

## NONPARAMETRIC ESTIMATION OF ADDITIVE MODELS

Joel L. Horowitz

*Department of Economics, Northwestern University*

*2001 Sheridan Road, Evanston, IL 60208, U.S.A.*

*E-mail: [joel-horowitz@northwestern.edu](mailto:joel-horowitz@northwestern.edu)*

This chapter is about nonparametric additive modeling of a conditional mean or quantile function. Nonparametric additive modeling relaxes the restrictive functional form assumptions of parametric modeling while avoiding many of the disadvantages of fully nonparametric estimation. The chapter reviews recently developed methods for estimating nonparametric additive models with and without link functions. The emphasis is on methods that avoid the curse of dimensionality and achieve a desirable property called oracle efficiency.

### 1. Introduction

Much empirical research in statistics and econometrics is concerned with estimating conditional mean or quantile functions. For example, labor economists are interested in estimating the mean wages of employed individuals conditional on characteristics such as years of work experience and education. The most frequently used estimation methods assume that the function of interest is known up to a set of constant parameters that can be estimated from data. Such models are called *parametric*. The use of a parametric model greatly simplifies estimation, statistical inference, and interpretation of the estimation results but is rarely justified by theoretical or other *a priori* considerations. Estimation and inference based on convenient but incorrect assumptions about the form of the conditional mean or quantile function can be highly misleading.

Many investigators attempt to minimize the risk of specification error by carrying out a *specification search* in which several different parametric models are estimated and conclusions are based on the one that appears to fit the data best. Specification searches may be unavoidable in some applications, but they have many undesirable properties and their use should be minimized. There is no guarantee that a specification search will include the correct model or a good approximation to it. If the search includes the correct model, there is no guarantee that it will be selected by the investigator's model selection criteria. Moreover, the search process invalidates the statistical theory on which inference is based.

The rest of this chapter describes methods for dealing with the problem of specification error by relaxing the assumptions about functional form that are made by parametric models. The possibility of specification error can be essentially eliminated through the use of nonparametric estimation methods. These methods assume that the function of interest is smooth but make no other assumptions about its shape or functional form. However, nonparametric methods have important disadvantages that seriously limit their usefulness in applications. One important problem is that the precision of a nonparametric estimator decreases rapidly as the dimension of the explanatory variable,  $X$ , increases. This phenomenon is called the *curse of dimensionality*. As a result of it, impracticably large samples are usually needed to obtain acceptable estimation precision if  $X$  is multidimensional, as it often is in applications. For example, a labor economist may want to estimate mean log wages conditional on years of work experience, years of education, and one or more indicators of skill levels, thus making the dimension of  $X$  at least three.

Another problem is that nonparametric estimates can be difficult to display, communicate, and interpret when  $X$  is multidimensional. Nonparametric estimates do not have simple analytic forms. If  $X$  is one- or two-dimensional, then the estimate of the function of interest can be displayed graphically, but only reduced-dimension projections can be displayed when  $X$  has three or more components. Many such displays and much skill in interpreting them can be needed to fully convey and comprehend the shape of an estimate.

A further problem with nonparametric estimation is that it does not permit extrapolation. For example, in the case of a conditional mean function it does not provide predictions of  $E(Y|x)$  at points  $x$  that are outside of the support (or range) of the random variable  $X$ . This is a serious drawback in policy analysis and forecasting, where it is often important to predict what might happen under conditions that do not exist in the available data. Finally, in nonparametric estimation, it can be difficult to impose restrictions suggested by economic or other theory. Matzkin (1994) discusses this issue.

A variety of methods are now available for overcoming the curse of dimensionality and other drawbacks of fully nonparametric estimation. These methods offer a compromise between the flexibility of fully nonparametric estimation and the precision of parametric models. They make assumptions about functional form that are stronger than those of a nonparametric model but less restrictive than those of a parametric model, thereby reducing (though not eliminating) the possibility of specification error. They permit greater estimation precision than do nonparametric methods when  $X$  is multidimensional, the estimation results are easier to display and interpret than in fully nonparametric estimation, and there are limited capabilities for extrapolation and imposing restrictions derived from economic or other theory models. Leading examples of such methods are semiparametric index models (Hristache et al. 2001; Hristache, Juditsky, and Spokoiny 2001; Ichimura 1993. Ichimura and Lee 1991, Klein and Spady 1993; Manski 1988; Powell, Stock, and Stoker), partially linear models (Engle et al. 1986; Robinson 1988; Stock 1989, 1991; Härdle, Liang, and Gao 2000), and nonparametric additive models. Nonparametric additive models are the subject of this chapter.

This methods described in this chapter assume that the conditional mean or quantile function of  $Y$  given  $X = x$  has the form

$$E(Y|X = x) \text{ or } Quantile(Y|X = x) = F[\mu + m_1(x^1) + \dots + m_d(x^d)]. \quad (1.1)$$

In this specification  $d$  is the dimension of  $X$  and  $x$ , and  $x^j$  ( $j=1, \dots, d$ ) is the  $j$ 'th component of  $x$ .  $F$  is a known function, possibly the identity function, called a *link function*;  $\mu$  is an unknown

constant; and  $m_1, \dots, m_d$  are unknown functions with scalar arguments. It turns out that in large samples, each of the additive components  $m_j$  can be estimated with the same accuracy that could be achieved if  $X$  were a scalar. Moreover, each  $m_j$  can be estimated with the accuracy it would have if the other  $m_j$ 's were known. This property is called *oracle efficiency*. It means that in large samples there is no accuracy penalty for a high dimensional  $X$  or not knowing the other  $m_j$ 's. However, (1.1) is less flexible than a fully nonparametric model and, therefore, carries a risk of specification error. This is the price that must be paid for avoiding the curse of dimensionality.

Sec. 2 of this chapter reviews kernel nonparametric estimation of conditional mean functions. Kernel estimation is used repeatedly in the rest of the chapter. Sec. 3 describes nonparametric estimation of additive models of conditional mean functions when  $F$  is the identity function. Sec. 4 extends the results of Sec. 3 to the case in which  $F$  is not necessarily the identity function. Sec. 5 treats nonparametric additive quantile regression, and Sec. 6 presents an empirical example. The presentation is informal. Technical details and proofs of results are available in the reference material that is cited throughout the chapter.

## 2. Kernel Estimation of a Conditional Mean Function

In nonparametric estimation of a conditional mean function,  $g(x) \equiv E(Y | X = x)$  is assumed to satisfy smoothness conditions such as differentiability, but no assumptions are made about its shape or the form of its dependence on  $x$ . Härdle (1990) and Fan and Gijbels (1996) provide detailed discussions of nonparametric estimation methods. Kernel estimation is an easily understood and frequently used method. To describe the kernel method, assume that  $X$  is a continuously distributed, random  $d$ -vector for some finite  $d \geq 1$ . Let  $\{Y_i, X_i : i = 1, \dots, n\}$  be a random sample of  $n$  observations of  $(Y, X)$ . Let  $K_1$  be a bounded function on  $[-1, 1]$  that satisfies

$$\int_{-1}^1 K_1(z) dz = 1$$

and

$$\int_{-1}^1 z^j K_1(z) dz = 0; \quad j = 1, \dots, r-1$$

for some integer  $r \geq 2$ . For a vector  $v \in [-1, 1]^d$  with components  $v^1, \dots, v^d$  define

$$K(v) = \prod_{j=1}^d K_1(v^j).$$

Let  $\{h_n\} \equiv \{h\}$  be a sequence of positive numbers (called bandwidths) that converges to 0 as  $n \rightarrow \infty$ . For each  $n = 1, 2, \dots$  and  $i = 1, \dots, n$  define the function  $w_{ni}(\cdot)$  by

$$w_{ni}(x) = \frac{K[(x - X_i)/h]}{\sum_{i=1}^n K[(x - X_i)/h]}. \quad (2.1)$$

Then the kernel nonparametric estimator of  $g(x)$  is

$$g_n(x) = \sum_{i=1}^n w_{ni}(x) Y_i. \quad (2.2)$$

The estimator  $g_n(x)$  is a weighted average of the observed values of  $Y$ . Observations  $Y_i$  for which  $X_i$  is close to  $x$  get higher weight than do observations for which  $X_i$  is far from  $x$ .

Härdle (1990) provides a detailed discussion of the statistical properties of kernel nonparametric estimators. It can be shown that as  $n \rightarrow \infty$ ,  $g_n(x)$  converges in probability to  $g(x)$  if  $h \rightarrow 0$  and  $nh_n \rightarrow \infty$ . Thus, if  $n$  is large,  $g_n(x)$  is likely to be very close to  $g(x)$ . However, the rate of convergence is slow if  $d$  is large. Specifically, if  $g$  has  $r$  bounded derivatives with respect to any combination of its arguments, then the fastest possible rate of convergence in probability of  $g_n(x)$  to  $g(x)$  is  $n^{-r/(2r+d)}$ . This occurs when  $h \propto n^{-1/(2r+d)}$ . Thus, the rate of convergence of  $g_n(x)$  to  $g(x)$  decreases as  $d$  increases. This is the curse of dimensionality, and it is unavoidable in nonparametric estimation. As a result of it, impracticably large samples are usually needed to obtain acceptable estimation precision if  $X$  is multidimensional.

Kernel estimators are asymptotically normal. It suffices for the purposes of this chapter to consider only the case of a scalar  $X$ . Define

$$\int_{-1}^1 z^r K(z) dz = A$$

and

$$\int_{-1}^1 [K(z)]^2 dz = B.$$

Let  $p$  denote the probability density function of  $X$ , and set  $h = cn^{-1/(2r+1)}$  for some finite  $c > 0$ . Then

$$n^{r/(2r+1)}[g_n(x) - g(x)] \rightarrow^d N(\mu_R, \sigma_R^2),$$

where

$$\mu_R = \frac{c^r}{p(x)} AD(x),$$

$$D(x) = \sum_{k=1}^s \frac{1}{k!} \frac{d^k}{dv^k} \{[g(v+x) - g(x)]p(x)\}_{v=0},$$

and

$$\sigma_R^2 = \frac{B\sigma^2(x)}{cp(x)}.$$

The asymptotic distribution of  $n^{r/(2r+1)}[g_n(x) - g(x)]$  is not centered at 0. Centering at 0 can be achieved, though at the cost of a slower rate of convergence, by choosing the bandwidth  $h$  to converge more rapidly than  $n^{-1/(2r+1)}$ .

