0, V_0^{-1} \Omega_0 V_0^{-1}).$$

# 8 Computation

Flexible modeling methods are computationally more demanding than traditional approaches. Among the various classes of flexible estimators, local methods tend to be the most computationally intensive, because they require solving separate problems at each point at which the density or function is evaluated. The computational burden is particularly great when cross-validation or bootstrap methods are used to select smoothing parameters and/or bootstrap methods are used to

evaluate the variation of the estimators. Because local density and regression estimators form the ingredients for many semiparametric procedures, the semiparametric methods can also be highly computationally intensive.

Fortunately, the processing speeds of today's computers make nonparametric and semiparametric modeling methods feasible in many applications with sample sizes of a few thousand, despite their additional computational burden. But when sample sizes get large, say on the order of 10,000 or more, then computing estimates and standard errors can become a major task, and time considerations may drive the choice of bandwidth selector and variance estimator. In such cases, one can take advantage of approximation methods that were suggested by Silverman (1982) and further studied in Fan and Marron (1994), Hall and Wand (1996), Jones and Lotwick (1984), Wand (1994) and others for speeding up computations in local regression and density estimation. These methods allow for great gains in speed and provide a way of controlling the accuracy of the approximation.

## 8.1 Description of an approximation method

The approximation method first grids the $x$-axis and computes the estimates only at grid points. Computation over grids is done efficiently using fast Fourier transformation. The method then interpolates to find function values between the grid-point estimates. The number of grid points, $M$, is chosen by the researcher. We first describe the most simple version of the binning method, in the context of obtaining a local linear regression estimate. Then we describe a fast Fourier implementation of the binning method, first for density estimation and then for local regression. The FF transformation effectively factors the data component and the bandwidth component in the frequency domain. This allows computation across different bandwidths to be done in a more efficient way, because the data component of the computation can be done only once and reused when computing the values at different bandwidths..

### 8.1.1 A simple binning estimator

Let $x_1 ... x_n$ denote $n$ actual data points at which we wish to evaluate the conditional mean function for the model

$$y = m(x) + \varepsilon.$$

The local linear regression estimator at a point $x$ is given by

$$\hat{E}_n(y_i|x) = \frac{\sum_{j=1}^n y_i K_j \sum_{k=1}^n K_k (x - x_k)^2 - \sum_{j=1}^n y_i K_j (x - x_j) \sum_{k=1}^n K_k (x - x_k)}{\sum_{j=1}^n K_j \sum_{k=1}^n K_k (x_i - x_k)^2 - \left[\sum_{j=1}^n K_j (x_i - x_j)\right]^2},$$

where $K_j = K((x - x_j)/h_n)$. Calculating the local regression estimator requires estimating terms of the form

$$\sum_{i=1}^n y_i (x_i - x)^l K((x_i - x_j)/h_n) \tag{8.1}$$

for $l = 0, 1, 2$ for the $n$ data points at which the function is evaluated.

The binning method reduces the computational burden of evaluating these kernel values by making a grid over the support of $x_1..x_n$ of equally spaced points, evaluating the function only at the grid points and interpolating to estimate the value of the function at other points. Denote the $N$ grid points by $z_1..z_N$. Binning can be implemented by first assigning each data point ($x_i$) and point of evaluation ($x$) to their nearest grid points ($z_{j'}$ and $z_j$, respectively) and approximating (8.1) by

$$\sum_{j'=1}^{N} \sum_{i \in I_{j'}} y_i (z_{j'} - z_j)^l K((z_{j'} - z_j)/h_n),$$

where $z_j$ are now the $N$ grid points of evaluation, $z_{j'}$ are the grid points to which the data points have been assigned and $I_{j'}$ are the set of indices that are binned into the $j'$th bin.

A consequence of choosing equally-spaced grid points is that the distance between $z_1$ and $z_3$ is the same as between $z_{N-2}$ and $z_N$, etc. Letting $\Delta$ denote the smallest distance between two grid points, we only need to evaluate the kernel at N values:

$$K(\Delta/h), K(2\Delta/h), K(3\Delta/h), ..., K(N\Delta/h)$$

which reduces the required number of evaluations of the kernel function to $N$ from $n^2$(the number required under a naive strategy of evaluating the kernel for each possible combination of data-points).

Fan and Marron (1994) introduce a modification of this simple binning idea, called linear binning. Linear binning assigns each data point or point of evaluation to multiple grid points, weighting each in proportion to their distance from the grid points. Fan and Marron (1994) show that for the linear binning estimator, the approximation error can be bounded by $\delta^4$, where $\delta$ is the bin or grid width. The FFT implementation described below uses the linear binning idea.

### 8.1.2 Fast Fourier transform (FFT) binning for density estimation

The binning method described above is adequate for many univariate estimation problems. But for multivariate as well as univariate estimation problems, a more efficient FFT implementation of binning is available. We describe how the FFT can be used to increase the efficiency of the binning estimator in the context of estimating a density, and then discuss how to apply it for local linear regression estimation. The FFT reduces the number of computations by taking advantage of periodicity in complex functions.

The Fourier transform of a density $g(t)$ is

$$\tilde{g}(s) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{ist} g(t) dt \tag{8.2}$$

Let $\hat{f}_n(x)$ be a standard kernel density estimator, $\hat{f}_n(x) = (nh_n)^{-1} \sum_{i=1}^{n} K((t - x_i)/h_n)$. The F-

transform of $\hat{f}_n(x)$ is

$$\begin{aligned}\widetilde{\hat{f}}_n(s) &= (2\pi)^{-1/2}(nh_n)^{-1}\sum_{i=1}^{n}\int e^{ist}K((t-x_i)/h_n)dt \\ &= (2\pi)^{-1/2}n^{-1}\sum_{i=1}^{n}\int e^{is(x_i+h_nu)}K(u)du \\ &= \{n^{-1}\sum_{i=1}^{n}e^{isx_i}\}\cdot\{(2\pi)^{-1/2}\int e^{ish_nu}K(u)du\}\end{aligned}$$

where the last two equalities follow after doing a change of variables $u = (t - x_j)/h_n$. The first term in brackets depends only on the data. The second is the F-transform of the $K(sh_n)$, which depends on the kernel and bandwidth choice. Under certain choices for $K$, there is an explicit solution for the second term. For example, if $K$ is normal it equals $(2\pi)^{-1/2}\exp\{(-s^2h_n^2)/2\}$.

The separation of (8.2) into two terms–one that depends solely on the data and one on the smoothing parameters–has a major computational advantage for algorithms, such as cross-validation, which require evaluating the function for several different bandwidth parameter, since the data component need to be evaluated only once.

To be able to quickly evaluate the data component, we wish to find an approximation to the first term, $(2\pi)^{-1/2}n^{-1}\sum_{i=1}^{n}e^{isx_i}$. Then $f_n(s)$ will be estimated by applying FF inversion to $\widetilde{\hat{f}}_n(s)$.

