Horowitz, Joel L.

Working Paper

Nonparametric additive models

cemmap working paper, No. CWP20/12

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Horowitz, Joel L. (2012) : Nonparametric additive models, cemmap working paper, No. CWP20/12, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2012.2012

This Version is available at:
https://hdl.handle.net/10419/64766

cemmap
centre for microdata methods and practice

Nonparametric Additive Models

Joel L. Horowitz

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP20/12

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Nonparametric Additive Models

Joel L. Horowitz

1. INTRODUCTION

Much applied research in statistics, economics, and other fields is concerned with estimation of a conditional mean or quantile function. Specifically, let (Y, X) be a random pair, where Y is a scalar random variable and X is a d-dimensional random vector that is continuously distributed. Suppose we have data consisting of the random sample $\{Y_i, X_i : i = 1, ..., n\}$. Then the problem is to use the data to estimate the conditional mean function $g(x) \equiv E(Y \mid X = x)$ or the conditional $\alpha-$quantile function $Q_\alpha(x)$. The latter is defined by $P[Y \le Q_\alpha(x) \mid X = x] = \alpha$ for some α satisfying $0 < \alpha < 1$. For example, the conditional median function is obtained if $\alpha = 0.50$.

One way to proceed is to assume that g or Q_α is known up to a finite-dimensional parameter θ, thereby obtaining a parametric model of the conditional mean or quantile function. For example, if g is assumed to be linear, then $g(x) = \theta_0 + \theta_1' x$, where θ_0 is a scalar constant and θ_1 is a vector that is conformable with x. Similarly, if Q_α is assumed to be linear, then $Q_\alpha(x) = \theta_0 + \theta_1' x$. Given a finite-dimensional parametric model, the parameter θ can be estimated consistently by least squares in the case of conditional mean function and by least absolute deviations in the case of the conditional median function $Q_{0.5}$. Similar methods are available for other quantiles. However, a parametric model is usually arbitrary. For example, economic theory rarely if ever provides one, and a misspecified parametric model can be

seriously misleading. Therefore, it is useful to seek estimation methods that do not require assuming a parametric model for g or Q_α.

Many investigators attempt to minimize the risk of specification error by carrying out a specification search. In a specification search, several different parametric models are estimated, and conclusions are based on the one that appears to fit the data best. However, there is no guarantee that a specification search will include the correct model or a good approximation to it, and there is no guarantee that the correct model will be selected if it happens to be included in the search. Therefore, the use of specification searches should be minimized.

The possibility of specification error can be essentially eliminated through the use of nonparametric estimation methods. Nonparametric methods assume that g or Q_α satisfies certain smoothness conditions, but no assumptions are made about the shape or functional form of g or Q_α. See, for example, Fan and Gijbels (1996), Härdle 1990, Pagan and Ullah (1999), Li and Racine (2007), and Horowitz (2009), among many other references. However, the precision of a nonparametric estimator decreases rapidly as the dimension of X increases. This is called the curse of dimensionality. As a consequence of it, impracticably large samples are usually needed to obtain useful estimation precision if X is multi-dimensional.

The curse of dimensionality can be avoided through the use of dimension-reduction techniques. These reduce the effective dimension of the estimation problem by making assumptions about the form of g or Q_α that are stronger than those made by fully nonparametric estimation but weaker than those made in parametric modeling. Single-index and partially linear models (Härdle, Gao, and Liang 2000, Horowitz 2009) and nonparametric additive models, the subject of this chapter, are examples of ways of doing this. These models

achieve greater estimation precision than do fully nonparametric models, and they reduce (but do not eliminate) the risk of specification error relative to parametric models.

In a nonparametric additive model, g or Q_α is assumed to have the form

$$(1) \qquad \left.\begin{array}{l} g(x) \\ \text{or} \\ Q_\alpha(x) \end{array}\right\} = \mu + f_1(x^1) + f_2(x^2) + ... + f_d(x^d),$$

where μ is a constant, x^j ($j = 1,...,d$) is the j'th component of the d-dimensional vector x, and $f_1,..., f_d$ are functions that are assumed to be smooth but are otherwise unknown and are estimated nonparametrically. Model (1) can be extended to

$$(2) \qquad \left.\begin{array}{l} g(x) \\ \text{or} \\ Q_\alpha(x) \end{array}\right\} = F[\mu + f_1(x^1) + f_2(x^2) + ... + f_d(x^d)],$$

where F is a strictly increasing function that may be known or unknown.

It turns out that under mild smoothness conditions, the additive components $f_1,..., f_d$ can be estimated with the same precision that would be possible if X were a scalar. Indeed, each additive component can be estimated as well as it could be if all the other additive components were known. This chapter reviews methods for achieving these results. Section 2 describes methods for estimating model (1). Methods for estimating model (2) with a known or unknown link function F are described in Section 3. Section 4 discusses tests of additivity. Section 5 presents an empirical example that illustrates the use of model (1), and Section 6 presents conclusions. Estimation of derivatives of the functions $f_1,..., f_d$ is important in some applications. Estimation of derivatives is not discussed in this chapter but is discussed by Severance-Lossin and Sperlich (1999) and Yang, Sperlich, and Härdle (2003). The discussion in this chapter is informal. Regularity conditions and proofs of results are available in the

references that are cited in the chapter. The details of the methods described here are lengthy, so most methods are presented in outline form. Details are available in the cited references.

2. METHODS FOR ESTIMATING MODEL (1)

We begin with the conditional mean version of model (1), which can be written as

(3) $$E(Y \mid X = x) = \mu + f_1(x^1) + f_2(x^2) + ... + f_d(x^d).$$

The conditional quantile version of (1) is discussed in Section 2.1.

Equation (3) remains unchanged if a constant, say γ_j, is added to f_j ($j = 1,...,d$) and μ is replaced by $\mu - \sum_{j=1}^{d} \gamma_j$. Therefore, a location normalization is needed to identify μ and the additive components. Let X^j denote the j'th component of the random vector X. Depending on the method that is used to estimate the f_j's, location normalization consists of assuming that $Ef_j(X^j) = 0$ or that

(4) $$\int f_j(v)dv = 0$$

for each $j = 1,...,d$.

Stone (1985) was the first to give conditions under which the additive components can be estimated with a one-dimensional nonparametric rate of convergence and to propose an estimator that achieves this rate. Stone (1985) assumed that the support of X is $[0,1]^d$, that the probability density function of X is bounded away from 0 on $[0,1]^d$, and that $Var(Y \mid X = x)$ is bounded on $[0,1]^d$. He proposed using least squares to obtain spline estimators of the f_j's

under the location normalization $Ef_j(X^j) = 0$. Let $\hat{f}_j$ denote the resulting estimator of f_j. For any function h on $[0,1]$, define

$$\|h\|^2 = \int_0^1 h(v)^2 dv.$$

Stone (1985) showed that if each f_j is p times differentiable on $[0,1]$, then $E\left(\left\|\hat{f}_j - f_j\right\|^2 \middle| X^1,...,X^d\right) = O_p[n^{-2p/(2p+1)}]$. This is the fastest possible rate of convergence. However, Stone's result does not establish pointwise convergence of $\hat{f}_j$ to f_j or the asymptotic distribution of $n^{p/(2p+1)}[\hat{f}_j(x) - f_j(x)]$.