Local linear modeling is another kernel-based method for estimating a conditional mean function nonparametrically. It amounts to approximating  $g(x)$  by a straight line in a small neighborhood of  $x$  and using least squares to estimate the intercept and slope of the line. Specifically let

$$(\beta_0, \beta_1) = \arg \min_{b_0, b_1} \sum_{i=1}^n [Y_i - b_0 - b_1'(X_i - x)]^2 K\left(\frac{X_i - x}{h}\right),$$

where  $\dim(b_1) = d$ . The local linear estimator of  $g(x)$  is  $\beta_0$ . If  $g(x)$  is twice continuously differentiable and  $h \propto n^{-1/5}$ , then the local linear

estimator converges in probability at the rate  $n^{-2/(4+d)}$ , which is the fastest possible rate under these conditions. A local linear estimator is often better behaved at the boundary of the support of  $X$  than an ordinary kernel estimator is. Fan and Gijbels (1996) provide a detailed discussion of the properties of local linear estimators.

### 3. Nonparametric Additive Models for Conditional Mean Functions

This section is concerned with estimating the model

$$E(Y | X = x) = \mu + m_1(x^1) + \dots + m_d(x^d), \quad (3.1)$$

where  $\mu$  and the  $m_j$ 's are defined as in Sec. 1. We describe three kinds of estimators of  $\mu$  and the additive components  $m_j$ . The first is called marginal integration. It is conceptually simple but can be hard to compute and is not oracle efficient. The other two are backfitting and a two-step method. The two-step method and a version of backfitting provide asymptotically normal, oracle efficient estimators. The two step method can be extended easily to estimation of an additive model with a link function or an additive conditional quantile function. Backfitting cannot be extended easily this way.

The property of oracle efficiency is discussed repeatedly in the remainder of this chapter. To define it precisely, suppose that  $m_2, \dots, m_d$  and  $\mu$  were known. Then  $m_1(x^1)$  could be estimated by carrying out the nonparametric regression of  $Y - \mu - m_2(X^2) - \dots - m_d(X^d)$  on  $X^1$ . Call the resulting estimator  $\tilde{m}_1(x^1)$ . Any estimator that has the same asymptotic distribution as  $\tilde{m}_1(x^1)$  is called oracle efficient. If such an estimator can be found, then asymptotically, there is no penalty for not knowing  $m_2, \dots, m_d$  and  $\mu$  when estimating  $m_1$ .

#### 3.1. Marginal integration

This section describes the marginal integration method for estimating  $\mu$  and the additive components  $m_j$  in (3.1). The data are assumed to consist of the simple random sample  $\{Y_i, X_i : i = 1, \dots, n\}$ . Observe that (3.1) is unchanged if each component  $m_j$  is replaced by  $m_j + \alpha_j$  for

some finite constant  $\alpha_j$  and  $\mu$  is replaced by  $\mu - \alpha_1 - \dots - \alpha_d$ . Therefore, a location normalization is needed to identify the  $m_j$ 's. In marginal integration, this is accomplished by setting

$$E[m_j(X^j)] = 0; j = 1, 2, \dots, d. \quad (3.2)$$

We now consider estimation of  $\mu$  and  $m_1$ . Other additive components can be estimated by swapping them with  $m_1$ .

Let  $X^{(-1)}$  be the vector consisting of all components of  $X$  except  $X^1$ . Let  $p_{-1}$  denote the probability density function of  $X^{(-1)}$ . Then (3.2) yields the following identifying relations (that is, mappings from the population distribution of the observable variables to  $\mu$  and  $m_1$ ):

$$\mu = E(Y) \quad (3.3)$$

and

$$m_1(x^1) = \int E(Y | X = x) p_{-1}(x^{(-1)}) dx^{(-1)} - \mu. \quad (3.4)$$

We can estimate  $\mu$  and  $m_1$  by replacing unknown population quantities on the right-hand sides of (3.3) and (3.4) with consistent estimators. This gives the following estimator of  $\mu$ :

$$\hat{\mu} = n^{-1} \sum_{i=1}^n Y_i.$$

To estimate  $m_1(x^1)$ , let  $\hat{g}(x^1, X^{(-1)})$  be a kernel or local linear estimator of  $E(Y | X^1 = x^1, X^{(-1)} = x^{(-1)})$ . Observe that the integral on the right-hand side of (3.4) is the average of  $E(Y | X^1 = x^1, X^{(-1)} = x^{(-1)})$  over  $X^{(-1)}$ . This can be estimated by the sample average of  $\hat{g}(x^1, X^{(-1)})$ . Thus, the estimator of  $m_1$  is

$$\hat{m}_1(x^1) = n^{-1} \sum_{i=1}^n \hat{g}(x^1, X_i^{(-1)}) - \hat{\mu},$$

where  $X_i^{(-1)}$  is the  $i$ 'th observation of  $X^{(-1)}$ .

The marginal integration estimator was first proposed by Linton and Nielsen (1995), who derived its asymptotic distributional properties for  $d = 2$ . The properties of the estimator for arbitrary but finite  $d$  have

been derived by Linton and Härdle (1996). They use the following kernel estimator for  $\hat{g}$  :

$$\hat{g}(x) = [\hat{P}(x)]^{-1} \sum_{i=1}^n Y_i K_1 \left( \frac{x^1 - X_i^1}{h_1} \right) K_2 \left( \frac{x^{(-1)} - X_i^{(-1)}}{h_2} \right), \quad (3.5)$$

where

$$\hat{P}(x) = \sum_{i=1}^n K_1 \left( \frac{x^1 - X_i^1}{h_1} \right) K_2 \left( \frac{x^{(-1)} - X_i^{(-1)}}{h_2} \right),$$

$K_1$  is a kernel function of a scalar argument,  $K_2$  is a kernel function of a  $d-1$  dimensional argument, and  $h_1$  and  $h_2$  are bandwidths. Linton and Härdle prove the following theorem.

**Theorem 3.1:** Assume that  $\text{Var}(Y|X=x) \equiv \sigma^2(x)$  is bounded and Lipschitz continuous.

The functions  $m_j$  are  $q$  times continuously differentiable for some integer  $q > d-1$ .

The density  $p_{-1}$  and the density of  $X$ ,  $p$ , are bounded away from 0 and  $q$  times continuously differentiable.

The kernel function  $K_1$  is bounded, non-negative, compactly supported, and Lipschitz continuous. It satisfies

$$\int_{-1}^1 K_1(z) dz = 1,$$

and

$$\int_{-1}^1 z K_1(z) dz = 0.$$

The kernel function  $K_2$  is bounded, compactly supported, and Lipschitz continuous. It satisfies

$$\int_{-1}^1 K_2(z) dz = 1$$

and

$$\int_{-1}^1 z^j K_2(z) dz = 0; \quad j = 1, \dots, q-1$$

The bandwidths satisfy  $h_1 = c_1 n^{-1/5}$  for some constant  $c_1 < \infty$ ,  $n^{2/5} h_2^q \rightarrow 0$  and  $n^{2/5} h_2^{d-1} \rightarrow \infty$  as  $n \rightarrow \infty$ .

Define

$$A = \int z^2 K_1(z) dz$$

and

$$B = \int [K_1(z)]^2 dz.$$

Then

$$n^{2/5} [\hat{m}_1(x^1) - m_1(x^1)] \rightarrow^d n[\beta_1(x^1), v_1(x^1)],$$

where

$$\beta_1 = c_1^2 A \left[ \frac{1}{2} m_1''(x^1) + m_1' \int \frac{\partial \log p(x)}{\partial x_1} p_{-1}(x^{(-1)}) dx^{(-1)} \right],$$

and

$$v_1(x^1) = c_1^{-1} B \int \sigma^2(x) \frac{p_{-1}^2(x^{(-1)})}{p(x)} dx^{(-1)}. \blacksquare$$

Under the assumptions of Theorem 3.1, the additive components  $m_j$  can be estimated with  $n^{-2/5}$  rates of convergence, but the number of derivatives that are needed to do this,  $(d-1)$ , is larger than the number needed in the one-dimensional case whenever  $d > 3$ . Thus, marginal integration suffers from a form of the curse of dimensionality in that more derivatives of the  $m_j$ 's and densities are needed to achieve the  $n^{-2/5}$  rate as  $d$  increases. In addition, the marginal integration estimator is not oracle efficient. For example, if  $\sigma^2(x) = \sigma^2$ , a constant, then

$$v_1(x^1) = c_1^{-1} B \sigma^2 \int \frac{p_{-1}^2(x^{(-1)})}{p(x)} dx^{(-1)}.$$

The variance of the oracle-efficient estimator that is obtained from the nonparametric regression of  $Y - m_2(X^2) - \dots - m_d(X^d)$  on  $X^1$  is  $c_1^{-1} B \sigma^2 / p_1(x^1) \leq v_1(x^1)$ , where  $p_1$  is the probability density function of  $X^1$  (Linton 1997).

Linton (1997) has shown that if  $d = 2$ , then an oracle efficient estimator can be obtained by taking one step from a suitable marginal integration estimator. Specifically, define  $\hat{U}_i = Y_i - \hat{\mu} - \hat{m}_2(X_i^2)$ . Then an oracle efficient estimator of  $m_1(x^1)$  can be obtained as the value of  $\beta_0$  in

$$(\beta_0, \beta_1) = \arg \min_{b_0, b_1} \sum_{i=1}^n [\hat{U}_i - b_0 - b_1'(X_i^1 - x^1)]^2 K_1\left(\frac{X_i^1 - x^1}{h_3}\right),$$

where  $h_3$  is a bandwidth parameter. Linton (1997) uses a local linear estimator for  $\hat{g}$  in (3.5). The bandwidths must satisfy  $nh_1^5 \rightarrow 0$ ,  $nh_2^5 \rightarrow 0$ ,  $nh_1h_2 \rightarrow \infty$ , and  $nh_3^5$  is bounded away from 0 and  $\infty$ . Fan, Mammen, and Härdle (1998) showed how to achieve oracle efficiency with arbitrary values of  $d$ . However, they require increasing smoothness of the additive components and density of  $X$  as  $d$  increases, so their method does not avoid the curse of dimensionality.

