ELSEVIER

Journal of Econometrics 79 (1997) 147–168

JOURNAL OF Econometrics

# Convergence rates and asymptotic normality for series estimators

Whitney K. Newey

*Department of Economics, MIT, E52-262D, Cambridge, MA, USA*

Received June 1995; received in revised form August 1996

**Abstract**

This paper gives general conditions for convergence rates and asymptotic normality of series estimators of conditional expectations, and specializes these conditions to polynomial regression and regression splines. Both mean-square and uniform convergence rates are derived. Asymptotic normality is shown for nonlinear functionals of series estimators, covering many cases not previously treated. Also, a simple condition for $\sqrt{n}$-consistency of a functional of a series estimator is given. The regularity conditions are straightforward to understand, and several examples are given to illustrate their application.

*Key words*: Nonparametric estimation; Series estimation; Convergence rates; Asymptotic normality
*JEL classification*: C14; C21

## 1. Introduction

Nonparametric estimation is useful in many econometric applications. Models often depend on a conditional expectation $E[y|x]$ that is unknown. For example, in demand analysis, one could model the demand function as a conditional expectation with unknown functional form (e.g. see Deaton, 1988, or Hausman and Newey, 1995). One way to estimate $E[y|x]$ is by least squares regression on $K$ approximating functions, where $K$ grows with the sample size, referred to here as series estimation. Series estimation is convenient for imposing certain restrictions on $E[y|x]$, such as additive separability (e.g. see Stone, 1985, or Andrews and Whang, 1990). Also, it is computationally convenient, because the data is summarized by a relatively few estimated coefficients.

Support was provided by the NSF for research for this paper.


*PII* S0304-4076(97)00011-0

Large sample properties of series estimators have been derived by Stone (1985), Cox (1988), Andrews and Whang (1990), Eastwood and Gallant (1991), Gallant and Souza (1991), Newey (1988, 1994a, b, 1995). This paper extends previous results on convergence rates and asymptotic normality in several ways. These extensions include convergence rates that are faster than some published ones (e.g. such as Cox, 1988) or have weaker side conditions than current results (e.g. than Newey, 1995). Also, asymptotic normality is shown for regression splines and/or nonlinear functionals that are not covered by the results of Andrews (1991), and a simple primitive condition for $\sqrt{n}$-consistency is given. In addition, the results are derived under an upper bound on the growth of number of terms $K$ that is less restrictive than previous results (e.g. those of Andrews 1991 or Newey 1995), requiring only $K^2/n \to 0$ for regression splines and $K^3/n \to 0$ for power series estimators of linear functionals. Also, all results here are derived for the case where the number of terms $K$ does not depend on the data, although some results could be extended to data dependent cases using the approaches of Newey (1995) for convergence rates and Eastwood and Gallant (1991) for asymptotic normality.

To describe a series estimator, let $g_0(x) = \mathrm{E}[y|x]$ denote the true conditional expectation and $g$ denote some function of $x$. Also, consider a vector of approximating functions

$$p^K(x) = (p_{1K}(x), \ldots, p_{KK}(x))', \tag{1}$$

having the property that a linear combination can approximate $g_0(x)$. Let $(y_i, x_i)$, $(i = 1, \ldots, n)$ denote the data. A series estimator of $g_0(x)$ is

$$\hat{g}(x) = p^K(x)'\hat{\beta}, \qquad \hat{\beta} = (P'P)^- P'Y, \quad P = [p^K(x_1), \ldots, p^K(x_n)]',$$

$$Y = (y_1, \ldots, y_n)', \tag{2}$$

where $B^-$ denotes any symmetric generalized inverse. Under conditions given below, $P'P$ will be nonsingular with probability approaching one, and hence $(P'P)^-$ will be the standard inverse.

Series estimators are convenient for imposing certain types of restrictions. For example, suppose that $\mathrm{E}[y|x]$ is additively separable in two subvectors $x_a$ and $x_b$ of $x$, so that

$$\mathrm{E}[y|x] = g_a(x_a) + g_b(x_b). \tag{3}$$

This restriction can be imposed by including in $p^K(x)$ functions that depend either on $x_a$ or $x_b$, but not on both. Additivity of $\hat{g}(x)$ then results from it being a linear combination of these functions of the separate arguments. Another example is a partially linear model, where one of the components is a linear function, as in

$$\mathrm{E}[y|x] = x_a'\gamma_0 + g_b(x_b). \tag{4}$$

This restriction can be imposed by including only $x_a$ and functions of $x_b$ in $p^K(x)$. The value of imposing either of these restrictions is that efficiency improvements result, in the sense that convergence rates are faster. The partially linear model is particularly convenient when $x_a$ has a large number of elements, e.g. when $x_a$ consists of categorical variables. We explicitly consider both types of restrictions in this paper.

There are several different types of series estimators that could be used, such as Fourier series, power series, and splines. Fourier series are not considered here because, as in Andrews (1991), it is difficult to derive primitive conditions except when $g_0(x)$ is periodic, which is not relevant for most econometric applications.[1] Primitive conditions will be given in Sections 5 and 6 for power series and regression splines.

## 2. Convergence rates

The results will follow from a few primitive, easy to interpret conditions. The first condition is

*Assumption 1.* $(y_1,x_1),\ldots,(y_n,x_n)$ are i.i.d. and $\mathrm{Var}(y|x)$ is bounded.

The bounded conditional variance assumption is difficult to relax without affecting the convergence rates. For the next condition let $\|B\| = [\mathrm{trace}(B'B)]^{1/2}$ be the Euclidean norm for a matrix B. Also, let $\mathscr{X}$ be the support of $x_i$.

*Assumption 2.* For every $K$ there is a nonsingular constant matrix $B$ such that for $P^K(x) = B_p^K(x)$; (i) the smallest eigenvalue of $\mathrm{E}[P^K(x_i)P^K(x_i)']$ is bounded away from zero uniformly in $K$ and; (ii) there is a sequence of constants $\zeta_0(K)$ satisfying $\sup_{x\in\mathscr{X}} \|P^K(x)\| \leq \zeta_0(K)$ and $K = K(n)$ such that $\zeta_0(K)^2K/n \to 0$ as $n \to \infty$.

This condition imposes a normalization on the approximating functions, bounding the second moment matrix away from singularity, and restricting the magnitude of the series terms. This condition is useful for controlling the convergence in probability of the sample second moment matrix of the approximating functions to its expectation, in the Euclidean norm. Newey (1988) and Andrews (1991) also impose bounds on the smallest eigenvalue of the second moment matrix. Primitive conditions are given below for regression splines and power series when the density of $x$ is bounded away from zero, with $\zeta_0(K)$ equal to

[1] Primitive conditions for periodic functions and evenly spaced data are given in Eastwood and Gallant (1991), but these conditions do not approximate those in most economic applications. Also, primitive conditions for Gallant's Fourier flexible form seem to require unrealistically fast approximation rates, as discussed in Gallant and Souza (1991).

$C\sqrt{K}$ and $CK$ respectively, for a constant $C$, leading to the restrictions $K^2/n \to 0$ or $K^3/n \to 0$ mentioned in the introduction.