Since the work of Stone (1985), there have been many attempts to develop estimators of the f_j's that are pointwise consistent with the optimal rate of convergence and are asymptotically normally distributed. Oracle efficiency is another desirable property of such estimators. Oracle efficiency means that the asymptotic distribution of the estimator of any additive component f_j is the same as it would be if the other components were known.

Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1990) proposed an estimation method called backfitting. This method is based on the observation that

$$f_k(x^k) = E[Y - \mu - \sum_{j \neq k} f_j(x^j) \mid X = (x^1,...,x^d)].$$

If μ and the f_j's for $j \neq k$ were known, then f_k could be estimated by applying nonparametric regression to $Y - \mu - \sum_{j \neq k} f_j(X^j)$. Backfitting replaces the unknown quantities by preliminary estimates. Then each additive component is estimated by nonparametric regression, and the preliminary estimates are updated as each additive component is estimated. In principle, this process continues until convergence is achieved. Backfitting is implemented in many statistical software packages, but theoretical investigation of the statistical properties of backfitting estimators is difficult. This is because these estimators are outcomes of an iterative process, not the solutions to optimization problems or systems of equations. Opsomer and Ruppert (1997)

and Opsomer (2000) investigated the properties of a version of backfitting and found, among other things, that strong restrictions on the distribution of X were necessary to achieve results and that the estimators are not oracle efficient. Other methods described below are oracle efficient and have additional desirable properties. Compared to these estimators, backfitting is not a desirable approach, despite its intuitive appeal and availability in statistical software packages.

The first estimator of the f_j's that was proved to be pointwise consistent and asymptotically normally distributed was developed by Linton and Nielsen (1995) and extended by Linton and Härdle (1996). Tjøstheim and Auestad (1994) and Newey (1994) present similar ideas. The method is called marginal integration and is based on the observation that under the location normalization $Ef_j(X^j) = 0$, $\mu = E(Y)$ and

$$f_j(x^j) = \int E(Y \mid X = x) p_{-j}(x^{(-j)}) dx^{(-j)} - \mu, \tag{5}$$

where $x^{(-j)}$ is the vector consisting of all components of x except x^j and p_{-j} is the probability density function of $X^{(-j)}$. The constant μ is estimated consistently by the sample analog

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

To estimate, say, $f_1(x^1)$, let $\hat{g}(x^1, x^{(-1)})$ be the following kernel estimator of $E(Y \mid X^1 = x^1, X^{(-1)} = x^{(-1)})$:

$$\hat{g}(x^1, x^{(-1)}) = \hat{P}(x^1, x^{(-1)})^{-1} \sum_{i=1}^{n} Y_i K_1\left(\frac{x^1 - X_i^1}{h_1}\right) K_2\left(\frac{x^{(-1)} - X_i^{(-1)}}{h_2}\right), \tag{6}$$

where

$$\hat{P}(x^1, x^{(-1)}) = \sum_{i=1}^{n} K_1\left(\frac{x^1 - X_i^1}{h_1}\right) K_2\left(\frac{x^{(-1)} - X_i^{(-1)}}{h_2}\right), \tag{7}$$

K_1 is a kernel function of a scalar argument, K_2 is a kernel function of a $d-1$ dimensional argument, $X_i^{(-1)}$ is the i'th observation of $X^{(-1)}$, and h_1 and h_2 are bandwidths. The integral on the right-hand side of (5) is the average of $E(Y \mid X^1 = x^1, X^{(-1)} = x^{(-1)})$ over $X^{(-1)}$ and can be estimated by the sample average of $\hat{g}(x^1, X^{(-1)})$. The resulting marginal integration estimator of f_1 is

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

Linton and Härdle (1996) give conditions under which $n^{2/5}[\hat{f}_1(x^1) - f_1(x^1)] \to^d N[\beta_{1,MI}(x^1), V_{1,MI}(x^1)]$ for suitable functions $\beta_{1,MI}$ and $V_{1,MI}$. Similar results hold for the marginal integration estimators of the other additive components. The most important condition is that each additive component is at least d times continuously differentiable. This condition implies that the marginal integration estimator has a form of the curse of dimensionality, because maintaining an $n^{-2/5}$ rate of convergence in probability requires the smoothness of the additive components to increase as d increases. In addition, the marginal integration estimator is not oracle efficient and can be hard to compute.

There have been several refinements of the marginal integration estimator that attempt to overcome these difficulties. See, for example, Linton (1997), Kim, Linton, and Hengartner (1999), and Hengartner and Sperlich (2005). Some of these refinements overcome the curse of dimensionality, and others achieve oracle efficiency. However, none of the refinements is both free of the curse of dimensionality and oracle efficient.

The marginal integration estimator has a curse of dimensionality because, as can be seen from (6) and (7), it requires full-dimensional nonparametric estimation of $E(Y \mid X = x)$ and the probability density function of X. The curse of dimensionality can be avoided by imposing additivity at the outset of estimation, thereby avoiding the need for full-dimensional nonparametric estimation. This cannot be done with kernel-based estimators, such as those used in marginal integration, but it can be done easily with series estimators. However, it is hard to establish the asymptotic distributional properties of series estimators. Horowitz and Mammen

(2004) proposed a two-step estimation procedure that overcomes this problem. The first step of the procedure is series estimation of the f_j's. This is followed by a backfitting step that turns the series estimates into kernel estimates that are both oracle efficient and free of the curse of dimensionality.

Horowitz and Mammen (2004) use the location normalization (4) and assume that the support of X is $[-1,1]^d$. Let $\{\psi_k : k = 1, 2, ...\}$ be an orthonormal basis for smooth functions on $[-1,1]$ that satisfies (4). The first step of the Horowitz-Mammen (2004) procedure consists of using least squares to estimate μ and the generalized Fourier coefficients $\{\theta_{jk}\}$ in the series approximation

$$(8) \qquad E(Y \mid X = x) \approx \mu + \sum_{j=1}^{d} \sum_{k=1}^{\kappa} \theta_{jk} \psi_k(x^j),$$

where κ is the length of the series approximations to the additive components. In this approximation, f_j is approximated by

$$f_j(x^j) \approx \sum_{k=1}^{\kappa} \theta_{jk} \psi_k(x^j).$$

Thus, the estimators of μ and the θ_{jk}'s are given by

$$\{\tilde{\mu}, \tilde{\theta}_{jk} : j = 1, ..., d; k = 1, ..., \kappa\} = \arg\min_{\mu, \theta_{jk}} \sum_{i=1}^{n} \left[Y_i - \mu - \sum_{j=1}^{d} \sum_{k=1}^{\kappa} \theta_{jk} \psi_k(X_i^j) \right]^2,$$