Kim, Linton, and Hengartner (1999) proposed a modified marginal integration estimator that achieves oracle efficiency if the additive components and densities have enough derivatives. The modified estimator is also easier to compute than the original marginal integration estimator is. To describe the modified estimator, let  $m_{-1} = m_2 + \dots + m_d$ , and let  $p_1$  denote the probability density function of  $X^1$ . Define

$$w(x) = \frac{p_1(x^1)p_{-1}(x^{(-1)})}{p(x)}.$$

Then

$$E[w(X)m_{-1}(X^{(-1)}) | X^1 = x^1] = 0.$$

Therefore,

$$\mu + m_1(x^1) = E[Yw(X) | X^1 = x^1], \quad (3.6)$$

and  $m_1(x^1)$  can be estimated by replacing the expectation and  $w$  on the right-hand side of (3.6) with sample analogs. The resulting estimate of  $m_1(x^1)$  is

$$\hat{m}_1(x^1) = \frac{1}{nh} \sum_{i=1}^n Y_i K\left(\frac{x^1 - X_i^1}{h}\right) \frac{\hat{p}_{-1}(X_i^{(-1)})}{\hat{p}(X_i)} - \hat{\mu},$$

where  $\hat{p}$  and  $\hat{p}_{-1}$  are kernel estimators of  $p$  and  $p_{-1}$ , respectively. Achieving oracle efficiency requires an additional step. Define

$$\hat{U}_i = Y_i - \hat{\mu} - \hat{m}_2(X_i^2) - \dots - \hat{m}_d(X_i^d),$$

where  $\hat{m}_j$  is the modified marginal integration estimator of  $m_j$  and

$$S = \{x \in \mathbb{R}^d : \underline{b}_j + h \leq x^j \leq \bar{b}_j - h; j = 1, \dots, d\},$$

where  $\underline{b}_j$  and  $\bar{b}_j$ , respectively are the lower and upper bounds of the support of  $x^j$ . The oracle efficient estimator of  $m_1(x^1)$  is  $\alpha_0$  in

$$(\alpha_0, \dots, \alpha_{q-1}) = \min_{a_0, \dots, a_{q-1}} \sum_{i=1}^n [\hat{U}_i - \sum_{j=0}^{q-1} a_j (X_i^1 - x^1)^j]^2 K_1\left(\frac{x^1 - X_i^1}{h_1}\right) I(X_i \in S),$$

where  $h_1$  is a bandwidth and  $q$  is the number of derivatives that the  $m_j$ 's and the densities have. Kim, et al. (1999) show that  $\alpha_0$  is an oracle efficient estimator of  $m_1$  (that is, it has the same asymptotic distribution that it would have if  $m_2, \dots, m_d$  and  $\mu$  were known) if the following conditions are satisfied.

The kernel is bounded, symmetrical about 0, supported on  $[-1, 1]$ , and satisfies

$$\int_{-1}^1 z^j K_1(z) dv = 0; \quad j = 1, \dots, q-1.$$

The functions  $m_j$  ( $j = 1, \dots, d$ ) and  $p$  are  $q-1$  times continuously differentiable, where  $q \geq (d-1)/2$ .

The density function  $p$  is bounded away from 0 and  $\infty$  and supported on  $[\underline{b}_1, \bar{b}_1] \times \dots \times [\underline{b}_d, \bar{b}_d]$ .

$\text{Var}(Y | X = x)$  is Lipschitz continuous and bounded away from 0 and  $\infty$ .

The bandwidths satisfy  $h_1 = cn^{-1/(2q+1)}$  for some positive constant  $c < \infty$  and  $h = o[n^{-1/(2q+1)}]$  as  $n \rightarrow \infty$ .

The result of Kim, et. al. (1999) shows that an oracle efficient estimator can be obtained from a marginal integration estimator if  $q$  is sufficiently large, but the curse of dimensionality remains. That is, the number of derivatives that the  $m_j$ 's and  $p$  must have increases as  $d$  increases. Secs. 3.2 and 3.3 describe methods that overcome the curse of dimensionality. The method described in Section 3.3 achieves oracle efficiency for any fixed  $q \geq 2$ , regardless of  $d$ .

Hengartner and Sperlich (2005) found a way to modify the marginal integration estimator so as to overcome the curse of dimensionality, though the resulting estimator is not oracle efficient. To describe their method, let  $m_{-1}(x^{(-1)}) = m_2(x^2) + \dots + m_d(x^d)$ . Let  $\pi_1$  and  $\pi_{-1}$  be sufficiently smooth density functions on  $\mathbb{R}$  and  $\mathbb{R}^{d-1}$ , respectively. Define  $\pi = \pi_1\pi_{-1}$ . Hengartner's and Sperlich's idea is to use the location normalization

$$\int m_1(x^1)\pi_1(x^1)dx^1 = 0 \quad (3.7)$$

and

$$\int m_{-1}(x^{(-1)})\pi_{-1}(x^{(-1)})dx^{(-1)} = 0 \quad (3.8)$$

instead of (3.2). The normalization (3.7)-(3.8) makes it possible to use the smoothness of  $\pi_1$  and  $\pi_{-1}$  to reduce the bias of the marginal integration estimator instead of using the smoothness of the  $m_j$ 's. Therefore, the  $m_j$ 's do not need to be as smooth as in marginal integration methods based on (3.2), thereby overcoming the curse of dimensionality.

Now let  $K_1$  be a kernel function of a scalar argument and  $K_2$  be a kernel function of a  $d-1$  dimensional argument. Let  $\hat{p}$  be a kernel estimator of  $p$ . Let  $h_1$  and  $h_2$  be bandwidths. Define

$$\tilde{g}(x) = \frac{1}{nh_1h_2^{d-1}} \sum_{i=1}^n \frac{Y_i}{\hat{p}(X_i)} K_1\left(\frac{x^1 - X_i^1}{h_1}\right) K_2\left(\frac{x^{(-1)} - X_i^{(-1)}}{h_2}\right).$$

Observe that  $\hat{g}(x)$  is a kind of kernel estimator of  $E(Y|X=x)$ . The estimator of  $m_1(x^1)$  is

$$\hat{m}_1(x^1) = \int \tilde{g}(x)q_{-1}(x^{-1})dx^{-1} - \int \tilde{g}(x)q(x)dx. \quad (3.9)$$

This estimator can be understood intuitively by observing that if  $g$  replaces  $\tilde{g}$  in (3.9), then the location normalization (3.7)-(3.8) implies that the right-hand side of (3.9) equals  $m_1(x^1)$ . The following theorem, which is proved in Hengartner and Sperlich (2005), gives the asymptotic behavior of the estimator (3.9).

**Theorem 3.2:** Assume that:

- (a) The conditional mean function  $g(x)$  is  $s$  times continuously differentiable in  $x^1$ . The conditional variance function  $\sigma^2(x)$  is finite and Lipschitz continuous.
- (b) The density function  $p$  is compactly supported, Lipschitz continuous, and bounded away from 0 and  $\infty$  in the interior of the support.
- (c) The density of  $X^{(-1)}$  conditional on  $X^1$  is bounded away from 0 everywhere in the support of  $X$ .
- (d) The density  $\pi$  is continuous and bounded away from 0 and  $\infty$  on its support, which is contained in the support of  $X$ . Moreover, the density  $\pi_1$  has  $s+1$  continuous, bounded derivatives.
- (e) The kernels  $K_1$  and  $K_2$  are compactly supported and Lipschitz continuous.  $K_1$  satisfies

$$\int_{-1}^1 z^j K_1(z) dz = 0; \quad j = 1, \dots, s-1$$

$$\int_{-1}^1 z^s K_1(z) dz = A > 0$$

$$\int_{-1}^1 K_1(z)^2 dz = B.$$

- (f) The bandwidths satisfy  $h_1 = n^{-1/(2s+1)}$ ,  $h_2 = o(1)$ , and  $nh_2^d \rightarrow \infty$  as  $n \rightarrow \infty$ .

Then

$$(nh_1)^{1/2}[\hat{m}_1(x^1) - m_1(x^1)] \rightarrow^d N[\beta(x^1), v_1(x^1)],$$

where

$$\begin{aligned} \beta_1(x^1) &= A \left[ \frac{1}{p_1(x^1)} \frac{d^2}{dx_1^s} m_1(x^1) - \int m_1(z) \frac{d^2}{dz^2} \pi_1(z) dz \right], \\ v_1(x^1) &= B \frac{\omega(x^1, x^1)}{p_1(x^1)}, \\ \omega(z^1, x^1) &= \int [\sigma^2(x) + g(x)^2] \frac{\pi_{-1}(x^{(-1)}) p_1(x^1)}{p(x)} dx^{(-1)} \\ &\quad - \left[ \int g(z) \frac{p(z) p_1(x^1)}{p_1(z^1) p(x^1, z^{(-1)})} \pi_{-1}(z^{(-1)}) dz^{(-1)} \right]^2, \end{aligned}$$

and  $p_1$  is the density of  $x^1$ . ■

Theorem 3.2 imposes no smoothness requirements on  $\pi_{-1}$ . Therefore, this density can be set equal to the Dirac delta function centered at any  $x^{(-1)}$ . This yields

$$m_1(x^1) = \tilde{g}(x^1, x^{(-1)}) - \int \tilde{g}(z, x^{(-1)}) \pi_1(z) dz$$

for any  $x^{(-1)}$  in the support of  $X^{(-1)}$ .

The assumptions of Theorem 3.2 do not require  $s$  to increase as  $d$  increases. Therefore, the estimator (3.9) avoids the curse of dimensionality. The estimator is not oracle efficient, however, and it is not known whether oracle efficiency can be obtained by taking an additional step as in Linton (1997) and Kim, et al. (1999). Hengartner and Sperlich (2005) do not explore how the density  $\pi$  should be chosen in applications.