For controlling the bias of the estimator it is useful specify a rate of approximation for the series. To do so, let $(\lambda_1, \ldots, \lambda_r)' = \lambda$ denote a vector of nonnegative integers, having the same dimension as $x$, let $|\lambda| = \sum_{j=1}^{r} \lambda_j$, and $d$ be any nonnegative integer. For a vector of functions $h(x)$ define the vector of partial derivatives $\partial^\lambda h(x) = \partial^{|\lambda|} h(x)/\partial x_1^{\lambda_1} \cdots \partial x_r^{\lambda_r}$ and let $|g|_d = \max_{|\lambda| \leq d} \sup_{x \in \mathcal{X}} |\partial^{|\lambda|} g(x)|$, where the absence of an $x$ argument for $g$ (and for $p^K$ below) denotes the entire function rather than its value at a point.

*Assumption 3.* For an integer $d \geq 0$ there are $\alpha, \beta_K$ such that $|g_0 - p^{K\prime}\beta_K|_d = O(K^{-\alpha})$ as $K \to \infty$.

This condition requires that the uniform approximation error to the function and its derivatives up to order $d$ shrink at $K^{-\alpha}$. The integer $d$ will be specified below as the order of the derivative for which a uniform convergence rate is derived. The integer $\alpha$ is related to the smoothness of the function $g_0(x)$, the dimensionality of $x$, and the size of $d$. For example, for splines and power series and $d = 0$, this assumption will be satisfied with $\alpha = s/r$, where $s$ is the number of continuous derivatives of $g_0(x)$ that exist and $r$ is the dimension of $x$.

To state the general convergence rate result, some notation is required. Let $F_0(x)$ denote the cumulative distribution function of $x_i$ and for any nonnegative integer $d$ let

$$\zeta_d(K) = \max_{|\lambda| \leq d} \sup_{x \in \mathcal{X}} \|\partial^\lambda P^K(x)\|.$$

It will be assumed throughout that $\zeta_d(K) \geq 1$ for large enough $K$, and that $\zeta_d(K)$ exists whenever it appears in the assumptions. The following result gives a general result on mean-square and uniform convergence rates.

*Theorem 1. If Assumptions* 1–3 *are satisfied with* $d = 0$ *then*

$$\int [g_0(x) - \hat{g}(x)]^2 \, dF_0(x) = O_p(K/n + K^{-2\alpha}).$$

*Also, if Assumptions* 1–3 *are satisfied for some* $d \geq 0$ *then*

$$|\hat{g} - g_0|_d = O_p(\zeta_d(K)[\sqrt{K}/\sqrt{n} + K^{-\alpha}]).$$

This result is similar to Newey (1995), except that the requirement that $K$ does not depend on the data here and the requirement that $\zeta_0(K)^2 K/n \to 0$ is different. The mean square error result is different than Andrews and Whang (1990), in applying to integrated mean square error, rather than sample mean square error, and in imposing Assumption 2.

The conclusion for mean-square error leads to optimal convergence rates for power series and splines, i.e. that attain Stone's (1982) bound. The term $K/n$ essentially corresponds to a variance term and $K^{-2\alpha}$ to a bias term. When $K$ is chosen so that these two terms go to zero at the same rate, which occurs when $K$ goes to infinity at the same rate as $n^{1/(1+2\alpha)}$ (and the side condition $\zeta_0(K)^2K/n \to 0$ is satisfied), the convergence rate will be $n^{-\alpha/(1+2\alpha)}$. For power series and splines, where $\alpha = s/r$, the rate will be $n^{-r/(r+2s)}$, which equals Stone's (1982) bound.

The uniform convergence rates will not be optimal. For example, for splines $\zeta_d(K) = K^{(1/2)+d}$, so that Theorem 1 gives $|\hat{g} - g_0|_0 = O_p(K/\sqrt{n} + K^{1/2-s/r})$. This rate cannot attain Stone's (1982) bound on the best obtainable rate. Nevertheless, these uniform convergence rates improve on some in the literature, e.g. on Cox (1988). Also, it does not yet seem to be known whether it is possible to attain the optimal uniform convergence rates using a series estimator.

## 3. Asymptotic normality

There are many applications where a functional of a conditional expectation is of interest. For example, a common practice in demand analysis is estimation of the demand function in log linear form, where $y$ is the log of consumption, $x$ is a two dimensional vector with first argument equal to the log of price and second the log of income, and the estimated demand function is $e^{\hat{g}(x)}$. A functional of interest in demand analysis is approximate consumer surplus, equal to the integral of the demand function over a range of prices. For a fixed income $\bar{I}$ an estimator of this functional would be

$$\hat{\theta} = \int_{\underline{p}}^{\bar{p}} e^{\hat{g}(\ln t, \ln \bar{I})}\, dt. \tag{5}$$

An asymptotic normality result for this functional could be useful in constructing approximate confidence intervals and tests.

This section gives conditions for asymptotic normality of functionals of series estimates. In this section we focus on the 'slower than $1/\sqrt{n}$' case, and discuss the $\sqrt{n}$-consistent case in the next section. To describe the results, let $a(\cdot)$ be a vector functional of $g$, i.e. a mapping from a possible conditional expectation function to a real vector. The estimator will be assumed to take the form

$$\hat{\theta} = a(\hat{g}). \tag{6}$$

For example, the approximate consumer surplus estimator satisfies this equation with $a(g) = \int_{\underline{p}}^{\bar{p}} e^{g(\ln t, \ln \bar{I})}\, dt$. The true value corresponding to this estimator

will be

$$\theta_0 = a(g_0), \tag{7}$$

where $g_0$ denotes the true conditional expectation function.

To use $\hat{\theta}$ for approximate inference procedures, it is important to have an asymptotic variance estimator. Such can be formed from a delta-method estimator of the variance of $\hat{\theta}$ as a function of the estimated coefficients $\hat{\beta}$. Let

$$\hat{A} = \partial a(p^K'\beta)/\partial\beta|_{\beta=\hat{\beta}},$$

when $\hat{A}$ exists, and otherwise let $\hat{A}$ be any vector with the same dimension as $\beta$. The regularity conditions given below will imply that $\hat{A}$ exists with probability that approaches one in large samples. Let

$$\hat{V} = \hat{A}'\hat{Q}^-\hat{\Sigma}\hat{Q}^-\hat{A}, \qquad \hat{Q} = P'P/n,$$

$$\hat{\Sigma} = \sum_{i=1}^{n} p^K(x_i)p^K(x_i)'[y_i - \hat{g}(x_i)]^2/n. \tag{8}$$

This estimator is just the usual one for a nonlinear function of least squares coefficients. The vector $\hat{A}$ is a Jacobian term, and $\hat{Q}^-\hat{\Sigma}\hat{Q}^-$ is the White (1980) estimator of the least squares asymptotic variance for a possibly misspecified model. This estimator will lead to correct asymptotic inferences because it accounts properly for variance, and because bias will be small relative to variance under the regularity conditions discussed below.

Some additional conditions are important for the asymptotic normality result.

*Assumption 4.* $\mathrm{E}[\{y - g_0(x)\}^4|x]$ is bounded, and $\mathrm{Var}(y|x)$ is bounded away from zero.

This assumption requires that the fourth conditional moment of the error is bounded, strengthening Assumption 1. The next one is a smoothness condition on $a(g)$, requiring that it be approximated sufficiently well by a linear functional when $g$ is close to $g_0$.