where X_i^j is the j'th component of the vector X_i. Let $\tilde{f}_j$ denote the resulting estimator of μ and f_j ($j = 1, ..., d$). That is,

$$\tilde{f}_j(x^j) = \sum_{k=1}^{\kappa} \tilde{\theta}_{jk} \psi_k(x^j).$$

Now let K and h, respectively, denote a kernel function and a bandwidth. The second-step estimator of, say, f_1 is

$$(9) \qquad \hat{f}_1(x^1) = \left[\sum_{i=1}^{n} K\left(\frac{x^1 - X_i^1}{h}\right)\right]^{-1} \sum_{i=1}^{n} [Y_i - \tilde{f}_{-1}(X_i^{(-1)})] K\left(\frac{x^1 - X_i^1}{h}\right),$$

where $X_i^{(-1)}$ is the vector consisting of the i'th observations of all components of X except the first and $\tilde{f}_{-1} = \tilde{f}_2 + ... + \tilde{f}_d$. In other words, $\hat{f}_1$ is the kernel nonparametric regression of $Y - \tilde{f}_{-1}(X^{(-1)})$ on X^1. Horowitz and Mammen (2004) give conditions under which $n^{2/5}[\hat{f}_1(x^1) - f_1(x^1)] \to^d N[\beta_{1,HM}(x^1), V_{1,HM}(x^1)]$ for suitable functions $\beta_{1,HM}$ and $V_{1,HM}$. Horowitz and Mammen (2004) also show that the second-step estimator is free of the curse of dimensionality and oracle efficient. Freedom from the curse of dimensionality means that the f_j's need to have only two continuous derivatives, regardless of d. Oracle efficiency means that the asymptotic distribution of $n^{2/5}[\hat{f}_1(x^1) - f_1(x^1)]$ is the same as it would be if the estimator $\tilde{f}_{-1}$ in (9) were replaced with the true (but unknown) sum of additive components, f_{-1}. Similar results apply to the second-step estimators of the other additive components. Thus, asymptotically, each additive component f_j can be estimated as well as it could be if the other components were known. Intuitively, the method works because the bias due to truncating the series approximations to the f_j's in the first estimation step can be made negligibly small by making κ increase at a sufficiently rapid rate as n increases. This increases the variance of the $\tilde{f}_j$'s, but the variance is reduced in the second estimation step because this step includes

averaging over the $\tilde{f}_j$'s. Averaging reduces the variance enough to enable the second-step estimates to have an $n^{-2/5}$ rate of convergence in probability.

There is also a local linear version of the second step estimator. For estimating f_1, this consists of choosing b_0 and b_1 on minimize

$$S_n(b_0, b_1) = (nh)^{-1} \sum_{i=1}^{n} [Y_i - \tilde{\mu} - b_0 - b_1(X_i^1 - x^1) - \tilde{f}_{-1}(X_i^{(-1)}]^2 K\left(\frac{X_i^1 - x^1}{h}\right).$$

Let $(\hat{b}_0, \hat{b}_1)$ denote resulting value of (b_0, b_1). The local linear second-step estimator of $f_1(x^1)$ is $\hat{f}_1(x^1) = \hat{b}_0$. The local linear estimator is pointwise consistent, asymptotically normal, oracle efficient, and free of the curse of dimensionality. However, the mean and variance of the asymptotic distribution of the local linear estimator are different from those of the Nadaraya-Watson (or local constant) estimator (9). Fan and Gijbels (1996) discuss the relative merits of local linear and Nadaraya-Watson estimators.

Mammen, Linton, and Nielsen (1999) developed an asymptotically normal, oracle-efficient estimation procedure for model (1) that consists of solving a certain set of integral equations. Wang and Yang (2007) generalized the two-step method of Horowitz and Mammen (2004) to autoregressive time-series models. Their model is

$$Y_t = \mu + f_1(X_t^1) + ... + f_d(X_t^d) + \sigma(X_t^1, ..., X_t^d)\varepsilon_t; \quad t = 1, 2, ...,$$

where X_t^j is the j'th component of the d-vector X_t, $E(\varepsilon_t \mid X_t) = 0$, and $E(\varepsilon_t^2 \mid X_t) = 1$. The explanatory variables $\{X_t^j : j = 1, ..., d\}$ may include lagged values of the dependent variable Y_t. The random vector (X_t, ε_t) is required to satisfy a strong mixing condition, and the additive components have two derivatives. Wang and Yang (2007) propose an estimator that is like that

of Horowitz and Mammen (2004), except the first step uses a spline basis that is not necessarily orthogonal. Wang and Yang (2007) show that their estimator of each additive component is pointwise asymptotically normal with an $n^{-2/5}$ rate of convergence in probability. Thus, the estimator is free of the curse of dimensionality. It is also oracle efficient. Nielsen and Sperlich (2005) and Wang and Yang (2007) discuss computation of some of the foregoing estimators.

Song and Yang (2010) describe a different two-step procedure for obtaining oracle efficient estimators with time-series data. Like Wang and Yang (2007), Song and Yang (2010) consider a nonparametric, additive, autoregressive model in which the covariates and random noise component satisfy a strong mixing condition. The first estimation step consists of using least squares to make a constant-spline approximation to the additive components. The second step is like that of Horowitz and Mammen (2004) and Wang and Yang (2007), except a linear spline estimator replaces the kernel estimator of those papers. Most importantly, Song and Yang (2010) obtain asymptotic uniform confidence bands for the additive components. They also report that their two-stage spline estimator can be computed much more rapidly than procedures that use kernel-based estimation in the second step. Horowitz and Mammen (2004) and Wang and Yang (2007) obtained pointwise asymptotic normality for their estimators but did not obtain uniform confidence bands for the additive components. However, the estimators of Horowitz and Mammen (2004) and Wang and Yang (2007) are, essentially, kernel estimators. Therefore, these estimators are multivariate normally distributed over a grid of points that are sufficiently far apart. It is likely that uniform confidence bands based on the kernel-type estimators can be obtained by taking advantage of this multivariate normality and letting the spacing of the grid points decrease slowly as n increases.

2.1 Estimating a Conditional Quantile Function

This section describes estimation of the conditional quantile version of (1). The discussion concentrates on estimation of the conditional median function, but the methods and results also apply to other quantiles. Model (1) for the conditional median function can be estimated using series methods or backfitting, 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. Like the marginal integration estimator for a conditional mean function, the marginal integration estimator for a conditional median or other conditional quantile function is asymptotically normally distributed but suffers from the curse of dimensionality.