The next two sections describe estimation methods that avoid the curse of dimensionality and, in some cases, achieve oracle efficiency. The method of Sec. 3.3 is extended in Secs. 4-5 to estimation of an additive model with a non-identity link function and estimation of a conditional quantile function.

### 3.2. Backfitting

Backfitting is an estimation procedure for model (3.1) that is implemented in many statistical software packages. To describe the procedure, define

$$W_i^j = Y_i - \mu - \sum_{k \neq j} m_k(X_i^k)$$

for  $j = 1, \dots, d$ . Write model (3.1) as

$$W_i^j = m_j(X_i^j) + U_i \quad (3.10)$$

If  $W_j^i$  were known, then oracle efficiency could be achieved by estimating  $m_j$  in (3.10) nonparametrically, but  $W_i^j$  is not known. To obtain a feasible estimator, let  $\hat{\mu}^0, \hat{m}_2^0, \dots, \hat{m}_d^0$  be preliminary estimates of  $\mu, m_2, \dots, m_d$ . Set

$$\hat{W}_{i,0}^1 = Y_i - \hat{\mu}^0 - \sum_{j=2}^d \hat{m}_j^0(X_i^j).$$

Backfitting now proceeds as follows.

1. Estimate  $m_1$  by nonparametric regression of  $\hat{W}_{i,0}^1$  on  $X_i^1$ . Denote the resulting estimate by  $\hat{m}_1^1$ .
2. Set  $\hat{W}_{i,1}^2 = Y_i - \hat{\mu}^0 - \hat{m}_1^1(X_i^1) - \sum_{j=3}^d \hat{m}_j^0(X_i^j)$ .
3. Estimate  $m_2$  by nonparametric regression of  $\hat{W}_{i,1}^2$  on  $X_i^2$ . Denote the resulting estimate by  $\hat{m}_2^1$ .
4. Set  $\hat{W}_{i,1}^3 = Y_i - \hat{\mu}^0 - \hat{m}_1^1(X_i^1) - \hat{m}_2^1(X_i^2) - \sum_{j=4}^d \hat{m}_j^0(X_i^j)$ .
5. Estimate  $m_3$  by nonparametric regression of  $\hat{W}_{i,1}^3$  on  $X_i^3$ . Denote the resulting estimate by  $\hat{m}_3^1$ .
6. Continue until all additive components have been estimated. Then return to step 1 but with  $\hat{m}_j^1$  ( $j = 2, \dots, d$ ) in place of  $\hat{m}_j^0$ .
7. Iterate steps 1-5 until convergence is achieved.

This estimation procedure was first proposed by Buja, Hastie, and Tibshirani (1989) and further developed by Hastie and Tibshirani (1990). Opsomer and Ruppert (1997) and Opsomer (2000) investigated the statistical properties of backfitting and found, among other things, that it is not oracle efficient. Linton, Mammen, and Nielsen (1999) found a

way to modify backfitting to achieve an estimator that is oracle efficient, asymptotically normal, and avoids the curse of dimensionality. However, this method is very complicated, so we do not describe it here. Instead, we describe a two-step estimator that is simpler, easier to implement, asymptotically normal, and oracle efficient. Moreover, as will be seen in Secs. 4 and 5, the two-step method can be extended easily to models with non-identity link functions and estimation of conditional quantile functions.

### 3.3. Two-step, oracle-efficient estimation

This section describes an estimation procedure that was developed by Horowitz and Mammen (2004). The procedure does not use  $d$ -dimensional nonparametric regression and, thereby, avoids the curse of dimensionality. Estimation takes place in two stages. In the first stage, ordinary least squares is used to obtain a series approximation to each  $m_j$ . The first stage procedure imposes the additive structure of (3.1), thereby avoiding the need for  $d$ -dimensional nonparametric estimation. This is what enables the Horowitz-Mammen estimator to avoid the curse of dimensionality. The first-stage estimates are inputs to the second stage. Let  $\tilde{\mu}, \tilde{m}_2, \dots, \tilde{m}_d$  denote the first-stage estimates of  $\mu, m_2, \dots, m_d$ . Then second-stage estimate of  $m_1$  is obtained by carrying out the kernel nonparametric regression of  $Y - \tilde{\mu} - \tilde{m}_2(X^2) - \dots - \tilde{m}_d(X^d)$  on  $X^1$ . One can also use a local linear estimator, which has better behavior near the boundaries of the support of  $X$  and adapts better to non-uniform designs (Fan and Gijbels 1996). See Horowitz and Mammen (2004) for details. In large samples, the second-stage estimator has the structure of an ordinary kernel estimator, so deriving its pointwise rate of convergence and asymptotic distribution is relatively easy.

#### 3.3.1. Informal description of the estimator

Assume that the support of  $X$  is  $\mathcal{X} \equiv [-1, 1]^d$ , and normalize  $m_1, \dots, m_d$  so that

$$\int_{-1}^1 m_j(v) dv = 0; \quad j = 1, \dots, d.$$

For any  $x \in \mathbb{R}^d$  define  $m(x) = m_1(x^1) + \dots + m_d(x^d)$ , where  $x^j$  is the  $j$ 'th component of  $x$ . Let  $\{\psi_k : k = 1, 2, \dots\}$  denote a basis for smooth functions on  $[-1, 1]$ . A precise definition of "smooth" and conditions that the basis functions must satisfy are given in Sec. 3.3.2. These conditions include:

$$\int_{-1}^1 \psi_k(v) dv = 0; \quad (3.11)$$

$$\int_{-1}^1 \psi_j(v) \psi_k(v) dv = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise;} \end{cases} \quad (3.12)$$

and

$$m_j(x^j) = \sum_{k=1}^{\infty} \theta_{jk} \psi_k(x^j) \quad (3.13)$$

for each  $j = 1, \dots, d$ , each  $x^j \in [0, 1]$ , and suitable coefficients  $\{\theta_{jk}\}$ . For any positive integer  $\kappa$ , define

$$\Psi_{\kappa}(x) = [1, \psi_1(x^1), \dots, \psi_{\kappa}(x^1), \psi_1(x^2), \dots, \psi_{\kappa}(x^2), \dots, \psi_1(x^d), \dots, \psi_{\kappa}(x^d)]'. \quad (3.14)$$

Then for  $\theta_{\kappa} \in \mathbb{R}^{\kappa d + 1}$ ,  $\Psi_{\kappa}(x)' \theta_{\kappa}$  is a series approximation to  $\mu + m(x)$ . Section 3 gives conditions that  $\kappa$  must satisfy. These require that  $\kappa \rightarrow \infty$  at an appropriate rate as  $n \rightarrow \infty$ .

To obtain the first-stage estimators of the  $m_j$ 's, let  $\{Y_i, X_i : i = 1, \dots, n\}$  be a random sample of  $(Y, X)$ . Let  $\hat{\theta}_{n\kappa}$  be the solution to the ordinary least squares estimation problem

$$\text{minimize: } S_{n\kappa}(\theta) \equiv n^{-1} \sum_{i=1}^n [Y_i - \Psi_{\kappa}(X_i)' \theta]^2,$$

where  $\Theta_{\kappa} \subset \mathbb{R}^{\kappa d + 1}$  is a compact parameter set. The first-stage series estimator of  $\mu + m(x)$  is

$$\tilde{\mu} + \tilde{m}(x) = \Psi_{\kappa}(x)' \hat{\theta}_{n\kappa},$$

where  $\tilde{\mu}$  is the first component of  $\hat{\theta}_{n\kappa}$ . The estimator of  $m_j(x^j)$  for any  $j=1,\dots,d$  and any  $x^j \in [0,1]$  is the product of  $[\psi_1(x^j), \dots, \psi_\kappa(x^j)]$  with the appropriate components of  $\hat{\theta}_\kappa$ . There is no curse of dimensionality in this estimator because all the estimated functions have scalar arguments.

To obtain the second-stage estimator of (say)  $m_1(x^1)$ , let  $\tilde{X}_i$  denote the  $i$ 'th observation of  $\tilde{X} \equiv (X^2, \dots, X^d)$ . Define  $\tilde{m}_{-1}(\tilde{X}_i) = \tilde{m}_2(X_i^2) + \dots + \tilde{m}_d(X_i^d)$ , where  $X_i^j$  is the  $i$ 'th observation of the  $j$ 'th component of  $X$  and  $\tilde{m}_j$  is the series estimator of  $m_j$ . Let  $K$  be a probability density function on  $[-1,1]$ , and define  $K_h(v) = K(v/h)$  for any real, positive constant  $h$ . Conditions that  $K$  and  $h$  must satisfy are given in Sec. 3.3.2. These include  $h \rightarrow 0$  at an appropriate rate as  $n \rightarrow \infty$ . The second-stage estimate of  $m_1(x^1)$  is

$$\hat{m}_1(x^1) = \left[ \sum_{i=1}^n K_h(x^1 - X_i^1) \right]^{-1} \sum_{i=1}^n [Y_i - \tilde{m}_{-1}(\tilde{X}_i)] K_h(x^1 - X_i^1).$$

Sec. 3.3.2 gives conditions under which  $\hat{m}_1(x^1) - m_1(x^1) = O_p(n^{-2/5})$  and  $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)]$  is asymptotically normally distributed for any finite  $d$  when and the  $m_j$ 's are twice continuously differentiable.

### 3.3.2. Asymptotic properties of the two-stage estimator

This section begins by stating the assumptions that are used to prove that the two-stage estimator is asymptotically normal and oracle efficient.