*Assumption 5.* Either (a) $a(g)$ is linear in $g$, or; (b) for $d$ as in Assumption 3, $\zeta_d(K)^4K^2/n \to 0$ and there exists a function $D(g;\tilde{g})$ that is linear in $g$ and such that for some $C$, $\varepsilon > 0$ and all $\tilde{g}, \bar{g}$ with $|\tilde{g} - g_0|_d < \varepsilon$, $|\bar{g} - g_0|_d < \varepsilon$, it is true that $\|a(g) - a(\tilde{g}) - D(g - \tilde{g};\tilde{g})\| \leq C(|g - \tilde{g}|_d)^2$ and $\|D(g;\tilde{g}) - D(g;\bar{g})\| \leq L|g|_d|\tilde{g} - \bar{g}|_d$.

The interpretation of $D(g;\tilde{g})$ is that it is a functional derivative of $a(g)$. Indeed, this assumption implies that $a(g)$ is Frechet differentiable in $g$ with respect to the norm $|g|_d$.

This condition is often straightforward to verify. In the consumer surplus example it is easy to check that it will be satisfied with $d=0$ and

$$D(g;\tilde{g})=\int_{\underline{p}}^{\bar{p}} g(\ln t,\ln \bar{I})e^{\tilde{g}(\ln t,\ln \bar{I})}\,dt, \tag{9}$$

as long as $[\ln \underline{p},\ln \bar{p}]x\{\ln \bar{I}\}$ is contained in the support for $x$. To see that this is so, note that for $|\tilde{g}-g_0|_0<\varepsilon$ and $|g-g_0|_0<\varepsilon$, $\tilde{g}$ and $g$ will be uniformly bounded on the range of integration. Also, for any scalar $z$, $d^j e^z/dz^j=e^z$, so that a mean-value expansion of $e^z$ around some other point $\tilde{z}$ gives $|e^z-e^{\tilde{z}}-e^{\tilde{z}}(z-\tilde{z})|\leq C|z-\tilde{z}|^2$ for some constant $C$ when $z$ and $\tilde{z}$ are in some bounded set. Therefore,

$$\begin{aligned}\|a(g)-a(\tilde{g})-D(g-\tilde{g};\tilde{g})\| &\leq \int_{\underline{p}}^{\bar{p}} C|g(\ln p,\ln \bar{I})-\tilde{g}(\ln p,\ln \bar{I})|^2\,dp\\ &\leq C(\bar{p}-\underline{p})(|g-\tilde{g}|_0)^2.\end{aligned}$$

The next requirement imposes some continuity conditions on the derivative. Let $D(g)=D(g;g_0)$ denote the derivative at $\tilde{g}=g_0$.

*Assumption 6.* $a(g)$ is a scalar, there exists $C$ such that $|D(g)|\leq C|g|_d$ for $d$ from Assumption 3, and there exists $g_K(x)=p^K(x)'\tilde{\beta}_K$ such that $E[g_K(x)^2]\to 0$ and $D(g_K)$ is bounded away from zero.

This assumption says that the derivative is continuous in $|g|_d$, but *not* in the mean-square norm $(E[g(x)^2])^{1/2}$. The lack of mean-square continuity will imply that the estimator $\hat{\theta}$ is not $\sqrt{n}$-consistent, and is also a useful regularity condition. Another restriction imposed is that $a(g)$ is a scalar, which is general enough to cover many cases of interest. When $a(g)$ is a vector asymptotic normality with an estimated covariance matrix would follow from Assumption $J$ (iii) of Andrews (1991), which is difficult to verify. In contrast, Assumption 6 is a primitive condition, that is relatively easy to verify.

For example, for the consumer surplus estimator previously discussed where $p^K(x)$ is a power series, suppose that $x$ is continuously distributed with compact support and bounded density, and that the set $\mathscr{P}=\{(\ln p,\ln \bar{I})\colon \underline{p}\leq p\leq \bar{p}\}$ is contained in this support. Then there exists a sequence of continuous functions $g_J(x)$ such that $g_J(x)$ is equal to $1/(\bar{p}-\underline{p})$ on the line $\mathscr{P}$, is uniformly bounded, and converges to zero everywhere else. For this sequence, $D(g_J)=\int_{\underline{p}}^{\bar{p}} g_J(\ln p,\ln \bar{I})\times \exp[g_0(\ln p,\ln \bar{I})]\,dp\geq C\int_{\underline{p}}^{\bar{p}} g_J(\ln p,\ln \bar{I})\,dp\geq \tilde{C}$ for constants $C$, $\tilde{C}>0$ and $E[g_J(x)^2]\to 0$. By the Weierstrass approximation theorem, any $g_J(x)$ can be approximated uniformly by a polynomial $q_J(x)$, so that $D(q_J(x))>C/2$, and $E[q_J(x)^2]\to 0$. Since a polynomial is a linear combination of a power series it follows that Assumption 6 is satisfied in this example.

To state the asymptotic normality result it is useful to work with an asymptotic variance formula. Let $\sigma^2(x) = \mathrm{Var}(y \mid x)$ and

$$A = (D(p_{1K}), \ldots, D(p_{KK}))'.$$

The asymptotic variance formula is

$$V_K = A'Q^{-1}\Sigma Q^{-1}A, \qquad Q = \mathrm{E}[p^K(x)p^K(x)'],$$
$$\Sigma = \mathrm{E}[p^K(x)p^K(x)'\sigma(x)^2]. \tag{10}$$

*Theorem 2. If Assumptions* 1–6 *are satisfied and* $\sqrt{n}K^{-\alpha} \to 0$ *then* $\hat{\theta} = \theta_0 + \mathrm{O}_p(\zeta_d(K)/\sqrt{n})$ *and*

$$\sqrt{n}V_K^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0,1), \qquad \sqrt{n}\hat{V}^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0,1).$$

This theorem includes as a special case linear functionals, which were studied by Andrews (1991). In that case the conditions of Theorem 2 are quite simple relative to Andrews (1991) conditions. Also, the restriction $\zeta_0(K)^2K/n \to 0$ is all that is needed in the linear case, because Assumption 5(b) is not required. This leads to the requirement that $K^3/n \to 0$ for power series (where $\zeta_0(K) = CK$), which is weaker than the $K^6/n \to 0$ requirement of Andrews (1991). One reason for this contrast, is that the regressors are assumed to be i.i.d. here, while Andrews (1991) general results allows for the regressors to not be identically distributed.

This result only gives an upper bound $\mathrm{O}_p(\zeta_d(K)/\sqrt{n})$ on the convergence rate for $\hat{\theta}$. This bound may not be sharp. We know it will not be sharp in the $\sqrt{n}$-consistent case considered next, where $d = 0$. Whether it is sharp for other cases is still an open question.

## 4. $\sqrt{n}$-Consistency

The key condition for $\sqrt{n}$-consistency is that the derivative $D(g)$ be mean-square continuous, as specified in the following hypothesis.

*Assumption 7.* There is $v(x)$ with $\mathrm{E}[v(x)v(x)']$ finite and nonsingular such that $D(g_0) = \mathrm{E}[v(x)g_0(x)]$, $D(p_{kK}) = \mathrm{E}[v(x)p_{kK}(x)]$ for all $k$ and $K$, and there is $\tilde{\beta}_K$ with $\mathrm{E}[\|v(x) - \tilde{\beta}_K p^K(x)\|^2] \to 0$.

This condition allows for $a(g)$ to be a vector. It requires a representation of $a(g)$ as an expected outer product, when $g$ is equal to the truth or any of the approximating functions, and for the functional $v(x)$ in the outer product representation to be approximated in mean-square by some linear combination of the functions. This condition and Assumption 6 are mutually exclusive, and together cover most cases of interest (i.e. they seem to be exhaustive).