Horowitz and Lee (2005) proposed a two-step estimation procedure that is similar to that of Horowitz and Mammen (2004) for conditional mean functions. The two-step method is oracle efficient and has no curse of dimensionality. The first step of the method of Horowitz and Lee (2005) consists of using least absolute deviations (LAD) to estimate μ and the θ_{jk}'s in the series approximation (8). That is,

$$\{\tilde{\mu}, \tilde{\theta}_{jk} : j = 1,...,d; k = 1,...,\kappa\} = \arg\min_{\mu,\theta_{jk}} \sum_{i=1}^{n} \left| Y_i - \mu - \sum_{j=1}^{d} \sum_{k=1}^{\kappa} \theta_{jk} \psi_k(X_i^j) \right|,$$

As before, $\tilde{f}_j$ denote the first-step estimator of f_j. The second-step of the method of Horowitz and Lee (2005) is of a form local-linear LAD estimation that is analogous to the second-step of the method of Horowitz and Mammen (2004). For estimating f_1, this step consists of choosing b_0 and b_1 to minimize

$$S_n(b_0, b_1) = (nh)^{-1} \sum_{i=1}^{n} | Y_i - \tilde{\mu} - b_0 - b_1(X_i^1 - x^1) - \tilde{f}_{-1}(X_i^{(-1)} | K\left(\frac{X_i^1 - x^1}{h} \right),$$

where h is a bandwidth, K is a kernel function, and $\tilde{f}_{-1} = \tilde{f}_2 + ... + f_d$. Let $(\hat{b}_0, \hat{b}_1)$ denote resulting value of (b_0, b_1). The estimator of $f_1(x^1)$ is $\hat{f}_1(x^1) = \hat{b}_0$. Thus, the second-step estimator of any additive component is a local linear conditional median estimator. Horowitz and Lee (2005) give conditions under which $n^{2/5}[\hat{f}_1(x^1) - f_1(x^1)] \to^d N[\beta_{1,HL}(x^1), V_{1,HL}(x^1)]$ for suitable functions $\beta_{1,HL}$ and $V_{1,HL}$. Horowitz and Lee (2005) also show that that $\hat{f}_1$ is free of the curse of dimensionality and oracle efficient. Similar results apply to the estimators of the other f_j's.

3. METHODS FOR ESTIMATING MODEL (2)

This section describes methods for estimating model (2) when the link function F is not the identity function. Among other applications, this permits extension of methods for nonparametric additive modeling to settings in which Y is binary. For example, an additive binary probit model is obtained by setting

(10) $$P(Y = 1 | X = x) = \Phi[\mu + f_1(x^1) + ... + f_d(X^d)],$$

where Φ is the standard normal distribution function. In this case, the link function is $F = \Phi$. A binary logit model is obtained by replacing Φ in (10) with the logistic distribution function.

Section 3.1 treats the case in which F is known. Section 3.2 treats bandwidth selection for one of the methods discussed in Section 3.1 Section 3.3 discusses estimation when F is unknown.

3.1 Estimation with a Known Link Function

In this section, it is assumed that the link function F is known. A necessary condition for point identification of μ and the f_j's is that F is strictly monotonic. Given this requirement, it can be assumed without loss of generality that F is strictly increasing. Consequently, $F^{-1}[Q_\alpha(x)]$ is the α conditional quantile of $F^{-1}(Y)$ and has a nonparametric additive form. Therefore, quantile estimation of the additive components of model (2) can be carried out by applying the methods of Section 2.1 to $F^{-1}(Y)$. Accordingly, the remainder of this section is concerned with estimating the conditional mean version of model (2).

Linton and Härdle (1996) describe a marginal integration estimator of the additive components in model (2). As in the case of model (1), the marginal integration estimator has a curse of dimensionality and is not oracle efficient. The two-step method of Horowitz and Mammen (2004) is also applicable to model (2). When F has a Lipschitz continuous second derivative and the additive components are twice continuously differentiable, it yields asymptotically normal, oracle efficient estimators of the additive components. The estimators have an $n^{-2/5}$ rate of convergence in probability and no curse of dimensionality.

The first step of the method of Horowitz and Mammen (2004) is nonlinear least squares estimation of truncated series approximations to the additive components. That is, the generalized Fourier coefficients of the approximations are estimated by solving

$$\{\tilde{\mu}, \tilde{\theta}_{jk} : j = 1,...,d; k = 1,...,\kappa\} = \arg\min_{\mu,\theta_{jk}} \sum_{i=1}^{n} \left\{ Y_i - F\left[\mu + \sum_{j=1}^{d} \sum_{k=1}^{\kappa} \theta_{jk} \psi_k(x^j) \right] \right\}^2 .$$

Now set

$$\tilde{f}_j(x^j) = \sum_{k=1}^{K} \tilde{\theta}_{jk} \psi_k(x^j).$$

A second-step estimator of $f_1(x^1)$, say can be obtained by setting

$$\tilde{\tilde{f}}_1(x^1) = \arg\min_b \sum_{i=1}^{n} \left\{ Y_i - F\left[\tilde{\mu} + b + \sum_{j=2}^{d} \tilde{f}_j(X_i^j) \right] \right\}^2 K\left(\frac{x^1 - X_i^1}{h} \right),$$

where, as before, K is a kernel function and h is a bandwidth. However, this requires solving a difficult nonlinear optimization problem. An asymptotically equivalent estimator can be obtained by taking one Newton step from $b_0 = \tilde{f}_1(x^1)$ toward $\tilde{\tilde{f}}_1(x^1)$. To do this, define

$$S'_{n1}(x^1, f) = -2\sum_{i=1}^{n} \left\{ Y_i - F[\mu + f_1(x^1) + f_2(X_i^2) + \ldots + f_d(X_i^d)] \right\}$$

$$\times F'[\mu + f_1(x^1) + f_2(X_i^2) + \ldots + f_d(X_i^d)] K\left(\frac{x^1 - X_i^1}{h} \right)$$

and

$$S''_{n1}(x^1, f) = 2\sum_{i=1}^{n} F'[\mu + f_1(x^1) + f_2(X_i^2) + \ldots + f_d(X_i^d)]^2 K\left(\frac{x^1 - X_i^1}{h} \right)$$

$$-2\sum_{i=1}^{n} \{ Y_i - F[\mu + f_1(x^1) + f_2(X_i^2) + \ldots + f_d(X_i^d)] \}$$

$$\times F''[\mu + f_1(x^1) + f_2(X_i^2) + \ldots + f_d(X_i^d)] K\left(\frac{x^1 - X_i^1}{h} \right).$$

The second-step estimator is

$$\hat{f}_1(x^1) = \tilde{f}_1(x^1) - S'_{n1}(x^1 \tilde{f}) / S''_{n1}(x^1, \tilde{f}).$$

Horowitz and Mammen (2004) also describe a local-linear version of this estimator.

Liu, Yang, and Härdle (2011) describe a two-step estimation method for model (2) that is analogous to the method of Wang and Yang (2007) but uses a local pseudo log-likelihood objective function based on the exponential family at each estimation stage instead of a local least squares objective function. As in Wang and Yang (2007), the method of Liu, Yang, and Härdle (2011) applies to an autoregressive model in which the covariates and random noise satisfy a strong mixing condition. Yu, Park, and Mammen (2008) proposed an estimation method for model (2) that is based on numerically solving a system of nonlinear integral equations. The method is more complicated than that of Horowitz and Mammen (2004), but the results of Monte Carlo experiments suggest that the estimator of Yu, Park, and Mammen (2008) has better finite-sample properties than that of Horowitz and Mammen (2004), especially when the covariates are highly correlated.