The following additional notation is used. For any matrix  $A$ , define the norm  $\|A\| = [\text{trace}(A'A)]^{1/2}$ . Define  $U = Y - \mu + m(X)$ ,  $V(x) = \text{Var}(U | X = x)$ ,  $Q_\kappa = E[\Psi_\kappa(X)\Psi_\kappa(X)']$ , and  $\Upsilon_\kappa = Q_\kappa^{-1} E\{[m(X)]^2 V(X)\Psi_\kappa(X)\Psi_\kappa(X)'\} Q_\kappa^{-1}$  whenever the latter quantity exists.  $Q_\kappa$  and  $\Upsilon_\kappa$  are  $d(\kappa) \times d(\kappa)$  positive semidefinite matrices, where  $d(\kappa) = \kappa d + 1$ . Let  $\lambda_{\kappa, \min}$  denote the smallest eigenvalue of  $Q_\kappa$ . Let  $Q_{\kappa, ij}$  denote the  $(i, j)$  element of  $Q_\kappa$ . Define  $\zeta_\kappa = \sup_{x \in \mathcal{X}} \|\Psi_\kappa(x)\|$ . Let  $\{\theta_{jk}\}$  be the coefficients of the series expansion (2.3). For each  $\kappa$  define

$$\theta_\kappa = (\mu, \theta_{11}, \dots, \theta_{1\kappa}, \theta_{21}, \dots, \theta_{2\kappa}, \dots, \theta_{d1}, \dots, \theta_{d\kappa})'.$$

The assumptions are:

A1: The data,  $\{(Y_i, X_i): i=1, \dots, n\}$ , are an *iid* random sample from the distribution of  $(Y, X)$ , and  $E(Y | X = x) = \mu + m(x)$  for almost every  $x \in \mathcal{X} \equiv [-1, 1]^d$ .

A2: (i) The support of  $X$  is  $\mathcal{X}$ . (ii) The distribution of  $X$  is absolutely continuous with respect to Lebesgue measure. (iii) The probability density function of  $X$  is bounded, bounded away from zero, and twice continuously differentiable on  $\mathcal{X}$ . (iv) There are constants  $c_V > 0$  and  $C_V < \infty$  such that  $c_V \leq \text{Var}(U | X = x) \leq C_V$  for all  $x \in \mathcal{X}$ . (v) There is a constant  $C_U < \infty$  such that  $E |U|^j \leq C_U^{j-2} j! E(U^2) < \infty$  for all  $j \geq 2$ .

A3: (i) There is a constant  $C_m < \infty$  such that  $|m_j(v)| \leq C_m$  for each  $j=1, \dots, d$  and all  $v \in [-1, 1]$ . (ii) Each function  $m_j$  is twice continuously differentiable on  $[-1, 1]$ .

A4: (i) There are constants  $C_Q < \infty$  and  $c_\lambda > 0$  such that  $|Q_{\kappa, ij}| \leq C_Q$  and  $\lambda_{\kappa, \min} > c_\lambda$  for all  $\kappa$  and all  $i, j=1, \dots, d(\kappa)$ . (ii) The largest eigenvalue of  $\Psi_\kappa$  is bounded for all  $\kappa$ .

A5: (i) The functions  $\{\psi_\kappa\}$  satisfy (2.1) and (2.2). (ii) There is a constant  $c_\kappa > 0$  such that  $\zeta_\kappa \geq c_\kappa$  for all sufficiently large  $\kappa$ . (iii)  $\zeta_\kappa = O(\kappa^{1/2})$  as  $\kappa \rightarrow \infty$ . (iv) There are a constant  $C_\theta < \infty$  and vectors  $\theta_{\kappa 0} \in \Theta_\kappa \equiv [-C_\theta, C_\theta]^{d(\kappa)}$  such that  $\sup_{x \in \mathcal{X}} |\mu + m(x) - \Psi_\kappa(x)' \theta_{\kappa 0}| = O(\kappa^{-2})$  as  $\kappa \rightarrow \infty$ . (v) For each  $\kappa$ ,  $\theta_\kappa$  is an interior point of  $\Theta_\kappa$ .

A6: (i)  $\kappa = C_\kappa n^{4/15+\nu}$  for some constant  $C_\kappa$  satisfying  $0 < C_\kappa < \infty$  and some  $\nu$  satisfying  $0 < \nu < 1/30$ . (ii)  $h = C_h n^{-1/5}$  for some constant  $C_h$  satisfying  $0 < C_h < \infty$ .

A7: The function  $K$  is a bounded, continuous probability density function on  $[-1, 1]$  and is symmetrical about 0.

The assumption that the support of  $X$  is  $[-1, 1]^d$  entails no loss of generality as it can always be satisfied by carrying out monotone increasing transformations of the components of  $X$ , even if their support before transformation is unbounded. For practical computations, it suffices to transform the empirical support to  $[-1, 1]^d$ . Assumption A2 precludes the possibility of treating discrete covariates, though they can be handled inelegantly by conditioning on them. Differentiability of the density of  $X$  (Assumption A2(iii)) is used to insure that the bias of our estimator converges to zero sufficiently rapidly. Assumption A2(v) restricts the thickness of the tails of the distribution of  $U$  and is used to

prove consistency of the first-stage estimator. Assumption A3 defines the sense in which  $m_j$ 's must be smooth. A4 insures the existence and non-singularity of the covariance matrix of the asymptotic form of the first-stage estimator. This is analogous to assuming that the information matrix is positive definite in parametric maximum likelihood estimation. Assumption A4(i) implies A4(ii) if  $U$  is homoskedastic. Assumptions A5(iii) and A5(iv) bound the magnitudes of the basis functions and insure that the errors in the series approximations to the  $m_j$ 's converge to zero sufficiently rapidly as  $\kappa \rightarrow \infty$ . These assumptions are satisfied by spline and (for periodic functions) Fourier bases. Assumption A6 states the rates at which  $\kappa \rightarrow \infty$  and  $h \rightarrow 0$  as  $n \rightarrow \infty$ . The assumed rate of convergence of  $h$  is asymptotically optimal for one-dimensional kernel mean-regression when the conditional mean function is twice continuously differentiable. The required rate for  $\kappa$  insures that the asymptotic bias and variance of the first-stage estimator are sufficiently small to achieve an  $n^{-2/5}$  rate of convergence in the second stage. The  $L_2$  rate of convergence of a series estimator of  $m_j$  is maximized by setting  $\kappa \propto n^{1/5}$ , which is slower than the rates permitted by A6(i) (Newey (1997)). Thus, A6(i) requires the first-stage estimator to be undersmoothed. Undersmoothing is needed to insure sufficiently rapid convergence of the bias of the first-stage estimator. We show that the first-order performance of our second-stage estimator does not depend on the choice of  $\kappa$  if A6(i) is satisfied. See Theorem 3.4. Optimizing the choice of  $\kappa$  would require a rather complicated higher-order theory and is beyond the scope of this paper, which is restricted to first-order asymptotics.

We now state two theorems that give the asymptotic properties of the two-stage estimator. Theorem 3.3 gives the asymptotic behavior of the first-stage series estimator under assumptions A1-A6(i). Theorem 3.4 gives the properties of the second-stage estimator. For  $i = 1, \dots, n$ , define  $U_i = Y_i - \mu + m(X_i)$  and  $b_{\kappa 0}(x) = \mu + m(x) - \Psi_{\kappa}(x)' \theta_{\kappa 0}$ . Let  $\|v\|$  denote the Euclidean norm of any finite-dimensional vector  $v$ .

**Theorem 3.3:** Let A1-A6(i) hold. Then

$$(a) \quad \lim_{n \rightarrow \infty} \left\| \hat{\theta}_{n\kappa} - \theta_{\kappa 0} \right\| = 0$$

almost surely,

$$(b) \quad \hat{\theta}_{n\kappa} - \theta_{\kappa 0} = O_p(\kappa^{1/2}/n^{1/2} + \kappa^{-2}),$$

and

$$(c) \quad \sup_{x \in \mathcal{X}} |\tilde{m}(x) - m(x)| = O_p(\kappa/n^{1/2} + \kappa^{-3/2}).$$

In addition,

$$(d) \quad \hat{\theta}_{n\kappa} - \theta_{\kappa 0} = n^{-1} Q_{\kappa}^{-1} \sum_{i=1}^n \Psi_{\kappa}(X_i) U_i \\ + n^{-1} Q_{\kappa}^{-1} \sum_{i=1}^n \Psi_{\kappa}(X_i) b_{\kappa}(X_i) + R_n,$$

where  $\|R_n\| = O_p(\kappa^{3/2}/n + n^{-1/2})$ . ■

Now let  $f_X$  and  $f_1$ , respectively, denote the probability density function of  $X$  and  $X^1$ . For  $j = 0, 1$ , define

$$S'_{n1}(x^1, m) = -2 \sum_{i=1}^n [Y_i - \mu - m_1(x^1) - m_{-1}(\tilde{X}_i)] K_h(x^1 - X_i^1).$$

Also define

$$A_K = \int_{-1}^1 v^2 K(v) dv,$$

$$B_K = \int_{-1}^1 K(v)^2 dv,$$

$$g(x^1, \tilde{x}) =$$

$$(\partial^2 / \partial \zeta^2) \{ [m_1(\zeta + x^1) + m_{-1}(\tilde{x})] - [m_1(x^1) + m_{-1}(\tilde{x})] \} f_X(\zeta + x^1, \tilde{x}) \Big|_{\zeta=0},$$

and

$$\beta_1(x^1) = C_h^2 A_K f_1(x^1)^{-1} \int g(x^1, \tilde{x}) f_X(x^1, \tilde{x}) d\tilde{x}.$$

**Theorem 3.4:** Let A1-A6 hold. Then

- (a)  $\hat{m}_1(x^1) - m_1(x^1) = -[2nhf_1(x^1)]^{-1} S'_{n01}(x^1, m) + o_p(n^{-2/5})$  uniformly over  $|x^1| \leq 1-h$  and  $\hat{m}_1(x^1) - m_1(x^1) = O_p[(\log n)^{1/2} n^{-2/5}]$  uniformly over  $|x^1| \leq 1$ .
- (b)  $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)] \rightarrow^d N[\beta_1(x^1), V_1(x^1)]$ .
- (c) If  $j \neq 1$ , then  $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)]$  and  $n^{2/5}[\hat{m}_j(x^j) - m_j(x^j)]$  are asymptotically independently normally distributed. ■

Theorem 3.4(a) implies that asymptotically,  $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)]$  is not affected by random sampling errors in the first stage estimator. In fact, the second-stage estimator of  $m_1(x^1)$  has the same asymptotic distribution that it would have if  $m_2, \dots, m_d$  were known and local linear estimation were used to estimate  $m_1(x^1)$  directly. This is the oracle efficiency property. Parts (b) and (c) of Theorem 3.4 imply that the estimators of  $m_1(x^1), \dots, m_d(x^d)$  are asymptotically independently normally distributed.