A sufficient condition for Assumption 7 is that the functional $a(g)$ be mean-square continuous in $g$ over some linear domain that includes the truth and the approximating functions, and that the approximation functions form a basis for this domain. The outer product representation in Assumption 7 will then follow from the Riesz representation theorem. This condition is somewhat like Van der Vaart's (1991) condition for $\sqrt{n}$-consistent estimability of functionals, except that his mean-square continuity hypothesis pertains to the set of scores rather than a set of conditional expectations. Also, here it is a sufficient condition for $\sqrt{n}$-consistency of a particular estimator, rather than a necessary condition for existence of such an estimator. A similar condition was also used by Newey (1994a) to derive primitive conditions for $\sqrt{n}$-consistency of series estimators, under stronger regularity conditions.

There are many interesting examples of functionals that satisfy this condition. One example is an average consumer surplus estimator, like that of Eq. (5), but integrated over income. Consider the functional

$$a(g) = \int_{\underline{I}}^{\bar{I}} v(I) \int_{\underline{p}}^{\bar{p}} \exp(g(\ln p, \ln I))\, \mathrm{d}p\, \mathrm{d}I, \tag{11}$$

where $v(I)$ is some weight function, with $\int v(I)\,\mathrm{d}I = 1$. This is an average of consumer surplus over different income values. An estimator of this functional could be used as a summary of consumer surplus as a function of income. Similarly to Eq. (9), the derivative of this functional is

$$D(g) = \int v(I, p) g(\ln p, \ln I)\, \mathrm{d}p\, \mathrm{d}I,$$

$$v(I, p) = 1(\underline{p} \leqslant p \leqslant \bar{p}, \underline{I} \leqslant I \leqslant \bar{I}) v(I) \exp(g_0(\ln p, \ln I)). \tag{12}$$

Assumption 7 will then be satisfied, with $v(x) = f(I, p)^{-1} v(I, p)$, where $f(I, p)$ is the density for $I$ and $p$, assumed to be bounded away from zero on $[\underline{p}, \bar{p}]x[\underline{I}, \bar{I}]$.

Another example is the coefficients $\gamma_0$ from the partially linear model of Eq. (4). Suppose that $\mathrm{E}[\mathrm{Var}(x_a \mid x_b)]$ is nonsingular, an identification condition for $\gamma_0$, and let $U = x_a - \mathrm{E}[x_a \mid x_b]$ and $v(x) = (\mathrm{E}[UU'])^{-1} U$. Then for $a(g) = \mathrm{E}[v(x) g(x)]$,

$$a(g_0) = \mathrm{E}[v(x) g_0(x)] = (\mathrm{E}[UU'])^{-1} (\mathrm{E}[U x_a'] \gamma_0 + \mathrm{E}[U g_b(x_b)]) = \gamma_0. \tag{13}$$

In this example $a(g)$ is a linear functional of $g$, and $a(g) = D(g)$ satisfies Assumption 7 by construction.

A third example is a weighted average derivative, where

$$a(g) = \int w(x) [\partial g(x)/\partial x]\, \mathrm{d}x, \qquad \int w(x)\, \mathrm{d}x = 1, \quad w(x) \geqslant 0. \tag{14}$$

This functional is useful for estimating scaled coefficients of an index model, as discussed in Stoker (1986), and can also be used to quantify the average slope

of a function. Assuming that $w(x)$ is zero outside some compact set and that $x$ is continuously distributed with density $f(x)$ that is bounded away from zero on the set where $w(x)$ is positive, integration by parts gives

$$a(g) = -\int [\partial w(x)/\partial x] g(x)\,dx = \mathrm{E}[v(x)g(x)],$$

$$v(x) = -f(x)^{-1}\partial w(x)/\partial x. \tag{15}$$

Therefore, Assumption 7 is satisfied for this functional, and hence Theorem 3 below will give $\sqrt{n}$-consistency of a series estimator of this functional.

The asymptotic variance of the estimator will be determined by the function $v(x)$ from Assumption 7. It will be equal to

$$V = \mathrm{E}[v(x)v(x)'\mathrm{Var}(y\,|\,x)].$$

*Theorem 3. If Assumptions* 1–5 *and* 7 *are satisfied for* $d = 0$, *and* $\sqrt{n}K^{-\alpha} \to 0$ *then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0, V), \quad \hat{V} \xrightarrow{p} V.$$

## 5. Power series

One particular type of series for which primitive regularity conditions can be specified is a power series. Suppose that $r$ is the dimension of $x$, let $\lambda = (\lambda_1, \ldots, \lambda_r)'$ denote a vector of nonnegative integers, i.e. a multi-index, with norm $|\lambda| = \sum_{j=1}^{r} \lambda_j$, and let $z^{\lambda} \equiv \prod_{j=1}^{r} (z_j)^{\lambda_j}$. For a sequence $(\lambda(k))_{\ell=1}^{\infty}$ of distinct such vectors, a power series approximation has

$$p_{kK}(x) = x^{\lambda(k)}. \tag{16}$$

It will be assumed that the multi-index sequence is ordered with degree $|\lambda(K)|$ increasing in $K$.

The theory to follow uses orthonormal polynomials, which may also have computational advantages. For the first step, replacing each power $x^{\lambda}$ by the product of orthonormal polynomials of order corresponding to components of $\lambda$, with respect to some distribution may lead to reduced collinearity, particularly when the distribution closely matches that of the data. The estimator will be numerically invariant to such a replacement, because $|\lambda(\ell)|$ is monotonically increasing.

For power series primitive conditions for Assumptions 2 and 3 can be specified. Assumption 2 will be a consequence of the following condition.

*Assumption 8.* The support of $x$ is a Cartesian product of compact connected intervals on which $x$ has a probability density function that is bounded away from zero.

This assumption can be relaxed by specifying that it only holds for a component of the distribution of $x$ (which would allow points of positive probability in the support of $x$), but it appears difficult to be more general. Under this condition one can use well known properties of orthogonal polynomials to obtain the explicit bound $\zeta_d(K) = K^{1+2d}$, which is important for obtaining explicit convergence rates and conditions for asymptotic normality.

To be specific about Assumption 3 we need uniform approximation rates for power series. These rates will follow from the following smoothness assumption.

*Assumption 9.* $g_0(x) = \mathrm{E}[y \mid x]$ is continuously differentiable of order $s$ on the support of $x$.

It follows from this condition and Lorentz (1986) that for $d = 0$, the approximation rate of Assumption 3 is $\alpha = s/r$, where $r$ is the dimension of $x$. A literature search has not yet revealed a corresponding result for $d > 0$, where derivatives are also approximated, except in two cases. When $x$ is univariate, so $r = 1$, then it is well known that Assumption 3 will be satisfied with $\alpha = s - d$. Also, when $g_0(x)$ is analytical it is known that Assumption 3 will hold for any $d$ with $\alpha$ equal to an arbitrarily large positive number. Thus, Theorem 1 could be applied to obtain uniform convergence rates for power series in these cases, with $\zeta_d(K) = K^{1+2d}$ and $\alpha$ either equal to $s - d$, in the univariate case, or $\alpha$ equal to any positive number, in the analytical $g_0(x)$ case. For simplicity, we limit attention to $d = 0$ rather than spelling out the different cases in a Theorem.

The first result for polynomials gives convergence rates.