3.2 Bandwidth Selection for the Two-Step Estimatorof Horowitz and Mammen (2004)

This section describes a penalized least squares (PLS) method for choosing the bandwidth h in the second step of the procedure of Horowitz and Mammen (2004). The method is described here for the local-linear version of the method, but similar results apply to the local constant version. The method described in this section can be used with model (1) by setting F equal to the identity function.

The PLS method simultaneously estimates the bandwidths for second-step estimation of all the additive components f_j ($j = 1,...,d$). Let $h_j = C_j n^{-1/5}$ be the bandwidth for $\hat{f}_j$. The PLS method selects the C_j's that minimize an estimate of the average squared error (ASE):

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

where $\hat{f} = \hat{f}_1 + ... + \hat{f}_d$ and $\bar{h} = (C_1 n^{-1/5}, ..., C_d n^{-1/5})$. Specifically, the PLS method selects the C_j's to

$$\text{(11)} \quad \underset{C_1,...,C_d}{\text{minimize}} : PLS(\bar{h}) = n^{-1}\sum_{i=1}^{n}[Y_i - F[\tilde{\mu} + \hat{f}(X_i)]^2 + 2K(0)n^{-1}\sum_{i=1}^{n}\{F'[\tilde{\mu} + \hat{f}(X_i)]^2 \hat{V}(X_i)\}$$

$$\times \sum_{j=1}^{d}[n^{4/5} C_j \hat{D}_j(X_i^j)]^{-1},$$

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

$$D_j(x^j) = (nh_j)^{-1}\sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j}\right) F'[\tilde{\mu} + \hat{f}(X_i)]^2$$

and

$$\hat{V}(x) = \left[\sum_{i=1}^{n} K\left(\frac{X_i^1 - x^1}{h_1}\right) ... K\left(\frac{X_i^d - x^d}{h_d}\right)\right]^{-1}$$

$$\times \sum_{i=1}^{n} K\left(\frac{X_i^1 - x^1}{h_1}\right) ... K\left(\frac{X_i^d - x^d}{h_d}\right) \{Y_i - F[\tilde{\mu} + \hat{f}(X_i)]^2 .$$

The bandwidths for $\hat{V}$ may be different from those used for $\hat{f}$, because $\hat{V}$ is a full-dimensional nonparametric estimator. Horowitz and Mammen (2004) present arguments showing that the solution to (11) estimates the bandwidths that minimize ASE.

3.3 Estimation with an Unknown Link Function

This section is concerned with estimating model (2) when the link function F is unknown. When F is unknown, model (2) contains semiparametric single-index models as a

special case. This is important, because semiparametric single-index models and nonparametric additive models with known link functions are non-nested. In a semiparametric single-index $E(Y \mid X = x) = G(\theta' x)$ for some unknown function G and parameter vector θ. This model coincides with the nonparametric additive model with link function F only if the additive components are linear and $F = G$. An applied researcher must choose between the two models and may obtain highly misleading results if an incorrect choice is made. A nonparametric additive model with an unknown link function makes this choice unnecessary, because the model nests semiparametric single index models and nonparametric additive models with known link functions. A nonparametric additive model with an unknown link function also nests the multiplicative specification

$$E(Y \mid X = x) = F[f_1(x^1) f_2(x^2) \ldots f_d(x^d)].$$

A further attraction of model (2) with an unknown link function is that it provides an informal, graphical method for checking the additive and single-index specifications. One can plot the estimates of F and the f_j's. Approximate linearity of the estimate of F favors the additive specification (1), whereas approximate linearity of the f_j's favors the single-index specification. Linearity of F and the f_j's favors the linear model $E(Y \mid X) = \theta' X$.

Identification of the f_j's in model (2) requires more normalizations and restrictions when F is unknown than when F is known. First, observe that μ is not identified when f is unknown, because $F[\mu + f_1(x^1) + \ldots + f_d(x^d)] = F^*[f_1(x^1) + \ldots + f_d(x^d)]$, where the function F^* is defined by $F^*(v) = F(\mu + v)$ for any real v. Therefore, we can set $\mu = 0$ without loss of generality. Similarly, a location normalization is needed because model (2) remains unchanged

if each f_j is replaced by $f_j + \gamma_j$, where γ_j is a constant, and $F(v)$ is replaced by $F^*(v) = F(v - \gamma_1 - ... - \gamma_d)$. In addition, a scale normalization is needed because model (2) is unchanged if each f_j is replaced by cf_j for any constant $c \neq 0$ and $F(v)$ is replaced by $F^*(v) = F(v / c)$. Under the additional assumption that F is monotonic, model (2) with F unknown is identified if at least two additive components are not constant. To see why this assumption is necessary, suppose that only f_1 is not constant. Then conditional mean function is of the form $F[f_1(x^1) + \text{constant}]$. It is clear that this function does not identify F and f_1. The methods presented in this discussion use a slightly stronger assumption for identification. We assume that the derivatives of two additive components are bounded away from 0. The indices j and k of these components do not need to be known. It can be assumed without loss of generality that $j = d$ and $k = d - 1$.

Under the foregoing identifying assumptions, oracle-efficient, pointwise asymptotically normal estimators of the f_j's can be obtained by replacing F in the procedure of Horowitz and Mammen (2004) for model (2) with a kernel estimator. As in the case of model (2) with F known, estimation takes place in two steps. In the first step, a modified version of Ichimura's (1993) estimator for a semiparametric single-index model is used to obtain a series approximation to each f_j and a kernel estimator of F. The first-step procedure imposes the additive structure of model (2), thereby avoiding the curse of dimensionality. The first-step estimates are inputs to the second step. The second-step estimator of, say, f_1 is obtained by taking one Newton step from the first-step estimate toward a local nonlinear least-squares estimate. In large samples, the second-step estimator has a structure similar to that of a kernel nonparametric regression estimator, so deriving its pointwise rate of convergence and asymptotic

distribution is relatively easy. The details of the two-step procedure are lengthy. They are presented in Horowitz and Mammen (2011). The oracle-efficiency property of the two-step estimator implies that asymptotically, there is no penalty for not knowing F in a nonparametric additive model. Each f_j can be estimated as well as it would be if F and the other f_j's were known.