For large $n$, $(2\pi)^{-1/2}n^{-1}\sum_{i=1}^{n}e^{isx_i}$ converges to $(2\pi)^{-1/2}\int_{-\infty}^{\infty}e^{isx_i}g(x_i)dx_i$. Usually it is not possible to explicitly obtain the integral, but it can be approximated over a discrete set of points. Let $t_k = k\Delta$ denotes grid points over the interval $[-\infty, \infty]$, $\Delta$ a bin width, $k = -(N-1), ..., 0, ...(N-1)$, and let $g_k = g(t_k)$. The discrete FFT approximation to the integral evaluated at a point $s_n = n/(N\Delta)$, $n = -N/2, ..., N/2$ is

$$\tilde{g}(s_n) = (2\pi)^{-1/2}\sum_{k=-(N-1)}^{(N-1)}e^{is_nt_k}g_k\Delta,$$

The last expression can be written as

$$\tilde{g}(s_n) = (2\pi)^{-1/2}\Delta\sum_{k=-(N-1)}^{(N-1)}e^{\frac{ink\Delta}{N\Delta}}g_k.$$

We can use the fact that $e^{i\alpha}$ is a cyclical function to reduce the number of calculations to $N\log_2 N$. Writing the last expression as

$$(2\pi)^{-1/2}\Delta\left\{\sum_{k=-(N-1)}^{-1}e^{\frac{ink}{N}}g_k + g_0 + \sum_{k=1}^{(N-1)}e^{\frac{ink}{N}}g_k\right\}. \tag{8.3}$$

We now consider just the third term in brackets, since all the same considerations apply to the

first. We can write it as

$$\begin{aligned}\sum_{k=1}^{(N-1)} e^{\frac{ink}{N}} g_k &= \sum_{k=1}^{(\frac{N}{2}-1)} e^{\frac{in(2k+1)}{N}} g_{(2k+1)} + \sum_{k=1}^{(\frac{N}{2}-1)} e^{\frac{in(2k)}{N}} g_{(2k)} \\ &= e^{\frac{in}{N}} \sum_{k=1}^{(\frac{N}{2}-1)} e^{\frac{ink}{N/2}} g_{(2k+1)} + \sum_{k=1}^{(\frac{N}{2}-1)} e^{\frac{ink}{N/2}} g_{(2k)}.\end{aligned}$$

Repeat this process until the summation only includes one term:

$$\begin{aligned}&= e^{\frac{in}{N}} \sum_{k=1}^{(\frac{N}{4}-1)} e^{\frac{in(2k+1)}{N/2}} g_{(2(2k+1)+1)} + e^{\frac{in}{N}} \sum_{k=1}^{(\frac{N}{4}-1)} e^{\frac{in(2k)}{N/2}} g_{(2(2k)+1)} \\ &\quad + \sum_{k=1}^{(\frac{N}{4}-1)} e^{\frac{in(2k+1)}{N/2}} g_{(2(2k+1))} + \sum_{k=1}^{(\frac{N}{4}-1)} e^{\frac{in2k}{N/2}} g_{(2(2k))} \\ &= e^{\frac{in}{N}} e^{\frac{in}{N/2}} \sum_{k=1}^{(\frac{N}{4}-1)} e^{\frac{ink}{N/4}} g_{(4k+3)} + e^{\frac{in}{N}} \sum_{k=1}^{(\frac{N}{4}-1)} e^{\frac{ink}{N/4}} g_{(4k+1)} \\ &\quad + e^{\frac{in}{N/2}} \sum_{k=1}^{(\frac{N}{4}-1)} e^{\frac{ink}{N/4}} g_{(4k+2)} + \sum_{k=1}^{(\frac{N}{4}-1)} e^{\frac{ink}{N/4}} g_{(4k)} \\ &\quad \text{etc...}\end{aligned}$$

After making these substitutions, we get

$$= g_{(0)}(e^{in/N})^0 + g_{(1)}(e^{in/N})^1 + g_{(2)}(e^{in/N})^2 + \cdots + g_{(2^r)}(e^{in/N})^{2r},$$

where $2^r$ is the total number of grid points ($2^r = M$).

Consider the number of calculations required for each of these terms for $n = 0, ..., N/2$. (Negative terms are complex conjugates). Here

$$\begin{aligned} g_{(1)}(e^{in/N}) \ , \ n &= 0, .., N/2 \text{ requires N complex multiplications} \\ g_{(2)}(e^{in/N})^2 \ , \ n &= 0, .., N/2 \text{ requires N/2 complex multiplications} \\ g_{(3)}(e^{in/N})^3 \ , \ n &= 0, .., N/2 \text{ requires N/3 complex multiplications} \\ &\ . \\ &\ . \\ g_{(2^r)}(e^{in/N})^{2^r} \ , \ n &= 0, .., N/2 \text{ requires N/N complex multiplications.}\end{aligned}$$

Thus, we need no more than $N + N/2 + N/3 + ... + N/N = N \log_2 N$ complex multiplications.

**Making the grid** To implement the method described above, consider an interval $[a, b]$ in which the data lie. The FFT method imposes periodic boundary conditions, so the interval needs to be chosen large enough. For a normal kernel, it suffices to choose $a$ and $b$ that satisfy

$$\begin{aligned} a &< \min(x_j) - 3h_n \\ b &> \max(x_j) + 3h_n,\end{aligned}$$

where $h_n$ is the bandwidth.(Silverman, 1986) Also, let $M = 2^r$ for some integer $r$ denote the total number of grid points and let $\delta$ the bin width, $\delta = (b-a)/M$. The grid points are given by $t_k = a + k\delta$, for $k = 0, 1, ..., M-1$. If the data point falls onto the grid interval $[t_k, t_{k+1}]$, we assign a weight $\underline{\xi}_k = \delta^{-2}n^{-1}(t_{k+1} - x_j)$ to $t_k$ and a weight $\bar{\xi}_{k+1} = \delta^{-2}n^{-1}(x_j - t_k)$ to $t_{k+1}$. The weights over all the data points ($x_j$, $j = 1..n$) are accumulated at each grid point. Let

$$\begin{aligned}
\underline{\xi}_k &= \delta^{-2}n^{-1}\sum_{j=1}^{n}(t_{k+1} - x_j)1(x_j \in [t_k, t_{k+1}]) \\
\bar{\xi}_k &= \delta^{-2}n^{-1}\sum_{j=1}^{n}(x_j - t_{k-1})1(x_j \in [t_{k-1}, t_k])
\end{aligned}$$

$$\xi_k = \underline{\xi}_k + \bar{\xi}_k.$$

The $\xi_k$ weights satisfy $\sum_{k=0}^{M}\xi_k = \delta^{-1}$.