*Theorem 4. For power series, if Assumptions* 1, 8, *and* 9 *are satisfied and* $K^3/n \to 0$ *then*

$$\int [g_0(x) - \hat{g}(x)]^2 \, dF_0(x) = \mathrm{O}_\mathrm{p}(K/n + K^{-2s/r}),$$

$$|\hat{g} - g_0|_0 = \mathrm{O}(K[\sqrt{K}/\sqrt{n} + K^{-s/r}]).$$

This result gives mean-square and uniform approximation rates for $\hat{g}$. The results could be extended to uniform convergence rates for derivatives when $x$ is univariate or $g_0(x)$ is analytical, as discussed above.

As previously noted, the mean-square approximation rate attains Stone's (1982) bounds for $K = Cn^{r/(r+2s)}$ and $s \geq r$ (implying $K^3/n \to 0$). The mean-square rate obtained here is different than that of Andrews and Whang (1990), because it is for the population mean-square error rather than the sample mean-square error, which is appealing because it is a fixed norm rather than one that changes with the sample size and configuration of the regressors. On the other hand, Assumption 8 is stronger than the conditions imposed in Andrews and Whang (1990).

The uniform approximation rate does not appear to be optimal, although it seems to improve on rates existing in the literature, being faster than that of Cox (1988). One implication of this rate is that $\hat{g}$ will be uniformly consistent when $s \geq r$ and $K^3/n \to 0$.

This result can also be extended to additive models, such as the one in Eq. (3). A power series estimator could be constructed as described above, except that products of powers from both $x_a$ and $x_b$ are excluded. If each of $g_a(x_a)$ and $g_b(x_b)$ are continuously differentiable of order $s$, the exclusion of the (many) interaction terms will increase the approximation rate to $K^{-s/\bar{r}}$, where $\bar{r}$ is the maximum dimension of $x_a$ and $x_b$. The conclusion of Theorem 4 will then hold with $r$ replaced by $\bar{r}$. This gives mean-square convergence rates for polynomials like the regression spline rates of Stone (1985). An extension to the partially linear case would also be similar, with $r$ replaced by the dimension of $x_b$ and Assumption 8 only required to hold for $x_b$ rather than $x$. Because these extensions are straightforward, explicit results are not given.

Assumptions 8 and 9 are primitive conditions for Assumptions 2 and 3, so that asymptotic normality and $\sqrt{n}$-consistency results will follow under the other conditions of Sections 3 and 4. All of these other conditions are primitive, are straightforward to verify, except for Assumption 6. The following more primitive version of Assumption 6 is easier to check and will suffice for Assumption 6 for power series.

*Assumption 10.* $a(g)$ is a scalar and there exists a sequence of continuous functions $\{g_J(x)\}_{J=1}^{\infty}$ such that $a(g_J)$ is bounded away from zero but $\mathrm{E}[g_J(x)^2] \to 0$.

An asymptotic normality result for power series estimators is given by the following:

*Theorem 5. For power series, if Assumptions* 1 *and* 4 *are satisfied, Assumptions* 8–10 *are satisfied,* $\sqrt{n}K^{-s/r} \to 0$, *either* $K^6/n \to 0$ *or* $a(g)$ *is linear and* $K^3/n \to 0$, *then* $\hat{\theta} = \theta_0 + \mathrm{O}_{\mathrm{p}}(K/\sqrt{n})$ *and* $\sqrt{n}\hat{V}^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0,1)$.

The rate conditions of this result require that $s > 3r/2$ in the case of a linear functional, or that $s > 3r$ in the case of a nonlinear functional. In this sense a certain amount of smoothness of the regression function is required for asymptotic normality.

For the consumer surplus example, it has already been shown that Assumptions 4 and 10 are satisfied. Since it is nonlinear, and since the dimension of $x$ is 2 in this case, asymptotic normality will follow from Assumptions 8 and 9 and from $K^6/n \to 0$ and $\sqrt{n}K^{-s/2} \to 0$.

A result can also be formulated for $\sqrt{n}$-consistency of a functional of a polynomial regression estimator.

*Theorem 6. For power series, if Assumptions* 1 *and* 4 *are satisfied, Assumptions* 7–9 *are satisfied,* $\sqrt{n}K^{-s/r} \to 0$, *either* $K^6/n \to 0$ *or* $a(g)$ *is linear and* $K^3/n \to 0$, *then* $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0, V)$ *and* $\hat{V} \to V$.

The conditions of this result are very similar to those of Theorem 5, except that Assumption 10 has been replaced by Assumption 7. More primitive conditions for Assumption 7 are not given because it is already in a form which is straightforward to verify, as demonstrated by the examples of Section 4. The average consumer surplus estimator satisfies Assumptions 4 and 7 with $d = 0$, so that under Assumptions 8 and 9 and $K^6/n \to 0$ and $\sqrt{n}K^{-s/2} \to 0$ it will be $\sqrt{n}$-consistent. Also, the partially linear coefficients and weighted average derivative examples also satisfy the assumptions, and are linear functionals, so that they will be $\sqrt{n}$-consistent if $K^3/n \to 0$ and $\sqrt{n}K^{-s/r} \to 0$.

## 6. Regression splines

Regression splines are linear combinations of functions that are smooth piecewise polynomials of a given order with fixed knots (join points). Their properties as approximating functions are described in Powell (1981) and there use in series estimation has been considered by Stone (1985) and Chen (1988) among others. Spline approximating functions have attractive features relative to polynomials, being less sensitive to outliers and to bad approximation over small regions. They have the disadvantage, as far as the theory here is concerned, that support of $x$ must be known. A known support is required for knot placement. Alternatively, the results could also be obtained for estimators that only use data where $x$ is in some specified region, but the conclusion would have to be modified for that case. In particular, the uniform and mean-square convergence rates would only apply to the true function over the region of the data that is used, and asymptotic normality would only hold for functionals that did not depend on the function $g_0$ for $x$ outside the region. For simplicity, this section will only consider the case where the support is known and satisfies Assumption 8, where the following condition can be imposed without loss of generality.

*Assumption 11.* The support of $x$ is $[-1, 1]^r$.

When the support of $x$ is known and Assumption 8 is satisfied, $x$ can always be rescaled so that this condition holds.

To describe a spline, let $(x)_+ = 1(x > 0) \cdot x$. A spline basis for a univariate $m$th degree polynomial with $L - 1$ knots is

$$p_\ell(x, L) = \begin{cases} x^{\ell - 1}, & 1 \leqslant \ell \leqslant m + 1, \\ \{[x + 1 - 2(\ell - m - 1)/L]\}_+^m, & m + 2 \leqslant \ell \leqslant m + L. \end{cases} \tag{17}$$

A spline basis can be formed from products of these functions for the individual components of $x$. For $\{\lambda(k,K)\}$ the set of distinct $r$-tuples of nonnegative integers with $\lambda_j(k,K) \leqslant m+J$ for each $j$ and $\ell$, and $K=(m+J)^r$, let

$$p_{kK}(x) = \prod_{j=1}^{r} p_{\lambda_j(k)}(x_j, L) \quad (k=1,\ldots,K). \tag{18}$$

The theory to follow uses $B$-splines, which are a linear transformations of the above functions that have lower multicollinearity. Well known properties of $B$-splines lead to the explicit bound $\zeta_d(K) = K^{(1/2)+d}$. The low multicollinearity of $B$-splines and recursive formula for calculation also leads to computational advantages; e.g. see Powell (1981).