### 3.3.3. Bandwidth selection

This section describes a data-based method for selecting the second-stage bandwidth  $h$  in the second estimation stage. It simultaneously estimates the bandwidths for estimating all the functions  $m_j$  ( $j=1, \dots, d$ ). In general, the bandwidth can be different for different  $m_j$ 's. Accordingly, denote the bandwidth for estimating  $m_j$  by  $h_j$ . Assume that  $h_j = C_j n^{-1/5}$  for some finite constant  $C_j > 0$ . The method described here selects the  $C_j$ 's to minimize an estimate of the average squared estimation error (ASE), which is

$$ASE(\bar{h}) = n^{-1} \sum_{i=1}^n \{[\tilde{\mu} + \hat{m}(X_i) - [\mu + m(X_i)]]\}^2$$

where  $\bar{h} = (C_1 n^{-1/5}, \dots, C_d n^{-1/5})$ . Specifically, the method selects the  $C_j$ 's to

$$\begin{aligned} \text{minimize: } PLS(\bar{h}) = n^{-1} \sum_{i=1}^n \{Y_i - [\tilde{\mu} + \hat{m}(X_i)]\}^2 \\ \text{over } C_1, \dots, C_d \\ + 2K(0)n^{-1} \sum_{i=1}^n \hat{V}(X_i) \sum_{j=1}^d [n^{4/5} C_j \hat{D}_j(X_i^j)]^{-1}, \end{aligned} \quad (3.15)$$

where the  $C_j$ 's are restricted to a compact, positive interval that excludes 0,

$$\hat{D}_j(x^j) = \frac{1}{nh_j} \sum_{i=1}^n K_{h_j}(X_i^j - x^j),$$

and

$$\begin{aligned} \hat{V}(x) = & \left[ \sum_{i=1}^n K_{h_1}(X_i^1 - x^1) \dots K_{h_d}(X_i^d - x^d) \right]^{-1} \\ & \times \sum_{i=1}^n K_{h_1}(X_i^1 - x^1) \dots K_{h_d}(X_i^d - x^d) \{Y_i - [\tilde{\mu} + \hat{m}(X_i)]\}^2. \end{aligned}$$

The bandwidths used for  $\hat{V}$  may be different from those used for  $\hat{m}$  because  $\hat{V}$  is a full dimensional nonparametric estimator. Horowitz and Mammen (2004) show that the solution to (3.15) consistently estimates the bandwidths that minimize ASE.

#### 4. Estimation with a Non-Identity Link Function

This section extends the two-stage method of Section 3.3 to estimation of the model

$$E(Y | X = x) = F[\mu + m_1(x^1) + \dots + m_d(x^d)], \quad (4.1)$$

where  $F$  is a known function, called a link function, that is not necessarily the identity function. As in the case of an identity link function, the first estimation stage consists of obtaining a series estimator of  $F$ . The additive structure is imposed in this stage, thereby avoiding the curse of dimensionality. The second estimation stage consists of taking one Newton step from the first-stage estimate toward a local linear

or local constant estimate. Here, we describe only the local constant estimator. Horowitz and Mammen (2004) explain the local linear estimator. In large samples, the second-stage estimator has a structure similar to that of a local linear or constant estimate, so its asymptotic distribution can be obtained. In particular, the second-stage estimate is oracle efficient.

We now describe the first-stage estimator. Let  $\{\psi_k : k=1,2,\dots\}$  denote a basis for smooth functions on  $[-1,1]$ . Assume that  $\{\psi_k\}$  satisfies (3.11)-(3.13). Define  $\Psi_\kappa$  as in (3.14). Let  $\{Y_i, X_i : i=1,\dots,n\}$  be a random sample of  $(Y, X)$ . To obtain the first-stage estimator, let  $\hat{\theta}_{n\kappa}$  be a solution to

$$\underset{\theta \in \Theta_\kappa}{\text{minimize:}} \quad S_{n\kappa}(\theta) \equiv n^{-1} \sum_{i=1}^n \{Y_i - F[\Psi_\kappa(X_i)' \theta]\}^2,$$

where  $\Theta_\kappa \subset \mathbb{R}^{kd+1}$  is a compact parameter set. Thus, first-stage estimation with a non-identity link function is like estimation with an identity link function except that nonlinear least squares is used instead of ordinary least squares. As with an identity link function, the series estimator of  $\mu + m(x)$  is  $\tilde{\mu} + \tilde{m}(x) = \Psi_\kappa(x)' \hat{\theta}_{n\kappa}$ , where  $\tilde{\mu}$  is the first component of  $\hat{\theta}_{n\kappa}$ , and the estimator of  $m_j(x^j)$  for any  $j=1,\dots,d$  and any  $x^j \in [0,1]$  is the product of  $[\psi_1(x^j), \dots, \psi_\kappa(x^j)]$  with the appropriate components of  $\hat{\theta}_{n\kappa}$ .

We now describe the second-stage estimator of (say)  $m_1(x^1)$ . As in Sec. 3.3, let  $\tilde{X}_i$  denote the  $i$ 'th observation of  $\tilde{X} \equiv (X^2, \dots, X^d)$ , and define  $\tilde{m}_{-1}(\tilde{X}_i) = \tilde{m}_2(X_i^2) + \dots + \tilde{m}_d(X_i^d)$ , where  $\tilde{m}_j$  is the series estimator of  $m_j$ . Let  $K$  and  $h$  be the kernel and bandwidth, respectively. Define

$$\begin{aligned} S'_{n1}(x^1, \tilde{m}) = & -2 \sum_{i=1}^n \{Y_i - F[\tilde{\mu} + \tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)]\} \\ & \times F'[\tilde{\mu} + \tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)] K_h(x^1 - X_i^1) \end{aligned}$$

and

$$\begin{aligned}
S_{n1}''(x^1, \tilde{m}) &= 2 \sum_{i=1}^n F'[\tilde{\mu} + \tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)]^2 K_h(x^1 - X_i^1) \\
&\quad - 2 \sum_{i=1}^n \{Y_i - F[\tilde{\mu} + \tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)]\} \\
&\quad \times F''[\tilde{\mu} + \tilde{m}_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i)] K_h(x^1 - X_i^1)
\end{aligned}$$

The second-stage estimator of  $m_1(x^1)$  is

$$\hat{m}_1(x^1) = \tilde{m}_1(x^1) - S'_{n01}(x^1, \tilde{m}) / S''_{n01}(x^1, \tilde{m}). \quad (4.2)$$

The second stage estimators of  $m_2(x^2), \dots, m_d(x^d)$  are obtained similarly.

The estimator (4.2) can be understood intuitively by observing that if  $\tilde{\mu}$  and  $\tilde{m}_{-1}$  were the true values of  $\mu$  and  $m_{-1}$ , then  $m_1(x^1)$  could be estimated by the value of  $b$  that minimizes

$$S_{n1}(x^1, b) = \sum_{i=1}^n \{Y_i - F[\tilde{\mu} + b + \tilde{m}_{-1}(\tilde{X}_i)]\}^2 K_h(x^1 - X_i^1) \quad (4.3)$$

The estimator (4.2) is the result of taking one Newton step from the starting value  $b_0 = \tilde{m}_1(x^1)$ , toward the minimum of the right-hand side of (4.2).

Describing the asymptotic distributional properties of the second-stage estimator requires modifying the notation and assumptions of Section 3.3. As before, define  $U = Y - F[\mu + m(X)]$  and  $V(x) = \text{Var}(U | X = x)$ . Define  $A_K$  and  $B_K$  as in Section 3.3. Redefine

$$Q_\kappa = \mathbf{E}\{F'[\mu + m(X)]^2 \Psi_\kappa(X) \Psi_\kappa(X)'\},$$

and

$$\Upsilon_\kappa = Q_\kappa^{-1} \mathbf{E}\{F'[\mu + m(X)]^2 V(X) \Psi_\kappa(X) \Psi_\kappa(X)'\} Q_\kappa^{-1}.$$

Define  $\lambda_{\kappa, \min}$ ,  $Q_{\kappa, ij}$ ,  $\zeta_\kappa$ ,  $\{\theta_{jk}\}$ , and  $\theta_\kappa$  as in Section 3.2. In addition, define

$$\begin{aligned}
S'_{n1}(x^1, m) &= -2 \sum_{i=1}^n \{Y_i - F[\mu + m_1(x^1) + m_{-1}(\tilde{X}_i)]\} \\
&\quad \times F'[\mu + m_1(x^1) + m_{-1}(\tilde{X}_i)] K_h(x^1 - X_i^1), \\
D_0(x^1) &= 2 \int F'[\mu + m_1(x^1) + m_{-1}(\tilde{x})]^2 f_X(x^1, \tilde{x}) d\tilde{x}, \\
g(x^1, \tilde{x}) &= (\partial^2 / \partial \zeta^2) \{F[m_1(\zeta + x^1) + m_{-1}(\tilde{x})] \\
&\quad - F[m_1(x^1) + m_{-1}(\tilde{x})]\} f_X(\zeta + x^1, \tilde{x}) \Big|_{\zeta=0}, \\
\beta_1(x^1) &= 2C_h^2 A_K D_0(x^1)^{-1} \int g(x^1, \tilde{x}) F'[\mu + m_1(x^1) + m_{-1}(\tilde{x})] f_X(x^1, \tilde{x}) d\tilde{x}.
\end{aligned}$$

and

$$\begin{aligned}
V_1(x^1) &= \\
&\quad B_K C_h^{-1} D_0(x^1)^{-2} \int \text{Var}(U | x^1, \tilde{x}) F'[\mu + m_1(x^1) + m_{-1}(\tilde{x})]^2 f_X(x^1, \tilde{x}) d\tilde{x}.
\end{aligned}$$

Make the following assumptions in addition to those already made in Section 3.3.