Horowitz and Mammen (2007) present a penalized least squares (PLS) estimation procedure that applies to model (2) with an unknown F and also applies to a larger class of models that includes quantile regressions and neural networks. The procedure uses the location and scale normalizations $\mu = 0$, (4), and

$$\sum_{j=1}^{d} \int f_j^2(v)dv = 1. \tag{12}$$

The PLS estimator of Horowitz and Mammen (2007) chooses the estimators of F and the additive components to solve

$$\underset{\breve{F}, \breve{f}_1, ..., \breve{f}_d}{\text{minimize:}} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \breve{F}[\breve{f}_1(X_i^1) + ... + \breve{f}_d(X_i^d)]\} + \lambda_n^2 J(\breve{F}, \breve{f}_1, ..., \breve{f}_d) \tag{13}$$

subject to: (4), (12),

where $\{\lambda_n\}$ is a sequence of constants and J is a penalty term that penalizes roughness of the estimated functions. If F and the f_j's are k times differentiable, the penalty term is

$$J(\breve{F}, \breve{f}_1, ..., \breve{f}_d) = J_1^{\nu_1}(\breve{F}, \breve{f}_1, ..., \breve{f}_d) + J_2^{\nu_2}(\breve{F}, \breve{f}_1, ..., \breve{f}_d),$$

where ν_1 and ν_2 are constants satisfying $\nu_2 \geq \nu_1 > 0$,

$$J_1(\breve{F}, \breve{f}_1, ..., \breve{f}_d) = T_k(\breve{F})\left\{\sum_{j=1}^{d}[T_1^2(\breve{f}_j) + T_k^2(\breve{f}_j)]\right\}^{(2k-1)/4},$$

$$J_2(\breve{F}, \breve{f}_1, ..., \breve{f}_d) = T_1(\breve{F})\left\{\sum_{j=1}^{d}[T_1^2(\breve{f}_j) + T_k^2(\breve{f}_j)]\right\}^{1/4},$$

and

$$T_\ell^2(f) = \int f^{(\ell)}(v)^2 dv$$

for $0 \le \ell \le k$ and any function f whose ℓ'th derivative is square integrable. The PLS estimator can be computed by approximating $\breve{F}$ and the $\breve{f}_j$'s by B-splines and minimizing (13) over the coefficients of the spline approximation. Denote the estimator by $\hat{F}, \hat{f}_1, ..., \hat{f}_d$. Assume without loss of generality that the X is supported on $[0,1]^d$. Horowitz and Mammen (2007) give conditions under which the following result holds:

$$\int_0^1 [\hat{f}_j(v) - f_j(v)]^2 dv = O_p(n^{-2k/(2k+1)})$$

for each $j = 1, ..., d$ and

$$\int\left\{\hat{F}\left[\sum_{j=1}^{d} f_j(x^j)\right] - F\left[\sum_{j=1}^{d} f_j(x^j)\right]\right\}^2 dx^1 ... dx^d = O_p(n^{-2k/(2k+1)}).$$

In other words, the integrated squared errors of the PLS estimates of the link function and additive components converge in probability to 0 at the fastest possible rate under the assumptions. There is no curse of dimensionality. The available results do not provide an asymptotic distribution for the PLS estimator. Therefore, it is not yet possible to carry out statistical inference with this estimator.

4. TESTS OF ADDITIVITY

Models (1) and (2) are misspecified and can give misleading results if the conditional mean or quantile of Y is not additive. Therefore, it is useful to be able to test additivity. Several

tests of additivity have been proposed for models of conditional mean functions. These tests undoubtedly can be modified for use with conditional quantile functions, but this modification has not yet been carried out. Accordingly, the remainder of this section is concerned with testing additivity in the conditional mean versions of models (1) and (2). Bearing in mind that model (1) can be obtained from model (2) by letting F be the identity function, the null hypothesis to be tested is

$$H_0: E(Y \mid X = x) = F[\mu + f_1(x^1) + ... + f_d(x^d)].$$

The alternative hypothesis is

$$H_1: E(Y \mid X = x) = F[\mu + f(x)],$$

where there are no functions $f_1, ..., f_d$ such that

$$P[f(X) = f_1(X^1) + ... + f_d(X^d)] = 1.$$

Gozalo and Linton (2001) have proposed a general class of tests. Their tests are applicable regardless of whether F is the identity function. Wang and Carriere (2011) and Dette and von Lieres und Wilkau (2001) proposed similar tests for the case of an identity link function. These tests are based on comparing fully a fully nonparametric estimator of f with an estimator that imposes additivity. Eubank, Hart, Simpson and Stefanski (1995) also proposed tests for the case in which F is the identity function. These tests look for interactions among the components of X and are based on Tukey's (1949) test for additivity in analysis of variance. Sperlich, Tjøstheim and Yang (2002) also proposed a test for the presence of interactions among components of X. Other tests have been proposed by Abramovich, De Fesis, and Sapatinas (2009) and Derbort, Dette, and Munk (2002).

The remainder of this section outlines a test that Gozalo and Linton (2001) found though Monte Carlo simulation to have satisfactory finite sample performance. The test statistic has the form

$$\hat{\tau}_n = \sum_{i=1}^{n} \{F^{-1}[\hat{f}(X_i)] - [\hat{\mu} + \hat{f}_1(X_i^1) + ... + f_d(X_i^d)]\}^2 \pi(X_i),$$

where $\hat{f}(x)$ is a full-dimensional nonparametric estimator of $E(Y \mid X = x)$, $\hat{\mu}$ and the $\hat{f}_j$'s are estimators of μ and f_j under H_0, and π is a weight function. Gozalo and Linton (2001) use a Nadaraya-Watson kernel estimator for $\hat{f}$ and a marginal integration estimator for $\hat{\mu}$ and the $\hat{f}_j$'s. Dette and von Lieres und Wilkau (2001) also use these marginal integration estimators in their version of the test. However, other estimators can be used. Doing so might increase the power of the test or enable some of the regularity conditions of Gozalo and Linton (2001) to be relaxed. In addition, it is clear that $\hat{\tau}_n$ can be applied to conditional quantile models, though the details of the statistic's asymptotic distribution would be different from those with conditional mean models. If F is unknown, then $F^{-1}[f(x)]$ is not identified, but a test of additivity can be based on the following modified version of $\hat{\tau}_n$:

$$\hat{\tau}_n = \sum_{i=1}^{n} \{\hat{f}(X_i) - \hat{F}[\hat{\mu} + \hat{f}_1(X_i^1) + ... + f_d(X_i^d)]\}^2 \pi(X_i),$$

where $\hat{f}$ is a full-dimensional nonparametric estimator of the conditional mean function, $\hat{F}$ is a nonparametric estimator of F, and the $\hat{f}_j$'s are estimators of the additive components.

Gozalo and Linton (2001) give conditions under which a centered, scaled version of $\hat{\tau}_n$ is asymptotically normally distributed as $N(0,1)$. Dette and von Lieres und Wilkau (2001) provide similar results for the case in which F is the identity function. Gozalo and Linton (2001) and Dette and von Lieres und Wilkau (2001) also provide formulae for estimating the centering and scaling parameters. Simulation results reported by Gozalo and Linton (2001) indicate that using

the wild bootstrap to find critical values produces smaller errors in rejection probabilities under H_0 than using critical values based on the asymptotic normal distribution. Dette and von Lieres und Wilkau (2001) also used the wild bootstrap to estimate critical values.