The rate at which splines uniformly approximate a function is the same as that for power series, so the smoothness condition of Assumption 9 will be left unchanged. For splines there does not seem to be approximation rates for derivatives readily available in the literature, except for the univariate case, where $\alpha = s-d$ as with power series. For simplicity we will focus attention on the $d=0$ case.

*Theorem 7. For splines, if Assumptions* 1, 8, 9, *and* 11 *are satisfied then*

$$\int [g_0(x) - \hat{g}(x)]^2 \, dF_0(x) = O_p(K/n + K^{-2s/r}),$$

$$|\hat{g} - g_0|_0 = O(\sqrt{K}[\sqrt{K}/\sqrt{n} + K^{-s/r}]).$$

It is interesting to note that the uniform convergence rate for splines is faster than power series. It may be that this is an artifact of the proof, rather than some intrinsic feature of the two types of approximation, but more work is needed to determine that.

Asymptotic normality and consistency will follow similarly as for power series.

*Theorem 8. For splines, if Assumptions* 1 *and* 4 *are satisfied, Assumptions* 8–11 *are satisfied,* $\sqrt{n}K^{-s/r} \to 0$, *either* $K^4/n \to 0$ *or* $a(g)$ *is linear and* $K^2/n \to 0$, *then* $\hat{\theta} = \theta_0 + O_p(K/\sqrt{n})$ *and* $\sqrt{n}\hat{V}^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,1)$.

In comparison with Theorem 5, this asymptotic normality result for functionals of a spline estimator imposes less stringent conditions on the growth rate $K$. Consequently, the necessary smoothness conditions for asymptotic normality will be less severe, requiring only $s > 2r$ for a nonlinear functional or $s > r$ for a linear one. As in the case of polynomials, the consumer surplus functional satisfies the conditions of this result, so that it will be asymptotically normal for $K^4/n \to 0$ and $\sqrt{n}K^{-s/2} \to 0$.

A $\sqrt{n}$-consistency result for splines can also be formulated.

*Theorem 9. For splines, if Assumptions* 1 *and* 4 *are satisfied, Assumptions* 7–9 *and* 11 *are satisfied,* $\sqrt{n}K^{-s/r} \to 0$, *either* $K^4/n \to 0$ *or* $a(g)$ *is linear and* $K^2/n \to 0$, *then* $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0, V)$ *and* $\hat{V} \to V$.

The examples of Section 4 meet the conditions of this result, and so for regression splines, the average consumer surplus estimator will be $\sqrt{n}$-consistent if $K^4/n \to 0$ and $\sqrt{n}K^{-s/2} \to 0$, and the partially linear coefficients and weighted average derivative will be $\sqrt{n}$-consistent if $K^2/n \to 0$ and $\sqrt{n}K^{-s/r} \to 0$.

## 7. Conclusion

This paper has derived convergence rate and asymptotic normality results for series estimators. Primitive conditions were given for power series and regression splines. Further refinements of these results would be useful. It would be good to have results for estimation of derivatives of functions. These could easily be obtained from approximation rates for power series or splines. Also, it would be useful to know whether the uniform convergence rates could be improved on.

## Appendix. Proofs of Theorems

Throughout the Appendix, let $C$ denote a generic constant that may be different in different uses and $\sum_i = \sum_{i=1}^n$. Also, before proving Theorems 1–3, it is helpful to make the following observations that apply to each of those proofs. Since $\hat{g}$ is invariant to nonsingular linear transformations of $p^K(x)$, it can be assumed throughout that $B = I$, i.e. $p^K(x) = P^K(x)$. Furthermore, since $Q = \mathrm{E}[p^K(x_i)p^K(x_i)']$ has smallest eigenvalue bounded away from zero, it can be assumed throughout that $Q = I$. This is because, for a symmetric square root $Q^{-1/2}$ of $Q^{-1}$, $Q^{-1/2}p^K(x)$ is a nonsingular linear transformation of $p^K(x)$ satisfying

$$\tilde{\zeta}_d(K) = \sup_{x \in \mathscr{X}, |\lambda| \leq d} \|\partial^\lambda Q^{-1/2} p^K(x)\| \leq C\zeta_d(K).$$

Hence, the conditions imposed on $K$ in Assumptions 2 and 5 will be satisfied when $Q^{-1/2}p^K(x)$ replaces $p^K(x)$, and convergence rate bounds in terms of $\tilde{\zeta}_d(K)$ derived for this replacement will also hold for the original $\zeta_d(K)$.

*Proof of Theorem 1.* First, for $\hat{Q} = P'P/n$,

$$\mathrm{E}[\|\hat{Q} - I\|^2] = \sum_{k=1}^K \sum_{j=1}^K \mathrm{E}[\{\sum_{i=1}^n p_{kK}(x_i)p_{jK}(x_i)/n - I_{jk}\}^2]$$

$$\leqslant \sum_{k=1}^{K}\sum_{j=1}^{K} \mathrm{E}[p_{kK}(x_i)^2 p_{jK}(x_i)^2]/n$$

$$= \mathrm{E}[\sum_{k=1}^{K} p_{kK}(x_i)^2 \sum_{j=1}^{K} p_{jK}(x_i)^2]/n$$

$$\leqslant \zeta_0(K)^2 \operatorname{tr}(I)/n$$

$$\leqslant \zeta_0(K)^2 K/n \to 0,$$

so that

$$\|\hat{Q} - I\| = \mathrm{O_p}(\zeta_0(K)K^{1/2}/\sqrt{n}) = \mathrm{o_p}(1). \tag{A.1}$$

Furthermore, since the difference in smallest eigenvalues is bounded in absolute value by $\|\hat{Q} - I\|$, the smallest eigenvalue of $\hat{Q}$ converges in probability to one. Let $1_n$ be the indicator function for the smallest eigenvalue of $\hat{Q}$ being greater than $1/2$, so $\mathrm{Prob}(1_n = 1) \to 1$.

Next, for $G = (g_0(x_1), \ldots, g_0(x_n))'$ let $\varepsilon = Y - G$ and $X = (x_1, \ldots, x_n)$. Boundedness of $\mathrm{Var}(y \mid x)$ and independence of the observations implies $\mathrm{E}[\varepsilon\varepsilon' \mid X] \leqslant CI$, where the inequality denotes the usual positive semi-definite semi-order, with $A \leqslant B$ meaning that $B - A$ is positive semidefinite. Therefore, for $G = (g_1(x_1), \ldots, g_0(x_n))'$,

$$\begin{aligned}
\mathrm{E}[1_n\|\hat{Q}^{-1/2}P'\varepsilon/n\|^2|X] &= 1_n\mathrm{E}[\varepsilon' P(P'P)^{-1}P'\varepsilon \mid X]/n \\
&= 1_n\mathrm{E}[\operatorname{tr}\{P(P'P)^{-1}P'\varepsilon\varepsilon'\} \mid X]/n \\
&= 1_n \operatorname{tr}\{P(P'P)^{-1}P'\mathrm{E}[\varepsilon\varepsilon \mid X]\}/n \\
&\leqslant C 1_n \operatorname{tr}\{P(P'P)^{-1}P'\}/n \leqslant CK/n,
\end{aligned}$$

so by the Markov inequality, $1_n\|\hat{Q}^{-1/2}P'\varepsilon/n\| = \mathrm{O_p}(K^{1/2}/\sqrt{n})$. It follows that

$$\begin{aligned}
1_n\|\hat{Q}^{-1}P'\varepsilon/n\| &= 1_n\{(\varepsilon'P/n)\hat{Q}^{-1/2}\hat{Q}^{-1}\hat{Q}^{-1/2}P'\varepsilon/n\}^{1/2} \\
&\leqslant \mathrm{O_p}(1)1_n\|\hat{Q}^{-1/2}P'\varepsilon/n\| = \mathrm{O_p}(K^{1/2}/\sqrt{n}).
\end{aligned}$$