In this notation, we can write the binning approximation for $(2\pi)^{-1/2}n^{-1}\sum_{i=1}^{n}e^{is_n x_i}$ as

$$\begin{aligned}
&\approx (2\pi)^{-1/2}\sum_{k=0}^{M-1}\delta\xi_k e^{is_n t_k} \\
&= (2\pi)^{-1/2}\sum_{k=0}^{M-1}\delta\xi_k e^{is_n(a+k\delta)}.
\end{aligned}$$

$s_n$ are taken to be $s_n = n/M\delta$ for $n = -M/2, ..., M/2$:

$$\begin{aligned}
&= (2\pi)^{-1/2}\sum_{k=0}^{M-1}\delta\xi_k e^{i\frac{n}{M\delta}(a+k\delta)} \\
&= (2\pi)^{-1/2}e^{\frac{ia}{M\delta}}\{\sum_{k=0}^{M-1}\delta\xi_k e^{\frac{ink}{M}}\}.
\end{aligned}$$

This last expression is in the form needed to apply FFT. Jones and Lotwick (1983) show that the MISE of this approximation is $O(\delta^4)$.

## 8.2 Performance evaluation

In this section, we evaluate the gains in speed in a set-up where we are performing local linear regression and choosing smoothing parameters through least squares cross-validation (the LSCV method described in section 6). The computational method effectively factors the data component and the bandwidth component in the frequency domain, so that computation across different bandwidths can be done efficiently by reusing the data component of the computation. We show how these techniques work very well and make it feasible to do nonparametric and semiparametric estimation with sample sizes well over 100,000.

The following result is obtained for data generated by $y = \exp x$ without error, where $x$ has the standard normal distribution. Grids are constructed between -3 and 3. We estimate $E\{y|x\}$ at all

| | $\mathbf{n = 1000}$ | $\mathbf{n = 10,000}$ | $\mathbf{n = 100,000}$ |
|---|---|---|---|
| $M = 100$ | 1.02 sec<br>0.25% | 14.7 sec<br>0.27% | 185.8 sec<br>0.40% |
| $M = 500$ | 1.81 sec<br>0.047% | 11.5 sec<br>0.048% | 145.1 sec<br>0.049% |
| $M = 1000$ | 2.45 sec<br>0.021% | 11.5 sec<br>0.036% | 137.4 sec<br>0.041% |
| no approx. | 175.2 sec | 22257 sec | N/A |

Table 1: Speed/Accuracy comparisons

data points using the local linear regression method and use LSCV to select the globally optimum bandwidth. The machine we used is a DEC 5000/240.

Table 1 compares the speed and the average root percentage mean squared errors compared to the method without approximating (reported in the second row of each cell) for different size samples and for different grid sizes, $M$.

Speed does not necessarily increase with the gain in accuracy, because the computation involves optimization over the bandwidth. The time it takes for convergence, in our experience, goes down as $M$ increases. As one can see for the case of 10000 observations we can reduce the computation time to 0.036 % of the time it would otherwise take. For the case of 100,000 observations and for this workstation, the computation would have been a major task running over days as opposed to about 3 minutes with the approximation method.

## 9 Conclusions

In this chapter, we have reviewed recent advances in nonparametric and semiparametric estimation, with emphasis on applicability of methods in empirical research. Our discussion focused on the modeling and estimation of densities, conditional mean functions and derivatives of functions. The examples of section two illustrated how flexible modeling methods have been adopted in previous empirical studies, either as an estimation method in their own right or as a way of checking parametric modeling assumptions. Section three highlighted key concepts in semiparametric and nonparametric modeling that do not have counterparts in parametric modeling, such as the dependence of rates of convergence on the dimension of the estimation problem, the notion of models with an infinite number of parameters, the criteria used to define optimal convergence rates, and the existence of so-called "dimension-free" semiparametric estimators.

Section four of the chapter described a number of nonparametric approaches for estimating densities and conditional mean functions. Although nonparametric estimators are sometimes deemed infeasible because of slow convergence rates, they are nonetheless of keen interest because they form the building blocks of many semiparametric methods. We introduced some likelihood based and method of moments based approaches and presented a unifying framework for thinking about

how apparently different estimators relate to one another. The asymptotic distribution theory for the commonly used local polynomial regression estimator was also presented.

Section five studied application of a variety of semiparametric models that offer a middle ground between fully parametric and nonparametric approaches. By imposing some parametric restrictions, they typically achieve faster convergence rates than a nonparametric estimators. By remaining flexible with regard to certain aspects of the model, semiparametric estimators are consistent under a broader class of models than are fully parametric estimators. In some cases, flexibility can be achieved without sacrificing rates of convergence. However we note that semiparametric models are generally not embedded in a sequence of models in which arbitrary function can be approximated. It is desirable to consider such embedding and construct test against such sequences when semiparametric models are used. Stone's extended linear model provides such a framework for the additive separable models.

In section six we addressed questions that arise in implementing nonparametric methods, with regard to optimal choices of smoothing parameters and how best to implement trimming procedures. We reviewed a large and growing literature on bandwidth selectors for nonparametric density and regression estimators. Section six also considers the bandwidth selection problem in the context of semiparametric models, although that literature is still in its infancy. We described a few bandwidth selectors that have been proposed for index models and for the partially linear model.

Section seven presented a way to compute asymptotic variance of the semiparametric M-estimators. Section eight provided a brief introduction to some computational methods that have been introduced to ease the computational burden of nonparametric estimators when applied to large datasets. These methods show much promise, but their performance has yet to be widely studied in economic applications.

It is our hope that the topics of this chapter have provided an overview of how empirical researchers can best take advantage of recent developments in nonparametric and semiparametric modeling.

## References


[1] Abbring, J., Heckman, J. J. and E. Vytlacil (2006): Chapter in this Handbook.

[2] Ahmad, I. A. (1976): "On Asymptotic Properties Of An Estimate Of A Functional OF A Probability Density," *Scandinavian Actuarial Journal*, 176–181.

[3] Aït-Sahalia, Y. (1992): "The delta method for nonparametric kernel functionals," mimeo.

[4] Altonji, J. and H. Ichimura (1998): "Estimating Derivatives in Nonseparable Models with Limited Dependent Variables," mimeo.

[5] Ahn, H. and J. L. Powell (1993), "Semiparametric Estimation of Censored Sample Selection Models with A Nonparametric Selection Mechanism." *Journal of Econometrics*, 58, No. 1-2, 3–29.

[6] Ai, Chunrong, Richard Blundell and Xiaohong Chen (2000): "Semiparametric Engel Curves with Endogenous Expenditure," mimeo, UCL.

[7] Ai, Chunrong and Xiaohong Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

[8] Andrews, D. (1991): "Asymptotic Normality of Series Estimators for Various Nonparametric and Semiparametric Models, " *Econometrica*, 59, 307–345.

[9] Andrews, D. (1994): "Empirical Process Methods in Econometrics," *Handbook of Econometrics, Volume 4*. eds. R.F. Engle and D.L. McFadden, 2247–2294.

[10] Aradillas-Lopez, Andres, B. Honoré, and J. L. Powell (2005): "Pairwise Difference Estimation of Nonlinear Models with Nonparametric Functions," mimeo.