5. AN EMPIRICAL APPLICATION

This section illustrates the application of the estimator of Horowitz and Mammen (2004) by using it to estimate a model of the rate of growth of gross domestic product (GDP) among countries. The model is

$$G = f_T(T) + f_S(S) + U,$$

where G is the average annual percentage rate of growth of a country's GDP from 1960 to 1965, T is the average share of trade in the country's economy from 1960 to 1965 measured as exports plus imports divided by GDP, and S is the average number of years of schooling of adult residents of the country in 1960. U is an unobserved random variable satisfying $E(U \mid T, S) = 0$. The functions f_T and f_S are unknown and are estimated by the method of Horowitz and Mammen (2004). The data are taken from the dataset **Growth** in Stock and Watson (2011). They comprise values of G, T, and S for 60 countries.

Estimation was carried out using a cubic B-spline basis in the first step. The second step consisted of Nadaraya-Watson (local constant) kernel estimation with the biweight kernel. Bandwidths of 0.5 and 0.8 were used for estimating f_T and f_S, respectively.

The estimation results are shown in Figures 1-2. The estimates of f_T and f_S are

INSERT FIGURES 1 AND 2 HERE

nonlinear and differently shaped. The dip in f_S near $S = 7$ is almost certainly an artifact of random sampling errors. The estimated additive components are not well-approximated by

simple parametric functions such as quadratic or cubic functions. A lengthy specification search might be needed to find a parametric model that produces shapes like those in Figures 1-2. If such a search were successful, the resulting parametric models might provide useful compact representations of f_T and f_S but could not be used for valid inference.

6. CONCLUSIONS

Nonparametric additive modeling with a link function that may or may not be known is an attractive way to achieve dimension reduction in nonparametric models. It greatly eases the restrictions of parametric modeling without suffering from the lack of precision that the curse of dimensionality imposes on fully nonparametric modeling. This chapter has reviewed a variety of methods for estimating nonparametric additive models. An empirical example has illustrated the usefulness of the nonparametric additive approach. Several issues about the approach remain unresolved. One of these is to find ways to carry out inference about additive components based on the estimation method of Horowitz and Mammen (2007) that is described in Section 3.3. This is the most general and flexible method that has been developed to date. Another issue is the extension of the tests of additivity described in Section 5 to estimators other than partial integration and models of conditional quantiles. Finally, finding data-based methods for choosing tuning parameters for the various estimation and testing procedures remains an open issue.

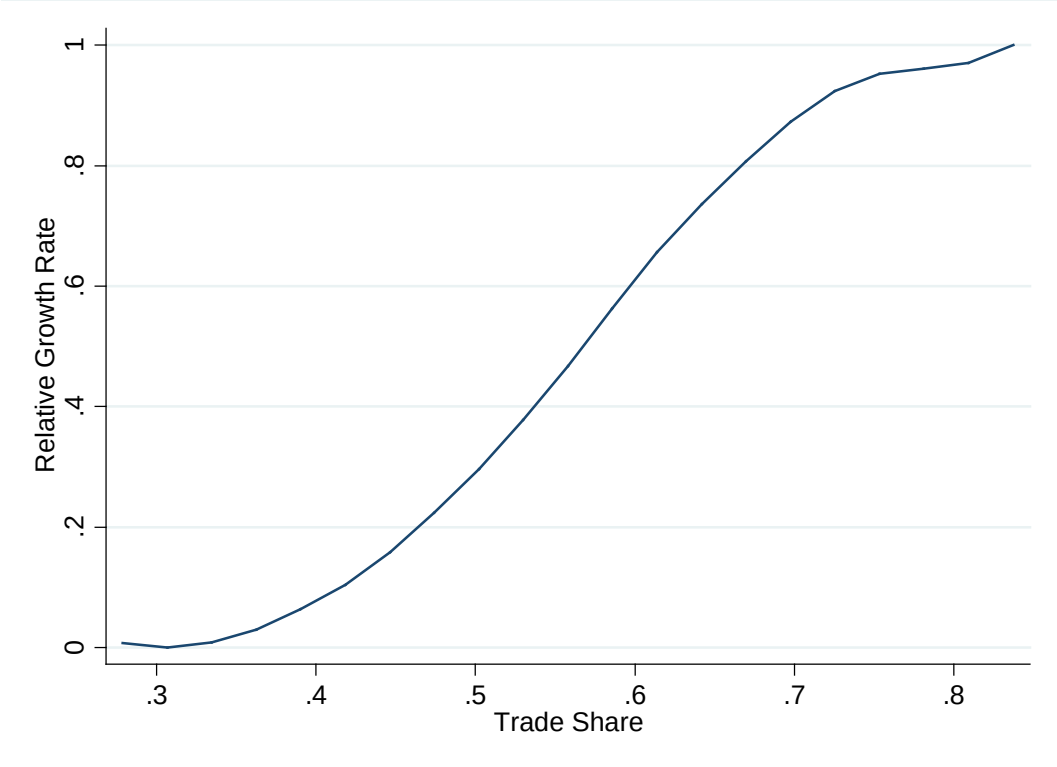

Figure 1: Additive component f_T in the growth model.

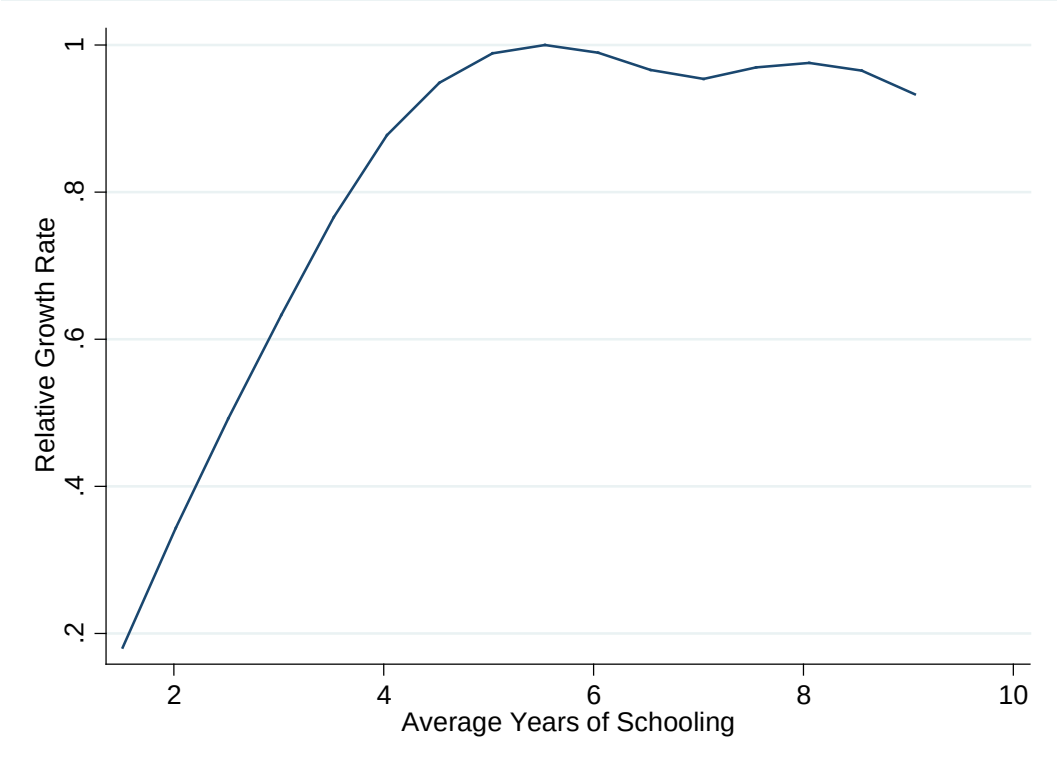

Figure 2: Additive component f_S in the growth model.