Let $\beta$ be as in Assumption 3. Then by $1_nP(P'P)^{-1}P'$ idempotent,

$$\begin{aligned}
1_n\|\hat{Q}^{-1}P'(g - P\beta)/n\| &\leqslant \mathrm{O_p}(1)1_n[(G - P\beta)'P(P'P)^{-1}P'(G - P\beta)/n]^{1/2} \\
&\leqslant \mathrm{O_p}(1)[(G - P\beta)'(Gg - P\beta)/n]^{1/2} = \mathrm{O_p}(K^{-\alpha}).
\end{aligned}$$

Therefore, by $1_n(\hat{\beta}-\beta)=1_n\hat{Q}^{-1}P'(y-G)/n+1_n\hat{Q}^{-1}P'(G-P\beta)/n$, it follows that

$$\begin{aligned}1_n\|\hat{\beta}-\beta\| &\leqslant 1_n\|\hat{Q}^{-1}P'\varepsilon/n\|+1_n\|\hat{Q}^{-1}P'(G-P\beta)/n\| \\ &= O_p(K^{1/2}/\sqrt{n}+K^{-\alpha}). \end{aligned} \tag{A.2}$$

Next, by the triangle inequality,

$$\begin{aligned}&1_n\int[\hat{g}(x)-g_0(x)]^2\,dF_0(x) \\ &= 1_n\int[p^K(x)'(\hat{\beta}-\beta)+p^K(x)'\beta-g_0(x)]^2dF_0(x) \\ &\leqslant 1_n\|\hat{\beta}-\beta\|^2+1_n\int[p^K(x)'\beta-g_0(x)]^2\,dF_0(x) \\ &= O_p(K/n+K^{-2\alpha})+O(K^{-2\alpha})=O_p(K/n+K^{-2\alpha}). \end{aligned} \tag{A.3}$$

Therefore, the first conclusion follows by $1_n=1$ with probability approaching one. Also, by the triangle and Cauchy–Schwartz inequalities,

$$\begin{aligned}1_n|\hat{g}-g_0|_d &\leqslant 1_n|p^{K\prime}(\hat{\beta}-\beta)|_d+|p^{K\prime}\beta-g_0|_d \\ &\leqslant \zeta_d(K)1_n\|\hat{\beta}-\beta\|+O(K^{-\alpha}) \\ &= O_p(\zeta_d(K)[(K^{1/2}/\sqrt{n})+K^{-\alpha}]). \end{aligned} \tag{A.4}$$

The second conclusion follows from this equation like the first follows from Eq. (A.3). QED.

*Proof of Theorem 2.* Note that Assumption 5(b) implies

$$\begin{aligned}\sqrt{n}[\zeta_d(K)(\sqrt{K}/\sqrt{n}+K^{-\alpha})]^2 &= \{(\zeta_d(K)^4K^2/n)^{1/2} \\ &\quad +(\sqrt{n}K^{-\alpha})^2[\zeta_d(K)^4/n]^{1/2}\}\to 0, \\ \zeta_d(K)^2(\sqrt{K}/\sqrt{n}+K^{-\alpha}) &= [\zeta_d(K)^4K/n]^{1/2} \\ &\quad +L[\zeta_d(K)^4/n]^{1/2}\sqrt{n}K^{-\alpha}\to 0. \end{aligned} \tag{A.5}$$

Let $F=V_K^{-1/2}=(A'\Sigma A)^{1/2}$. Then by Cauchy–Schwartz inequality, for $g_K$ as in Assumption 6, $|D(g_K)|=|A'\tilde{\beta}_K|\leqslant\|A\|\|\tilde{\beta}_K\|=\|A\|(E[g_K(x_i)^2])^{1/2}$, implying $\|A\|\to\infty$. Further, $\Sigma\geqslant CI$ by $\mathrm{Var}(y\,|\,x)$ bounded below and $Q=I$, so that $V_K=A'\Sigma A\geqslant C\|A\|^2$, and hence

$$|F|\leqslant C, \qquad \|FA\|^2=\mathrm{tr}(FA'AF)\leqslant\mathrm{tr}(CFV_KF)=C.$$

Then, for $1_n$ as in the proof of Theorem 1,

$$\begin{aligned} 1_n\|FA'\hat{Q}^{-1}\| &\leq 1_n\|FA'\| + 1_n\|FA'(\hat{Q}-I)\hat{Q}^{-1}\| \\ &\leq C + O_p(1)\|FA'\|\|\hat{Q}-I\| = O_p(1), \\ 1_n\|FA'\hat{Q}^{-1/2}\|^2 &\leq \|FA'\|^2 + 1_n \operatorname{tr}(FA'(\hat{Q}-I)\hat{Q}^{-1}AF) \\ &\leq C + C\|FA'\|1_n\|FA'\hat{Q}^{-1}\|\|\hat{Q}-I\| = O_p(1). \end{aligned} \tag{A.6}$$

Now, let $\bar{g}_K(x) = p^K(x)'\beta_K$ for $\beta_K$ from Assumption 3 and $\varepsilon = Y - G$. Then

$$\begin{aligned} 1_n\sqrt{n}V_K^{-1/2}(\hat{\theta}-\theta_0) &= 1_n\sqrt{n}F[a(\hat{g}) - a(g_0)] \\ &= 1_n\{\sqrt{n}F[a(\hat{g}) - a(g_0) - D(\hat{g}) + D(g_0)] \\ &\quad + FA'P'\varepsilon/\sqrt{n} + \sqrt{n}FA'(\hat{Q}^{-1} - I)P'\varepsilon/n \\ &\quad + \sqrt{n}FA'\hat{Q}^{-1}P'(G - P\beta_K)/n + \sqrt{n}F[D(\bar{g}_K) - D(g_0)]\}. \end{aligned} \tag{A.7}$$

By Assumptions 3 and 5 $1_n|\sqrt{n}F[D(\bar{g}_K) - D(g_0)]| \leq \sqrt{n}C|D(\hat{g}_K - g_0)| \leq C\sqrt{n}|\bar{g}_K - g_0|_d \leq C\sqrt{n}K^{-\alpha} \to 0$. Also, by the Cauchy–Schwartz inequality and Eq. (A.6), $1_n|\sqrt{n}FA'\hat{Q}^{-1}P'(g - P\beta_K)/n| \leq 1_n\|FA'\hat{Q}^{-1}P'/\sqrt{n}\|\|g - P\beta_K\| \leq 1_n\|FA'\hat{Q}^{-1/2}\|\sqrt{n}$ $\max_{i\leq n}|g_0(x_i) - \bar{g}_K(x_i)| \leq 1_n\|FA'\hat{Q}^{-1/2}\|\sqrt{n}|g_0 - \bar{g}_K|_0 = O_p(\sqrt{n}K^{-\alpha}) = o_p(1)$. Next, let $X = (x_1, \ldots, x_n)$ and $p_i = p^K(x_i)$. Then by $E[\varepsilon \mid X_n] = 0$, $E[1_n\|\sqrt{n}FA'(\hat{Q}^{-1/2} - I)P'\varepsilon/n\|^2|X_n] = 1_n \operatorname{tr}(FA'(\hat{Q}^{-1} - I)[\sum_i p_i p_i'\sigma^2(x_i)/n] \times (\hat{Q}^{-1} - I)AF') \leq C1_n \operatorname{tr}(FA'(\hat{Q}^{-1} - I)\hat{Q}(\hat{Q}^{-1} - I)AF) \leq C1_n \operatorname{tr}(FA'(\hat{Q} - I)\hat{Q}^{-1}(\hat{Q} - I)AF') \leq C\|FA\|^2\|\hat{Q} - I\|^2 \xrightarrow{p} 0$. Therefore, since this conditional expectation is $o_p(1)$, it follows that $1_n\|\sqrt{n}FA'(\hat{Q}^{-1/2} - I)P'\varepsilon/n\|^2 \xrightarrow{p} 0$.