[11] Arellano, M. and B. Honoré (2001): "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics Vol. 5*, edited by J. Heckman and E. Leamer, North Holland.

[12] Ashenfelter, O. (1978): "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics,* 60, 47–57.

[13] Ashenfelter, O. and D. Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648–660.

[14] Banks, J., Blundell, R. and A. Lewbel (1997): "Quadratic Engel Curves and Consumer Demand" in Review of Economics and Statistics, 79, 527–539.

[15] Barron, Andrew R. (1993): "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions in Information Theory*, 39, 930–945.

[16] Bassi, L. (1984): "Estimating the Effects of Training Programs with Nonrandom Selection, " *Review of Economics and Statistics*, 66, 36–43.

[17] Berlinet, Alain and Christine Thomas-Agnan (2003): *Reproducing kernel Hilbert Spaces in probability and statistics*, Kluwer Academic Publishers, Boston.

[18] Bickel, P. (1982): "On Adaptive Estimation," *Annals of Statistics*, 10, 647–671.

[19] Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore and London.

[20] Bierens, H. J. (1985): "Kernel estimators of regression functions", *Advances in Econometrics*, edited by T. Bewley, Cambridge University Press, New York.

[21] Blundell, R. and Duncan, A. (1998): "Kernel Regression in Empirical Microeconomics," *The Journal of Human Resources*, 33, 62–87.

[22] Blundell, R., M. Browning and I. A. Crawford (2003): "Nonparametric Engel Curves and Revealed Preference," *Econometrica*, 71, 205–240.

[23] Blundell, R., X. Chen, and D. Kristensen (2003): "Semi-Nonparametric IV Estimation of Shape Invariant Engel Curves," mimeo.

[24] Blundell, R. and J. Powell (2003): "Endogeneity in Semiparametric Binary Response Models," mimeo.

[25] Breiman, L. and Friedman, J. H. (1985): "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–619.

[26] Buchinsky, Moshe (1994): "Changes in the U.S. wage structure 1963-1987: Application of quantile regression," *Econometrica*, 62, 405–454.

[27] Buchinsky, Moshe (1995): "Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963-1987," *Journal of Econometrics*, 65, 109–154.

[28] Buchinsky, Moshe (1995): "Estimating the asymptotic covariance matrix for quantile regression models. A Monte Carlo study" *Journal of Econometrics*, 68, 303–338.

[29] Buchinsky, Moshe (1998): "The Dynamics of Changes in the Female Wage Distribution is the USA: A Quantile Regression Approach," *Journal of Applied Econometrics*, 13, 1–30.

[30] Buchinsky, Moshe and Jinyong Hahn (1998): "An alternative estimator for the censored quantile regression model," *Econometrica*, 66, 653–671.

[31] Cao, R., Cuevas, A. and Gonzalez-Mantiega, W. (1994) "A Comparative Study of Several Smoothing Methods in Density Estimation" *Computational Statistics and Data Analysis*, 17, 153–176.

[32] Chamberlain, Gary (1986a): "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 32, 189–218.

[33] Chamberlain, Gary (1986b): "Notes on Semiparametric Regression," unpublished manuscript, Harvard University.

[34] Chamberlain, Gary (1995): "Quantile regression, censoring, and the structure of wages," *Advances in Econometrics: Sixth World Congress Vol. 1*, edited by C. Sims, Cambridge University Press.

[35] Chaudhuri, P, (1991a): "Global nonparametric estimation of conditional quantile functions and their derivatives," *Journal of Multivariate Analysis*, 39, 246–269.

[36] Chaudhuri, P, (1991b): "Nonparametric estimates of regression quantiles and their local Bahadur representation," *Annals of Statistics*, 19, 760–777.

[37] Chen, Xiaohong, O. Linton, and I. Van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71, 1591–1608.

[38] Chen, Xiaohong, H. White (1999): "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions in Information Theory*, 45, 682-691

[39] Choi, K. (1992): "The Semiparametric Estimation of the Sample Selection Model Using Series Expansion and the Propensity Score," PhD dissertation, University of Chicago.

[40] Chui, Charles (1992): *An introduction to wavelets*, Academic Press.

[41] Cleveland, William S. and Clive Loader (1996): "Smoothing by Local Regression: Principles and Methods." unpublished manuscript, AT&T Bell Laboratories.

[42] Conley, Timothy G., Lars P. Hansen, Wen-Fang Liu (1997): "Bootstrapping the long run," *Macroeconomic Dynamics*, 1, 279–311.

[43] Cosslett, S. R. (1987), "Efficiency bounds for distribution free estimators for the binary choice and the censored regression model," *Econometrica*, 55, 559–585.

[44] Cosslett, S. R. (1991): "Semiparametric Estimation of a Regression Model with Sample Selectivity" in *Nonparametric and Semiparametric Methods in Economics and Statistics*: ed. by W.A. Barnett, J. Powell and G. Tauchen. Cambridge: Cambridge University Press, 175–197.

[45] Daubechies, Ingrid (1992): *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia.

[46] Davidson, R. and J. G. MacKinnon (1982): "Some Non-nested Hypotheses Tests and the Relations Among Them," *The Review of Economic Studies*, 49, 551–565.

[47] Deaton, Angus (1996): *Microeconometric Analysis for Development Policy: Approach to Analyzing Household Surveys*, World Bank/The Johns Hopkins University Press.

[48] Deaton, Angus and Serena Ng (1998): "Parametric and Nonparametric Approaches in Price and Tax Reform" in *Journal of the American Statistical Association*, 93, 900–910.

[49] Deaton, Angus and Christina Paxson (1998): "Economies of Scale, Household Size and the Demand for Food" in *Journal of Political Economy*, 106, 897–930.

[50] Delgado, Miguel A. and Robinson, Peter (1992): "Nonparametric and Semiparametric Methods for Economic Research" in *Journal of Economic Surveys*, 3, 201–249.

[51] DiNardo, J., Fortin, N. and Lemiex, T. (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64, 1001-1044.

[52] Dirac, PAM (1958), *The Principles of Quantum Mechanics*,4th Edition (Revised), Clarendon Press, Oxford.

[53] Doksum, K., D. Peterson, A. Samarov (2000): "On variable bandwidth selection in local polynomial regression," *Journal of Royal Statistical Society, Series B*, 62, 431–448.

[54] Engle, Robert, Clive W. Granger, J. Rice and A. Weiss (1986): "Semiparametric Estimates of the Relation between Weather and Electricity Demand," *Journal of the American Statistical Association*, 81, 310–320.

[55] Epanechnikov, VA (1969): "Non-parametric estimation of a multivariate probability density," *Theory of Probability and its Applications*, 14, 153–158.

[56] Eubank, R. L. (1999): *Nonparametric Regression and Spline Smoothing*, 2nd edition, New York: Dekker.