REFERENCES

Abramovich, F., I. De Fesis, and T. Sapatinas. 2009. "Optimal Testing for Additivity in Multiple Nonparametric Regression," *Annals of the Institute of Statistical Mathematics*, 61, pp. 691-714.

Buja, A., T. Hastie, and R. Tibshirani. 1989. "Linear Smoothers and Additive Models," *Annals of Statistics*, 17, pp. 453-555.

De Gooijer, J.G. and D. Zerom. 2003. "On Additive Conditional Quantiles with High Dimensional Covariates," *Journal of the American Statistical Association*, 98, pp. 135-146.

Dette, H. and C. von Lieres und Wilkau. 2001. "Testing Additivity by Kernel-Based Methods – What Is a Reasonable Test?" *Bernoulli*, 7, pp. 669-697.

Derbort, S., H. Dette, and A. Munk. 2002. "A Test for Additivity in Nonparametric Regression," *Annals of the Institute of Statistical Mathematics*, 54, pp. 60-82.

Eubank, R.L., J.D. Hart, D.G. Simpson, and L.A. Stefanski. 1995. "Testing for Additivity in Nonparametric Regression," *Annals of Statistics*, 23, pp. 1896-1920.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Gozalo, P.L. and O.B. Linton. 2001. "Testing Additivity in Generalized Nonparametric Regression Models with Estimated Parameters," *Journal of Econometrics*, 104, pp. 1-48.

Härdle,W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press

Härdle, W. H. Liang, and J. Gao. 2000. *Partially Linear Models*. New York: Springer.

Hastie, T.J. and R.J. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman and Hall.

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, pp. 246-272.

Horowitz, J.L. 2009. *Semiparametric and Nonparametric Methods in Econometrics*. New York: Springer.

Horowitz, J.L. and S. Lee. 2005. "Nonparametric Estimation of an Additive Quantile Regression Model," *Journal of the American Statistical Association*, 100, pp. 1238-1249.

Horowitz, J.L. and E. Mammen. 2004. "Nonparametric Estimation of an Additive Model with a Link Function," *Annals of Statistics*, 32, pp. 2412-2443.

Horowitz, J.L. and E. Mammen. 2007. "Rate-Optimal Estimation for a General Class of Nonparametric Regression Models with Unknown Link Functions," *Annals of Statistics*, 35, pp. 2589-2619.

Horowitz, J.L. and E. Mammen. 2011. "Oracle-Efficient Nonparametric Estimation of an Additive Model with an Unknown Link Function," *Econometric Theory*, 27, pp. 582-608.

Ichimura, H. 1993. "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics* 58, pp. 71-120.

Kim, W., O.B. Linton, and N.W. Hengartner. 1999. "A Computationally Efficient Oracle Estimator for Additive Nonparametric Regression with Bootstrap Confidence Intervals," *Journal of Computational and Graphical Statistics*, 8, pp. 278-297.

Li, Q. and J.S. Racine. 2007. *Nonparametric Econometrics*. Princeton: Princeton University Press.

Linton, O.B. (1997). "Efficient Estimation of Additive Nonparametric Regression Models," *Biometrika*, 84, pp. 469-473.

Linton, O. B. and W. Härdle. 1996. "Estimating Additive Regression Models with Known Links," *Biometrika*, 83, pp. 529-540.

Linton, O. B. and J. B. Nielsen. 1995. "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration," *Biometrika*, 82, pp. 93-100.

Liu, R., L. Yang, and W.K. Härdle. 2011. "Oracally Efficient Two-Step Estimation of Generalized Additive Model," SFB 649 discussion paper 2011-016, Humboldt-Universität zu Berlin, Germany.

Mammen, E., O. Linton, and J. Nielsen. 1999. "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions," *Annals of Statistics*, 27, pp. 1443-1490.

Newey, W.K. 1994. "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, pp. 233-253.

Nielsen, J.P. and S. Sperlich. 2005. "Smooth Backfitting in Practice," *Journal of the Royal Statistical Society, Series B*, 67, pp. 43-61.

Pagan, A. and A. Ullah. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Opsomer, J.D. 2000. "Asymptotic Properties of Backfitting Estimators," *Journal of Multivariate Analysis*, 73, pp. 166-179.

Opsomer, J.D. and D. Ruppert. 1997. "Fitting a Bivariate Additive Model by Local Polynomial Regression," *Annals of Statistics*, 25, pp. 186-211.

Severance-Lossin, E. and S. Sperlich. 1999. "Estimation of Derivatives for Additive Separable Models," *Statistics*, 33, pp. 241-265.

Song. Q. and L. Yang. 2010. "Oracally Efficient Spline Smoothing of Nonlinear Additive Autoregression Models with Simultaneous Confidence Band," *Journal of Multivariate Analysis*, 101, pp. 2008-2025.

Sperlich, S., D. Tjøstheim, and L. Yang. 2002. "Nonparametric Estimation and Testing of Interaction in Additive Models," *Econometric Theory*, 18, pp. 197-251.

Stone, C.J. 1985. "Additive Regression and Other Nonparametric Models," *Annals of Statistics*, 13, pp. 689-705.

Stock, J.H. and M.W. Watson. 2011. *Introduction to Econometrics*, 3rd edition. Boston: Pearson/Addison Wesley.

Tukey, J. 1949. "One Degree of Freedom Test for Non-Additivity," *Biometrics*, 5, pp. 232-242.

Wang, L. and L. Yang. 2007. "Spline-Backfitted Kernel Smoothing of Nonlinear Additive Autoregression Model," *Annals of Statistics*, 35, pp. 2474-2503.

Wang, X. and K.C. Carriere. 2011. "Assessing Additivity in Nonparametric Models – a Kernel-Based Method," *Canadian Journal of Statistics*, 39, pp. 632-655.

Yang, L., S. Sperlich, and W. Härdle. 2003. "Derivative Estimation and Testing in Generalized Additive Models," *Journal of Statistical Planning and Inference*, 115, pp. 521-542.

Yu, K., B.U. Park, and E. Mammen . 2008. "Smooth Backfitting in Generalized Additive Models," *Annals of Statistics*, 36, pp. 228-260.