A3: (iii) There are constants  $C_{F1} < \infty$ ,  $c_{F2} > 0$ , and  $C_{F2} < \infty$  such that  $F(v) \leq C_{F1}$  and  $c_{F2} \leq F'(v) \leq C_{F2}$  for all  $v \in [\mu - C_m d, \mu + C_m d]$ . (iv)  $F$  is twice continuously differentiable on  $[\mu - C_m d, \mu + C_m d]$ . (v) There is a constant  $C_{F3} < \infty$  such that  $|F''(v_2) - F''(v_1)| \leq C_{F3} |v_2 - v_1|$  for all  $v_2, v_1 \in [\mu - C_m d, \mu + C_m d]$ .

These assumptions impose smoothness restrictions on the link function  $F$ . They are satisfied automatically and, therefore, not needed if  $F$  is the identity function.

The properties of the first-stage estimator are given by the following theorem.

**Theorem 4.1:** Let A1-A6(i) hold. Then

$$(a) \lim_{n \rightarrow \infty} \left\| \hat{\theta}_{nk} - \theta_{k0} \right\| = 0$$

almost surely,

$$(b) \hat{\theta}_{n\kappa} - \theta_{\kappa 0} = O_p(\kappa^{1/2}/n^{1/2} + \kappa^{-2}),$$

and

$$(c) \sup_{x \in \mathcal{X}} |\tilde{m}(x) - m(x)| = O_p(\kappa/n^{1/2} + \kappa^{-3/2}).$$

In addition,

$$(d) \hat{\theta}_{n\kappa} - \theta_{\kappa 0} = n^{-1} Q_{\kappa}^{-1} \sum_{i=1}^n F'[\mu + m(X_i)] \Psi_{\kappa}(X_i) U_i \\ + n^{-1} Q_{\kappa}^{-1} \sum_{i=1}^n F'[\mu + m(X_i)]^2 \Psi_{\kappa}(X_i) b_{\kappa}(X_i) + R_n$$

where  $\|R_n\| = O_p(\kappa^{3/2}/n + n^{-1/2})$ . ■

The properties of the second-stage estimator are given by the next theorem.

**Theorem 4.2:** Let A1-A6 hold. Then

$$(a) \hat{m}_1(x^1) - m_1(x^1) = -[nhD_0(x^1)]^{-1} S'_{n1}(x^1, m) + o_p(n^{-2/5}), \quad \text{uniformly} \\ \text{over } |x^1| \leq 1-h \quad \text{and} \quad \hat{m}_1(x^1) - m_1(x^1) = O_p[(\log n)^{1/2} n^{-2/5}] \\ \text{uniformly over } |x^1| \leq 1.$$

$$(b) n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)] \rightarrow^d N[\beta_1(x^1), V_1(x^1)].$$

(c) If  $j \neq 1$ , then  $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)]$  and  $n^{2/5}[\hat{m}_j(x^j) - m_j(x^j)]$  are asymptotically independently normally distributed. ■

Theorem 4.2a implies that the second-stage estimator is oracle efficient.

## 5. Estimating a Conditional Quantile Function

This section is concerned with estimation of the unknown functions  $m_j$  in the model

$$Y = \mu + m_1(x^1) + \dots + m_d(x^d) + U_{\alpha}, \quad (5.1)$$

where  $U_\alpha$  is an unobserved random variable whose  $\alpha$  quantile conditional on  $X = x$  is zero for almost every  $x$ . Existing estimation methods include series estimation (Doksum and Koo 2000) and backfitting (Fan and Gijbels), but the rates of convergence and other asymptotic distributional properties of these estimators are unknown. De Gooijer and Zerom (2003) proposed a marginal integration estimator that is asymptotically normally distributed, but it begins with a  $d$ -dimensional nonparametric quantile regression and, therefore, suffers from a curse of dimensionality.

This section describes a two-stage estimator that was developed by Horowitz and Lee (2005). The estimator is asymptotically normal, oracle efficient, and has no curse of dimensionality. The estimator is similar in concept to the method of Horowitz and Mammen (2004) that is described in Sec. 3.3. However, there are enough differences between estimation of conditional quantile and conditional mean functions to require a separate treatment of quantile estimation.

Horowitz and Lee (2005) assume that the support of  $X$  is  $[-1,1]^d$  and use the location normalization

$$\int_{-1}^1 m_j(v)dv = 0.$$

Define the function  $\rho_\alpha(u) = |u| + (2\alpha - 1)u$  for  $0 < \alpha < 1$ . As in Sec. 3.2, the first stage in estimating a conditional quantile function is estimating the coefficients of a series approximation to  $\mu + m(x)$ . As in estimation of a conditional mean function, this is done by solving an optimization problem, but the objective function is different in conditional quantile estimation. Specifically let  $\hat{\theta}_{n\kappa}$  be a solution to

$$\underset{\theta}{\text{minimize}} S_{n\kappa}(\theta) = n^{-1} \sum_{i=1}^n \rho_\alpha[Y_i - \Psi_\kappa(X_i)' \theta],$$

where  $\Psi_\kappa$  is defined as in Section 3.2. The first-stage series estimator of  $\mu + m(x)$  is

$$\tilde{\mu} + \tilde{m}(x) = \Psi_\kappa(x)' \hat{\theta}_{n\kappa},$$

where  $\tilde{\mu}$  is the first component of  $\hat{\theta}_{n\kappa}$ .

To describe the second stage estimator of (say)  $m_1$ , assume that  $m_1$  is twice continuously differentiable on  $[-1,1]$ . The second stage then consists of local linear estimation. Specifically, using the notation of Section 3.2, the estimator of  $m_1(x^1)$  is defined as  $\hat{m}_1(x^1) = \hat{b}_0$ , where  $\hat{b}_n = (\hat{b}_0, \hat{b}_1)$  minimizes

$$S_n(b) = (nh)^{-1} \sum_{i=1}^n \rho_\alpha[Y_i - \tilde{\mu} - b_0 - b_1(X_i^1 - x^1) - \tilde{m}_{-1}(\tilde{X}_i)] K_h(X_i^1 - x^1).$$

Because quantiles of monotone transformations of  $Y$  are equal to monotone transformations of quantiles of  $Y$ , it is straightforward to extend the estimator of Horowitz and Lee to a model of the form

$$G(Y) = \mu + m_1(x^1) + \dots + m_d(x^d) + U_\alpha,$$

where  $G$  is a known, strictly increasing function and the  $\alpha$  quantile of  $U_\alpha$  conditional on  $X = x$  is zero. Estimation of the  $m_j$ 's can be carried out by replacing  $Y$  with  $G(Y)$  in the two-stage procedure. The  $\alpha$  quantile of  $Y$  conditional on  $X = x$  is estimated by  $G^{-1}[\tilde{\mu} + \hat{m}_1(x^1) + \dots + m_d(x^d)]$ .

Horowitz and Lee (2005) make the following assumptions to obtain the asymptotic properties of the two-stage estimator.

A1: The data,  $\{Y_i, X_i : i = 1, \dots, n\}$  are iid, and the  $\alpha$  quantile of  $Y$  conditional on  $X = x$  is  $\mu + m(x)$  for almost every  $x$ .

A2: The support of  $X$  is  $\mathcal{X} = [-1,1]^d$ . The distribution of  $X$  is absolutely continuous with respect to Lebesgue measure. The probability density function of  $X$ , denoted by  $f_X(x)$ , is bounded, bounded away from 0, twice differentiable in the interior of  $\mathcal{X}$ , and has continuous one-sided second derivatives at the boundary of  $\mathcal{X}$ .

A3: Let  $F(u|x)$  denote the distribution function of  $U_\alpha$  conditional on  $X = x$ . Then  $F(0|x) = \alpha$  for almost every  $x \in \mathcal{X}$ , and  $F(\cdot|x)$  has a probability density function  $f(\cdot|x)$ . There is a constant  $L_f < \infty$  such that  $|f(u_1|x) - f(u_2|x)| \leq L_f |u_1 - u_2|$  for all  $u_1$  and  $u_2$  in a neighborhood of 0 and all  $x \in \mathcal{X}$ . There are constants  $c_f > 0$  and  $C_f < \infty$  such that  $c_f \leq f(u|x) \leq C_f$  for all  $u$  in a neighborhood of 0 and all  $x \in \mathcal{X}$ .

A4: For each  $j = 1, \dots, d$ ,  $m_j$  is twice continuously differentiable in the interior of  $[-1, 1]$  and has continuous, one-sided second derivatives at the boundaries of  $[-1, 1]$ .

A5: Define  $\Phi_\kappa = E[f(0|X)\Psi_\kappa(X)\Psi_\kappa(X)']$ . The smallest eigenvalue of  $\Phi_\kappa$  is bounded away from 0 for all  $\kappa$ , and the largest eigenvalue is bounded for all  $\kappa$ .

A6: The basis functions satisfy (3.11) and (3.12). Moreover,  $\zeta_\kappa = O(\kappa^{1/2})$ , and  $\sup_{x \in \mathcal{X}} |\mu + m(x) - \Psi_\kappa(x)' \theta_{\kappa 0}| = O(\kappa^{-2})$ .

A7. (i)  $\kappa = C_\kappa n^\nu$  for some constant  $C_\kappa$  satisfying  $0 < C_\kappa < \infty$  and some  $\nu$  such that  $1/5 < \nu < 7/30$ . (ii) The bandwidth  $h = C_h n^{-1/5}$  for some finite, positive constant  $C_h$ .

A8: The kernel function  $K$  is a bounded, continuous probability density function on  $[-1, 1]$  and is symmetrical about 0.

Now define

$$\bar{\Psi}_\kappa(\tilde{x}) = [1, \underbrace{0, \dots, 0}_\kappa, \psi_1(x^2), \dots, \psi_\kappa(x^2), \dots, \psi_1(x^d), \dots, \psi_\kappa(x^d)]',$$

where  $\tilde{x} = (x^2, \dots, x^d)$ .

A9: The largest eigenvalue of  $E[\bar{\Psi}_\kappa(\tilde{X})\bar{\Psi}_\kappa(\tilde{X})' | X^1 = x^1]$  is bounded for all  $\kappa$  and each component of  $E[\bar{\Psi}_\kappa(\tilde{X})\bar{\Psi}_\kappa(\tilde{X})' | X^1 = x^1]$  is twice continuously differentiable with respect to  $x^1$ .

These assumptions are similar to those of the two-step estimator of a conditional mean function that is described in Section 3.3.