[57] Fan, J. (1992): "Design Adaptive Nonparametric Regression, " *Journal of the American Statistical Association,* 87, 998–1004.

[58] Fan, J. and I. Gijbels (1992): "Variable bandwidth and local linear regression smoothers," *Annals of Statistics*, 20, 2008–2036.

[59] Fan, J. and I. Gijbels (1995): "Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation," *Journal of Royal Statistical Society, Series B*, 57, 371–394.

[60] Fan, J. and I. Gijbels (1996): *Local Polynomial Modeling and Its Applications*. New York: Chapman and Hall.

[61] Fan, J., P. Hall, M. Martin, P. Patil (1996): "On local smoothing of nonparametric curve estimators," *Journal of American Statistical Association*, 91, 258–266.

[62] Fan, J. and Marron (1994): "Fast implementations of nonparametric curve estimation," *Journal of Computational and Graphical Statistics*, 3, 35–56.

[63] Faraway, J. J. and Jhun, M. (1990) "Bootstrap Choice of Bandwidth for Density Estimation" *Journal of the American Statistical Association*, 85, 1119–1122.

[64] Fraker, T. and R. Maynard (1987): "The Adequacy of Comparison Group Designs for Evaluations of Employment Related Programs," *The Journal of Human Resources*, 22, 194–227.

[65] Goldberger, A. (1968): *Topics in Regression Analysis*, Wiley, New York.

[66] Goldberger, A. (1982): "Abnormal Selection Bias," *Studies in Econometrics, Time Series and Multivariate Statistics*, ed. by S. Karlin, T. Amemiya and L.A. Goodman, Academic Press, New York.

[67] Gronau, R. (1973): "New Econometric Approaches to Fertility" in *Journal of Political Economy*, 81, 168–199.

[68] Gronau, R. (1973): "The Intrafamily Allocation of Time: The Value of the Housewive's Time" in *The American Economic Review*, 63, 634–651.

[69] Green, P. J. and B. W. Silverman (1994): *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall.

[70] Hall, Peter (1982): "Limit theorems for stochastic measures of the accuracy of density estimation," *Stochastic Process Applications*, 13, 11–25.

[71] Hall, Peter (1983): "Large Sample Optimality of Least-Squares Cross-Validation in Density Estimation" in *Annals of Statistics*, 11, 1156–1174.

[72] Hall, Peter and Joel L. Horowitz (1990): "Bandwidth Selection in Semiparametric Estimation of Censored Linear Regression Models," *Econometric Theory*, 6, 123–150.

[73] Hall, P., S. Sheather, M. Jones and J. Marron (1991): "On optimal data-based bandwidth selection in kernel density estimation." *Biometrika*, 78, 263–269.

[74] Hall, P. and M. P. Wand (1996): "On the accuracy of binned kernel density estimates," *Journal of Multivariate Analysis*, 56, 165–184.

[75] Hall, Peter and Marron, J. S. (1987): "Extent to which Least-Squares Cross-Validation Minimizes Integrated Squared Error in Nonparametric Density Estimation," *Probability Theory and Related Fields*, 74, 567–581.

[76] Hall, Peter and Marron, J. S. (1991) "Local minima in cross-validation functions," *Journal of the Royal Statistical Society, Series B*, 53, 245–252.

[77] Hall, P. and P. Patil (1995): "Formulae for mean integrated squared error of nonlinear wavelet-based density estimators," *Annals of Statistics*, 23, 905–928.

[78] Hansen, Lars P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators" in *Econometrica*, 50, 1029–1059.

[79] Härdle, Wolfgang (1990): *Applied Nonparametric Regression*, Cambridge University Press.

[80] Härdle, Wolfgang (1991): *Smoothing Techniques with Implementation in S*, Springer-Verlag, New York.

[81] Härdle, Wolfgang and Gasser, T. (1984): "Robust nonparametric function fitting," *Journal of the Royal Statistical Society, Series B*, 46, 42–51.

[82] Härdle, Wolfgang and Oliver Linton (1994): "Applied Nonparametric Methods," *Handbook of Econometrics, Vol. 4*, Elsevier, Amsterdam, 2295–2339.

[83] Härdle, Wolfgang and Linton, Oliver B. (1996): "Estimating Additive Regression with Known Links," *Biometrika*, 83, 529–540.

[84] Härdle, Wolfgang, Peter Hall and Hidehiko Ichimura (1993): "Optimal Semiparametric Estimation in Single Index Models," *Annals of Statistics*, 21, 157–178.

[85] Härdle, Wolfgang, Hildebrand, and M. Jerison (1991): "Empirical Evidence on the Law of Demand," *Econometrica*, 59, 1525–1549.

[86] Härdle, Wolfgang, J. Hart, J. S. Marron, and A. B. Tsybakov (1992): "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association*, 87, 218–226.

[87] Härdle, Wolfgang and A. B. Tsybakov (1993): "How sensitive are average derivatives?," *Journal of Econometrics*, 58, 31–48.

[88] Härdle, Wolfgang and Thomas Stoker (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of American Statistical Association*, 84, 986–995.

[89] Hasminskii, R. Z. and I. A. Ibragimov (1979): "On the nonparametric estimation of functionals," *Proceedings of the Second Prague Symposium on Asymptotic Satistics*, editors M. Mandl and M. Husková, 41–51, North-Holland, Amsterdam.

[90] Hastie, Trevor J. and Tibshirani, R. J. (1990): *Generalized Additive Models*, Chapman and Hall.

[91] Hausman, J. (1978): "Specification Tests in Econometrics" in *Econometrica*, 46, 1251–1272.

[92] Heckman, James J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

[93] Heckman, James J. (1980): "Addendum to Sample Selection Bias as specification Error," *Evaluation Studies Review Annual*, ed. by E. Stromsdorfer and G. Frakas. San Fransisco, Sage.

[94] Heckman, James J. (1990): "Varieties of Selection Bias," *American Economic Review*, 80, 313–328.

[95] Heckman, James J. and Joseph Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–880.

[96] Heckman, James J. and Richard Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions" in *Longitudinal Analysis of Labor Market Data*, eds. J. Heckman and B. Singer, Cambridge University Press.

[97] Heckman, James J. and Jeffrey Smith (1995): "Assessing the Case for Randomized Social Experiments", *Journal of Economic Perspectives*, 9, 85–100.

[98] Heckman, James J. , Hidehiko Ichimura, and Petra Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies*, 64, 605–654.

[99] Heckman, James J. , Hidehiko Ichimura, Jeffrey Smith, and Petra Todd (1998): "Nonparametric Characterization of Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

[100] Heckman, James J., Hidehiko Ichimura, and Petra Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

[101] Heckman, James J., Ichimura, Hidehiko and Petra Todd (1998): "Implementing the Partially Linear Regression Model," unpublished manuscript.

[102] Heckman, James J., Lochner, Lance and Petra Todd (2005): "Earnings functions, rates of return and treatment effects: The Mincer equation and beyond, forthcoming in *Handbook of Education Economics.*

[103] Hjort, N. L. and M. C. Jones (1996): "Local Parametric Nonparametric Density Estimation " in *Annals of Statistics*, 24, 1619–1647.

[104] Honoré, Bo and J. Powell (1994): "Pairwise difference estimators of censored and truncated regression models," *Journal of Econometrics*, 64, 241–278.

[105] Honoré, Bo and J. Powell (2005): "Pairwise difference estimation of nonlinear models," *Identification and Inference for Econometric Models,* edited by D. Andrews and J. Stock, Cambridge University Press, NY.

[106] Horowitz, Joel L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

[107] Horowitz, Joel L. (1998): *Semiparametric Methods in Econometrics: Lecture Notes in Statistics,* Vol. 131, New York and Heidelberg: Springer Verlag.

[108] Horowitz, Joel L. and W. Härdle (1996): "Direct semiparametric estimation of single-index models with discrete covariates," *Journal of American Statistical Association*, 91, 1632–1640.

[109] Horowitz, Joel L. and S. Lee (2005): "Nonparametric Estimation of an Additive Quantile Regression Model," *Journal of the American Statistical Association*, 100, 1238–1249.

[110] Huang, Jianhua Z. (1998): "Projection estimation in multiple regression with application to functional ANOVA models," *Annals of Statistics*, 26, 242–272.

[111] Huang, Jianhua Z. (2003): "Local asymptotics for polynomial spline regression," *Annals of Statistics*, 31, 1600–1635.

[112] Huang, Su-Yun (1999): "Density estimation by wavelet-based reproducing kernels," *Statistica Sinica*, 9, 137–151.

[113] Ichimura, Hidehiko (1993): "Semiparametric Least Squares Estimation of Single Index Models (SLS) and Weighted SLS Estimation of Single Index Models." *Journal of Econometrics,* 58, 71–120.

[114] Ichimura, Hidehiko (1995): "Asymptotic Distribution Theory for Semiparametric and Nonparametric Estimators with Data Dependent Smoothing Parameters," unpublished manuscript, University of Pittsburgh.

[115] Ichimura, Hidehiko and Lung-Fei Lee ( 1991): "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in *Nonparametric and Semiparametric Methods in Economics and Statistics,* ed. by W.A. Barnett, J. Powell, and G. Tauchen (Cambridge, England: Cambridge University Press) 3–49.

[116] Ichimura, H. and O. Linton (2005): "Asymptotic expansions for some semiparametric program evaluation estimators," *Identification and Inference for Econometric Models,* edited by D. Andrews and J. Stock, Cambridge University Press, NY.

[117] Ichimura, H. and S. Lee (2006): "Characterization of the asymptotic distribution of semiparametric M-estimators," mimeo, University of Tokyo.

[118] Jones, M. C. and S.J. Sheather (1991): "A reliable data-based bandwidth selection method for kernel density estimation" in *Journal of the Royal Statistical Society, Series B*, 53, 683–690.

[119] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996): "A Brief Survey of Bandwidth Selection for Density Estimation" in *Journal of the American Statistical Association*, 91, 401–407.

[120] Jones, M. C. and J. S. Marron and S. J. Sheather, (1992): "Progress in Data-Based Bandwidth Selection for Kernel Density Estimation," manuscript.

[121] Jones, M.C. and Lotwick, H.W. (1983), "On the errors involved in computing the empirical characteristic function" in *Journal of Statistical Computation and Simulation*, 17, 133–149.

[122] Jones, M.C. and Lotwick, H.W. (1984), "A Remark on Algorithm AS 176 Kernel Density Estimation using the Fast Fourier Transform" in *Applied Statistics*, 33, 120–122.

[123] Klein, R. W. and R. H. Spady (1993), "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387–421.

[124] Kim, Jeankyung and D. Pollard (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18, 191–219.

[125] Koenker, R. (2005): *Quantile Regression*, Cambridge University Press.

[126] Koenker, R. and G. Bassett (1978), "Regression Quantiles" in *Econometrica*, 46, 33–50.

[127] Lee, Myoung-Jae (1996): *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*, Springer-Verlag.

[128] Lewis, H. Gregg (1974): "Comments on Selectivity Biases in Wage Comparisons" in *Journal of Political Economy*, 82, 1145–1155.

[129] Linton, Oliver B. (1995): "Second Order Approximation in a Partially Linear Regression Model," *Econometrica*, 63, 1079–1113.

[130] Linton, Oliver B. (1995): "Estimation in semiparametric models: A review," *A Volume in Honor of C.R. Rao*, in P.C.B. Phillips and G.S. Maddala (eds.), Blackwell.

[131] Linton, Oliver B. (1996): "Edgeworth approximation for MINPIN estimators in semiparametric regressions models," *Econometric Theory*, 12, 30–60.

[132] Linton, Oliver B. (1997): "Efficient estimation of additive nonparametric regression models," *Biometrika*, 84, 469–474.

[133] Linton, Oliver B., Chen, Rong, Wang, Naiysin, and Härdle, Wolfgang (1997): "An Analysis of Transformations for Additive Nonparametric Regression" in *Journal of the American Statistical Association*, 92, 1512–1521.

[134] Linton, Oliver, B. and J. P. Nielsen (1995): "A kernel method of estimating structured nonparametric regression based on marginal integration," *Biometrika*, 82, 93–100.

[135] Loader, Clive (1995): "Old Faithful Erupts: Bandwidth Selection Reviewed", manuscript, AT&T Bell Laboratories.

[136] Loader, Clive (1996): "Local Likelihood Density Estimation" in *Annals of Statistics*, 24, 1602–1618.

[137] McFadden, Daniel L. (1985): Presidential address at the World Congress of the Econometric Society.

[138] Malinvaud, E. B. (1970): *Statistical Methods of Econometrics*, Amsterdam: North-Holland.

[139] Manski, Charles F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

[140] Manski, Charles F. (1985): "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator" *Journal of Econometrics*, 27, 313–333.

[141] Manski, Charles F. (1994): "Analog Estimation of Econometric Models" in *Handbook of Econometrics, Volume 4*, eds. R.F. Engle and D.L. McFadden, North-Holland.

[142] Manski, Charles F. and Daniel McFadden (1981): "Alternative Estimators and Sample Designs for Discrete Choice Analysis" in *Structural Analysis of Discrete Data with Economic Applications*, edited by C.F. Manski and D. McFadden (Cambridge, Mass.: MIT Press) 1–50.

[143] Masry, E (1996a): "Multivariate regression estimation: Local Polynomial fitting for time Series," *Stochastic Processes and their Applications*, 65, 81–101.

[144] Masry, E (1996b): "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates," *Journal of Time Series Analysis*, 17, 571–599.

[145] Matzkin, Rosa (1994): "Restrictions of Economic Theory in Nonparametric Methods" in *Handbook of Econometrics, Volume 4*. eds. R.F. Engle and D.L. McFadden, 2524–2554.