For  $j = 0, 1, 2$  define

$$\rho_j = \int_{-1}^1 v^j K(v) dv.$$

Let  $S_K$  be the  $2 \times 2$  matrix whose  $(i, j)$  component is  $\rho_{i+j-2}$ . Also, define  $e_1 = (1, 0)'$ . Set  $K_*(u) = e_1' S_K^{-1} (1, u)' K(u)$ . Let  $f_{X^1}$  denote the probability density function of  $X^1$ , and let  $f_1(u | x^1)$  denote the probability density function of  $U_\alpha$  conditional on  $X^1 = x^1$ . Finally, define

$$\beta_1(x^1) = 0.5 C_h^2 \left[ \int_{-1}^1 v^2 K_*(v) dv \right] m_1''(x^1)$$

and

$$V_1(x^1) = \left[ \int_{-1}^1 K_*(v)^2 dv \right] C_h^{-1} \alpha(1-\alpha) / [f_{X^1}(x^1) f_1(0|x^1)^2].$$

The main result of Horowitz and Lee (2005) is given by the following theorem.

**Theorem 5.1:** Let assumptions A1-A9 hold. Then as  $n \rightarrow \infty$  and any  $x^1$  satisfying  $|x^1| \leq 1-h$ , the following results hold:

- (a)  $|\hat{m}_1(x^1) - m_1(x^1)| = O_p(n^{-2/5})$
- (b)  $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)] \rightarrow^d N[\beta_1(x^1), V_1(x^1)]$
- (c) If  $j \neq 1$ , then  $n^{2/5}[\hat{m}_1(x^1) - m_1(x^1)]$  and  $n^{2/5}[\hat{m}_j(x^j) - m_j(x^j)]$  are asymptotically independently normally distributed for any  $x^j$  satisfying  $|x^j| \leq 1-h$ . ■

This theorem implies that the second-stage estimator achieves the optimal rate of convergence for a nonparametric estimator of a twice differentiable function. Only two derivatives are needed regardless of  $d$ , so there is no curse of dimensionality. Moreover, the second-stage estimator is oracle efficient. That is, it has the same asymptotic distribution as it would have if  $m_2, \dots, m_d$  were known.

## 6. An Empirical Example

This section illustrates the use of the estimator of Horowitz and Mammen (2004) by using it to estimate an earnings function. The specification of the model is

$$\log W = m_{EXP}(EXP) + m_{EDUC}(EDUC) + U,$$

where  $W$  is an individual's wage;  $EXP$  and  $EDUC$ , respectively, are the number of years of work experience and education that the individual has had; and  $U$  is an unobserved random variable satisfying  $E(U | EXP, EDUC) = 0$ . The functions  $m_{EXP}$  and  $m_{EDUC}$  are unknown and are estimated by the Horowitz-Mammen procedure. The data are taken from the 1993 Current Population Survey and consist of

observations on 3123 individuals. The estimation results are shown in Figure 1. The estimates of  $m_{EXP}$  (Figure 1a) and  $m_{EDUC}$  (Figure 1b) are nonlinear and differently shaped. They are not well approximated by simple parametric functions such as quadratic functions. A lengthy specification search might be needed to find a parametric model that produces the shapes shown in Fig. 1.

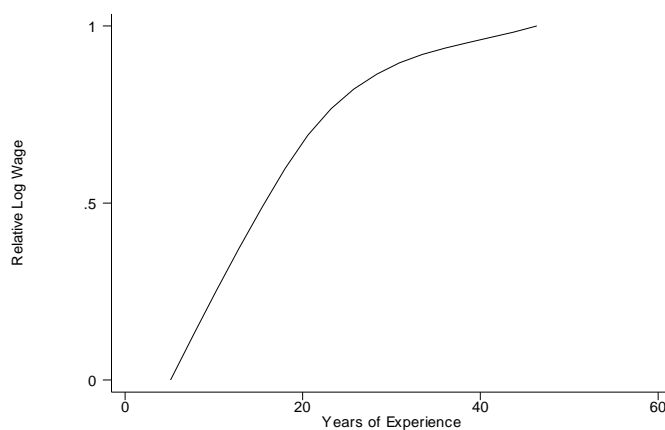


Fig. 1a: Nonparametric Estimate of  $m_{EXP}$

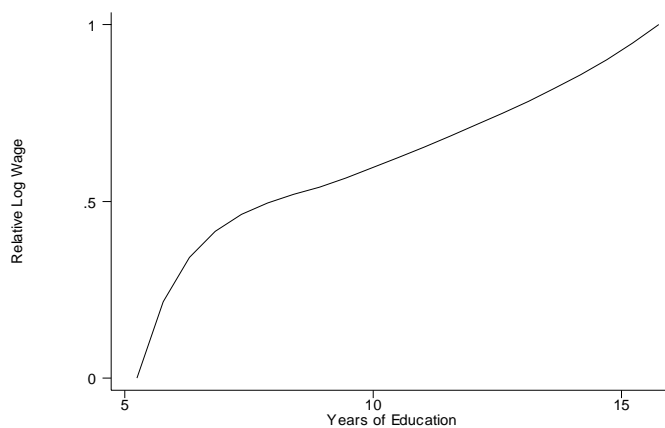


Fig 1b: Nonparametric Estimate of  $m_{EDUC}$

## References

1. Buja, A., T.J. Hastie, and R.J. Tibshirani (1989). Linear Smoothers and Additive Models, *Annals of Statistics*, 17, 453-555.
2. De Gooijer, J.G. and D. Zerom (2003). On Additive Conditional Quantiles with High-Dimensional Covariates, *Journal of the American Statistical Association*, 98, 135-146.
3. Doksum, K. and J.-Y. Koo (2000). On Spline Estimators and Prediction Intervals in Nonparametric Regression, *Computational Statistics and Data Analysis*, 35, 76-82.
4. Engle, R F, Granger C W J, Rice J, and Weiss A (1986). Semiparametric estimates of the relationship between weather and electricity sales. *Journal of the American Statistical Association* 81: 310-320.
5. Fan, J and Gijbels I (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.
6. Fan, J., E. Mammen, and W. Härdle (1998). Direct Estimation of Low Dimensional Components in Additive Models, *Annals of Statistics*, 26, 943-971.
7. Härdle, W. 1990 *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.
8. Härdle, W., H. Liang, and J. Gao (2000). *Partially Linear Models*, Springer-Verlag, New York.
9. Hastie, T J and Tibshirani R J 1990 *Generalized Additive Models*. Chapman and Hall, London.
10. Hengartner, N.W. and S. Sperlich (2005). Rate Optimal Estimation with the Integration Method in the Presence of Many Covariates, *Journal of Multivariate Analysis*, 95, 246-272.
11. Horowitz, J.L. (1998). *Semiparametric Methods in Econometrics*. Springer-Verlag, New York.
12. Horowitz J.L. and Härdle W. (1996). Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates, *Journal of the American Statistical Association* 91, 1632-1640.
13. Horowitz, J.L. and S. Lee (2005). Nonparametric Estimation of an Additive Quantile Regression Model, *Journal of the American Statistical Association*, forthcoming.
14. Horowitz, J.L. and E. Mammen (2004). Nonparametric Estimation of an Additive Model with a Link Function, *Annals of Statistics*, 32, 2412-2443.
15. Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001). Structure Adaptive Approach for Dimension Reduction, *Annals of Statistics*, 29, 1537-1566.
16. Hristache, M., Juditsky, A., and Spokoiny, V. (2001). Structure Adaptive Approach for Dimension Reduction, *Annals of Statistics*, 29, 1-32.
17. Ichimura, H. (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models, *Journal of Econometrics* 58, 71-120.
18. Ichimura, H. and Lee L.-F. (1991). Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation. In Barnett W A, Powell J, and Tauchen G (eds), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press, Cambridge, pp. 3-49.
19. Klein, R.W. and R.H. Spady (1993). An Efficient Semiparametric Estimator for Binary Response Model, *Econometrica*, 61, 387-412.

20. Linton, O.B. (1997). Efficient Estimation of Additive Nonparametric Regression Models, *Biometrika* 84, 469-473.
21. Linton, O.B. and Härdle, W. (1996). Estimating Additive Regression Models with Known Links, *Biometrika*, 83, 529-540.
22. Linton, O.B. and Nielsen J.P. (1995). A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration, *Biometrika*, 82, 93-100.
23. Mammen, E., Linton, O.B., and Nielsen, J.P. (1999). The Existence and Asymptotic Properties of Backfitting Projection Algorithm under Weak Conditions, *Annals of Statistics*, 27, 1443-1490.
24. Matzkin, R.L. (1994). Restrictions of Economic Theory in Nonparametric Methods. In Engle, R.F. and McFadden, D.L. (eds) *Handbook of Econometrics*, Vol. 4. North-Holland, Amsterdam, pp. 2523-2558.
25. Newey, W.K. (1997). Convergence Rates and Asymptotic Normality of Series Estimators, *Journal of Econometrics*, 79, 147-168.
26. Opsomer, J.D. (2000). Asymptotic Properties of Backfitting Estimators, *Journal of Multivariate Analysis*, 73, 166-179.
27. Opsomer, J.D. and D. Ruppert (1997). Fitting a Bivariate Additive Model by Local Polynomial Regression, *Annals of Statistics*, 25, 186-211.
28. Powell, J.L. (1994). Estimation of Semiparametric Models. In Engle, R.F. and McFadden, D.L. (eds) *Handbook of Econometrics*, Vol. 4. North-Holland, Amsterdam, pp. 2444-2521.
29. Powell, J.L., Stock, J.H., and Stoker, T.M. (1989). Semiparametric Estimation of Index Coefficients, *Econometrica*, 51, 1403-1430.
30. Robinson, P.M. (1988). Root- $N$ -Consistent Semiparametric Regression. *Econometrica*, 56, 931-954.
31. Stock, J.H. (1989). Nonparametric Policy Analysis, *Journal of the American Statistical Association*, 84, 567-575.
32. Stock, J.H. (1991). Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits. In Barnett, W.A., Powell, J., and Tauchen, G. (eds) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press, Cambridge, pp. 77-98.