[146] Moore, David S. and James W. Yackel (1977a): "Consistency Properties of Nearest Neighbor Density Function Estimators," *Annals of Statistics*, 5, 143–154.

[147] Moore, David S. and James W. Yackel (1977b): "Large Sample Properties of Nearest Neighbor Density Function Estimators," *Statistical Decision Theory and Related Topics*, editors S. S. Gupta and D. S. Moore, Academic Press, New York.

[148] Nadaraya, E. A. (1964): "On estimating regression" in *Theory of Probability and its Applications*, 10, 186–190.

[149] Newey, Whitney K. (1985): "Generalized method of moments specification tests," *Journal of Econometrics*, 29, 229–256.

[150] Newey, Whitney K. (1987): "Specification tests for distributional assumptions in the Tobit model," *Journal of Econometrics*, 34, 125–145.

[151] Newey, Whitney K. (1988): "Two-step series estimation of sample selection models," working paper, MIT.

[152] Newey, Whitney K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

[153] Newey, Whitney K. (1994): "The Asymptotic Variance of Semiparametric Estimators" *Econometrica*, 62, 1349–1382.

[154] Newey, Whitney K. (1994):"Kernel Estimation of Partial Means" in *Econometric Theory*, 10, 233–253.

[155] Newey, Whitney K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

[156] Newey, Whitney K. and James Powell (1990): "Efficient Estimation of Linear and Type I Censored Regression Models Under Conditional Quantile Restrictions," *Econometric Theory*, 6, 295–317.

[157] Newey, Whitney K., Powell, James and James Walker (1990): "Semiparametric estimation of selection models: some empirical results," *American Economic Review*, 80, 324–328.

[158] Newey, Whitney and Daniel McFadden (1994): "Large Sample Estimation and Hypothesis Testing" *Handbook of Econometrics, Volume 4*. eds. R.F. Engle and D.L. McFadden, 2113-2241.

[159] Newey, Whitney and Thomas Stoker (1993): "Efficiency of Weighted Average Derivative Estimators," *Econometrica*, 61, 1199–1223.

[160] Nishiyama, Y. and P. Robinson (2000): "Edgeworth expansions for semiparametric averaged derivatives," *Econometrica*, 68, 931–980.

[161] Nishiyama, Y. and P. Robinson (2001): "Studentization in Edgeworth expansions for estimates of semiparametric single index models," in *Nonlinear Statistical Modeling* edited by C. Hsiao, K. Morimune, and J. Powell, Cambridge University Press, UK.

[162] Nishiyama, Y. and P. Robinson (2005): "The Bootstrap and the Edgeworth Correction for Semiparametric Average Derivatives," *Econometrica*, 73, 903–948.

[163] Ochiai, Toshimitsu and Kanta Naito (2003): "Asymptotic theory for the multiscale wavelet density derivative estimator," *Communications in Statistics Theory and Methods*, 32, 1925–1950.

[164] Olley, Steve and Pakes, Ariel (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

[165] Pagan, Adrian and Ullah, Aman (1999): *Nonparametric Econometrics*, Cambridge University Press, Cambridge, MA.

[166] Pakes, Ariel and David Pollard (1989): "Simulation and Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

[167] Park, B. U. and Marron, J. S. (1990): "Comparison of data-driven bandwidth selectors" *Journal of the American Statistical Association*, 85, 66–72.

[168] Park, B. U., W. C. Kim, J. S. Marron (1994): "Asymptotically best bandwidth selectors in kernel density estimation," *Statistics and Probability Letters*, 19, 119–127.

[169] Park, Byeong U. and Turlach, Berwin A. (1992): "Reply to comments on 'Practical performance of several data driven bandwidth selectors," *Computational Statistics*, 7, 283–285.

[170] Parzen, E. (1962) "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, 33, 1065–1076.

[171] Pollard, David (1984): *Convergence of Stochastic Processes,* Springer-Verlag, New York.

[172] Pollard, David (1985): "New ways to prove central limit theorems," *Econometric Theory*, 1, 295–314.

[173] Powell, James (1984): "Least Absolute Deviations Estimator for the Censored Regression Model" in *Journal of Econometrics*, 25, 303–325.

[174] Powell, James (1987): "Semiparametric Estimation of Bivariate Latent Variable Models," SSRI, University of Wisconsin–Madison, Working Paper No. 8704.

[175] Powell, James (1994): "Estimation of Semiparametric Models." *Handbook of Econometrics, Volume 4*. eds. R.F. Engle and D.L. McFadden, 2443-2521.

[176] Powell, J. Stock and Thomas Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 6, 1403-1430.

[177] Powell, James and Thomas Stoker (1996): "Optimal bandwidth choice for density weighted averages," *Journal of Econometrics*, 75, 291–316.

[178] Prakasa-Rao (1983): *Nonparametric Functional Estimation*, (Orlando: Academic Press).

[179] Prewitt, K. and S. Lohr (2006): "Bandwidth selection in local polynomial regression using eigenvalues," *Journal of Royal Statistical Society, Series B*, 68, 135–154.

[180] Ritov, Y and P. Bickel (1990): "Achieving information bounds in non and semiparametric models," *Annals of Statistics*, 18, 925–938.

[181] Robinson, Peter M. (1983): "Nonparametric Estimators for Time Series," *Journal of Time Series Analysis*, 4, 185–207.

[182] Robinson, Peter M. (1988): "Root-N Consistent Nonparametric Regression," *Econometrica*, 56, 931–954.

[183] Robinson, Peter M. (1989): "Hypothesis testing in semiparametric and nonparametric models for econometric time series," *Review of Economic Studies*, 56, 511–534.

[184] Robinson, Peter M. (1991): "Automatic Frequency Domain Inference on Semiparametric and Nonparametric Models," *Econometrica*, 59, 1329–1363.

[185] Robinson, Peter M. (1995): "The normal approximation for semiparametric averaged derivatives," *Econometrica*, 63, 667–680.

[186] Rosenblatt, M. (1956) "Remarks on some Nonparametric Estimators of a Density Function" *Annals of Mathematical Statistics*, 27, 832-837.

[187] Rosenzweig, Mark and Wolpin, Kenneth I. (2000): "Natural Natural Experiments," *Journal of Economic Literature*, 38, 827–874.

[188] Rudemo (1982): "Empirical choice of histograms and kernel density estimators" in *Scandinavian Journal of Statistic*s, 9, 65–78.

[189] Ruppert, D. (1997): "Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation," *Journal of American Statistical Association*, 92, 1049–1062.

[190] Ruppert, D. and M. P. Wand (1994): "Multivariate locally weighted least squares regression," *Annals of Statistics*, 22, 1346–1370.

[191] Saito, S. (1989): *Theory for reproducing kernels and its applications*, Wiley, New York.

[192] Schmalensee, R. and Stoker, Thomas (1999): "Household Gasoline Demand in the United States," *Econometrica*, 67, 645–662.

[193] Schuster, E. F. and Gregory, C. G. (1981): "On the nonconsistency of maximum likelihood nonparametric density estimators," in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface,* Eddy, W. F. (editor), 295–298, Springer-Verlag, NY.

[194] Schweder, T. (1975): "Window estimation of the asymptotic variance of rank estimators of location," *Scandinavian Journal of Statistics*, 2, 113–126.

[195] Scott, David W. (1992): *Multivariate Density Estimation*, John Wiley & Sons, New York.

[196] Scott, D.W. and Terrell, G. R.(1987): "Biased and Unbiased Cross-Validation in Density Estimation" in *Journal of American Statistical Association*, 82, 1131–1146.

[197] Sherman, (1994a): "Maximal Inequalities for Degenerate U-Processes with Application to Optimization Estimators," *Annals of Statistics*, 22, 439–459.

[198] Sherman, R. (1994b): "U-Processes in the analysis of a generalized semiparametric regression estimator," *Econometric Theory*, 10, 372–395.

[199] Shiller, R.J. (1984): "Smoothness Priors and Nonlinear Regression," *Journal of the American Statistical Association,* 72, 420–423.

[200] Schoenberg, I. J. (1964): "Spline functions and the problem of graduation," *Proceedings of the National Academy of Sciences*, 52, 947–950.

[201] Silverman, B.W. (1982): "Kernel density estimation using the fast Fourier Transform Method," *Applied Statistics*, 31, 93–99.

[202] Silverman, B.W. (1982): "On the estimation of a probability density function by the maximum penalized likelihood method" *Annals of Statistics*, 10, 795–810.

[203] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[204] Smith, Jeffrey and Todd, Petra (2001): "Reconciling Conflicting Evidence on the Performance of Propensity Score Matching Estimators," *American Economic Review*, 91, 112–118.

[205] Smith, Jeffrey and Todd, Petra (2005): "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, 125, 305–353. Rejoinder 125, 365–375.

[206] Stein, Charles (1956): "Efficient nonparametric testing and estimation," *Proc. Third Berkeley Symp. Math. Statist. Prob.*, 1, 187–195, Univ California Press, Berkeley.

[207] Stern, Steven (1996): "Semiparametric estimates of the supply and demand effects of disability on labor force participation" in *Journal of Econometrics*, 71, 49–70.

[208] Stock, James (1991): "Nonparametric Policy Analysis: An application to estimating hazardous waste cleanup benefits" in *Nonparametric and Semiparametric Methods in Economics and Statistics,* ed. by W.A. Barnett, J. Powell, and G. Tauchen (Cambridge, England: Cambridge University Press) 77–98.

[209] Stoker, Thomas (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.

[210] Stoker, Thomas (1991): "Equivalence of Direct, Indirect, and Slope Estimators of Average Derivatives" in *Nonparametric and Semiparametric Methods in Economics and Statistics,* ed. by W.A. Barnett, J. Powell, and G. Tauchen (Cambridge, England: Cambridge University Press), 99–118.

[211] Stoker, Thomas (1996): "Smoothing bias in the measurement of marginal effects," *Journal of Econometrics*, 72, 49–84.

[212] Stone, Chales (1974): "Cross-validatory choice and assessment of statistical predictions (with discussion)" in *Journal of the Royal Statistical Society, Series B*, 36, 111–147.

[213] Stone, Charles (1977): "Consistent nonparametric regression (with discussion)," *Annals of Statistics*, 5, 595–645.

[214] Stone, Charles (1980): "Optimal rates of convergence for nonparametric estimators," *Annals of Statistics*, 8, 1348–1360.

[215] Stone, Charles (1982): "Optimal rates of convergence for nonparametric regression," *Annals of Statistics*, 10, 1040–1053.

[216] Stone, Charles (1984): "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *Annals of Statistics*, 12, 1285–1297.

[217] Stone, Charles, M. Hansen, C. Kooperberg, Y. Truong (1997): "Polynomial Splines and their Tensor Products in Extended Linear Modeling," *Annals of Statistics*, 25, 1371–1425.

[218] Tauchen, George (1985): "Diagnostic testing and evaluation of maximum likelihood models," *Journal of Econometrics*, 30, 415–443.

[219] Taylor, Charles (1989): "Bootstrap choice of the smoothing parameter in kernel density estimation" *Biometrika*, 76, 705–712.

[220] Tsybakov, A. B. (1982): "Robust estimates of a function," *Problems of information transmission*, 18, 190–201.

[221] Ullah, Aman and Vinod, H. D. (1993), *Nonparametric Econometrics,* Cambridge University Press.

[222] Van der Vaart, A. W., (1998) *Asymptotic Statistics*, Cambridge University Press.

[223] Van der Vaart, A. W. and Jon W. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

[224] Wahba, Grace (1984): "Partial Spline Models for the Semi-Parametric Estimation of Functions of Several Variables," in *Statistical Analysis of Time Series,* ed. by R.A. Bradley, J.S. Hunter, D.G. Kendall and G.S. Watson. Tokyo: Institute of Statistical Mathematics.

[225] Wahba, Grace (1990): *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia.

[226] Walter, G. and Blum, J. R. (1979): "Probability density estimation using delta sequences," *Annals of Statistics*, 7, 328–340.

[227] Wand, M. P. (1994): "Fast computation of multivariate kernel estimators," *Journal of Computational and Graphical Statistics*, 3, 433–445.

[228] Watson, G. S. (1964), "Smooth regression analysis" in *Sankhya, Series A*, 26, 357–372.

[229] Weinert, H. (1982): editor, *Reproducing kernel Hilbert spaces: Applications in Statistical Signal Processing*, Hutchinson Ross, Stroudsburg, PA.

[230] White, Halbert (1980), "Using least squares to approximate unknown regression functions," *International Economic Review*, 21, 149–170.

[231] Willis, R. (1986), "Wage determinants: a survey and reinterpretation of human capital earnings functions" in *Handbook of Labor Economics*, eds. Orley Ashenfelter and Richard Layard.

[232] Wu, De-Min (1974), "Alternative tests of independence between stochastic regressors and disturbances: Finite sample results," *Econometrica*, 42, 529–546.

[233] Yatchew, Adonis (1997): "An Elementary Estimator of the Partial Linear Model," *Economics Letters*, 57, 135–143.

[234] Yatchew, Adonis (1998): "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature*, 26, 669–721.

[235] Yatchew, Adonis (2003): *Semiparametric Regression for the Applied Econometrician*, Cambridge University Press.

[236] Yu, Keming and M. C. Jones (1998), "Local Linear Quantile Regression" *Journal of the American Statistical Association*, 93, 228-237.

[237] Zemanian, A. H. (1965), "Distribution theory and Transform Analysis," Dover Publications, Inc. New York.

[238] Zhou, S., X. Shen, and D. A. Wolfe (1998): "Local asymptotics for regression splines and confidence regions," *Annals of Statistics*, 26, 1760–1782.