Next, let $Z_{in} = FA'p_i\varepsilon_i/\sqrt{n}$, so that $\sum_i Z_{in} = FA'P'\varepsilon/\sqrt{n}$. Note that for each $n$, $Z_{in}$ $(i = 1, \ldots, n)$ is i.i.d. Also, $E[Z_{in}] = 0$, $\sum_i E[Z_{in}^2] = 1$, and

$$\begin{aligned} nE[1(|Z_{in}| > \varepsilon)Z_{in}^2] &= n\varepsilon^2 E[1(|Z_{in}/\varepsilon| > 1)(Z_{in}/\varepsilon)^2] \leq n\varepsilon^2 E[(Z_{in}/\varepsilon)^4] \\ &\leq n\varepsilon^2\|FA'\|^4\zeta_0(K)^2 E[\|p_i\|^2 E[\varepsilon_i^4 \mid x_i]]/n^2\varepsilon^4 \\ &\leq C\zeta_0(K)^2 K/n \to 0. \end{aligned} \tag{A.8}$$

Then by the Lindbergh–Feller central limit theorem, $\sum_i Z_{in} \xrightarrow{d} N(0, 1)$, so $1_n FAQ^{-1}P'\varepsilon/\sqrt{n} \xrightarrow{d} N(0, I)$ by $1_n \xrightarrow{p} 1$. Finally, by Theorem 1, Assumption 5, and Eq. (A.5), $|\sqrt{n}F[a(\hat{g}) - a(g_0) - D(\hat{g}) + D(g_0)]| \leq L\sqrt{n}(|\hat{g} - g_0|_d)^2 = O_p(L\sqrt{n}[\zeta_d(K)(\sqrt{K}/\sqrt{n} + K^{-\alpha})]^2) \xrightarrow{d} 0$. 0 Then by Eq. (A.7) and the triangle

inequality, $1_n\sqrt{n}V_K^{-1/2}(\hat{\theta}-\theta_0)\xrightarrow{d}N(0,1)$, so that the first conclusion follows by $(1-1_n)\sqrt{n}V_K^{-1/2}(\hat{\theta}-\theta_0)\xrightarrow{p}0$.

Next, consider case (a) of Assumption 5, where $a(g)$ is linear in $g$. Then $a(p^{K\prime}\beta)=A'\beta$, so $\hat{A}=A$. In case (b) of Assumption 5, $\zeta_d(K)[\sqrt{K}/\sqrt{n}+K^{-\alpha}]=[\zeta_d(K)^2K/n]^{1/2}(1+K^{-1/2}\sqrt{n}K^{-\alpha})\to 0$, so that by Theorem 1, $|\hat{g}-g_0|_d\xrightarrow{p}0$. Therefore, for the $\varepsilon$ from Assumption 5 and $\bar{1}_n$ equal to the indicator function for $|\hat{g}-g_0|_d<\varepsilon/2$ and $1_n=1$, $\text{Prob}(\bar{1}_n=1)\to 1$. Also, let $\hat{J}=(D(p_{1k};\hat{g}),\ldots,D(p_{KK};\hat{g}))'$. By Assumption 5, for any $\beta$ such that $|p^{K\prime}\beta-\hat{g}|<\varepsilon/2$, implying $|p^{K\prime}\beta-g_0|<\varepsilon$,

$$\begin{aligned}
&\bar{1}_n|a(p^{K\prime}\beta)-a(\hat{g})-\hat{J}'(\beta-\hat{\beta})|/\|\beta-\hat{\beta}\| \\
&\quad=\bar{1}_n|a(p^{K\prime}\beta)-a(\hat{g})-D(p^{K\prime}\beta;\hat{g})+D(\hat{g};\hat{g})|/\|\beta-\hat{\beta}\| \\
&\quad\leqslant\bar{1}_nC(|p^{K\prime}\beta-\hat{g}|_d)^2/\|\beta-\hat{\beta}\|\leqslant\bar{1}_nC\zeta_d(K)^2\|\beta-\hat{\beta}\|\to 0
\end{aligned}$$

as $\beta\to\hat{\beta}$, so that $\hat{A}$ exists and equals $\hat{J}$ when $\bar{1}_n=1$. Therefore, $\hat{A}$ exists with probability approaching one. Furthermore, by linearity of $D(g;\tilde{g})$ in $g$, Assumption 5, and Theorem 1,

$$\begin{aligned}
\bar{1}_n\|\hat{A}-A\|^2 &= \bar{1}_n(\hat{A}-A)'(\hat{A}-A) \\
&= \bar{1}_n|D((\hat{A}-A)'p^K;\hat{g})-D((\hat{A}-A)'p^K;g_0)| \\
&\leqslant C\bar{1}_n|(\hat{A}-A)'p^K|_d|\hat{g}-g_0|_d \\
&\leqslant \bar{1}_nC\|\hat{A}-A\|\zeta_d(K)|\hat{g}-g_0|_d. \qquad \text{(A.9)}
\end{aligned}$$

Eq. (A.5) and dividing through by $\|\hat{A}-A\|$ gives $\bar{1}_n\|\hat{A}-A\|\leqslant C\zeta_d(K)|\hat{g}-g_0|_d=O_p(\zeta_d(K)^2[\sqrt{K}/\sqrt{n}+K^{-\alpha}])\xrightarrow{p}0$. Therefore, $\bar{1}_n\|F\hat{A}\|\leqslant\bar{1}_n\|F\|\|\hat{A}-A\|+\|FA\|=O_p(1)$, and similarly $\bar{1}_n\|F\hat{A}\hat{Q}^{-1}\|=O_p(1)$.

Next, let $\hat{h}=\bar{1}_n\hat{Q}^{-1}\hat{A}F$ and $h=\bar{1}_nAF$. Then $\|\hat{h}\|=O_p(1)$ and

$$\begin{aligned}
\|\hat{h}-h\| &\leqslant\bar{1}_n\|F\hat{A}'\hat{Q}^{-1}(I-\hat{Q})\|+\bar{1}_n\|F(\hat{A}-A)'\| \\
&\leqslant\bar{1}_n\|F\hat{A}'\hat{Q}^{-1}\|\|I-\hat{Q}\|+\bar{1}_n|F|\|\hat{A}-A\|\xrightarrow{p}0.
\end{aligned}$$

Since $\Sigma\leqslant CI$, the largest eigenvalue of $\Sigma$ is bounded above, and hence by $h'\Sigma h=\bar{1}_n$

$$\begin{aligned}
\bar{1}_n|\hat{h}'\Sigma\hat{h}-1| &= |\hat{h}'\Sigma\hat{h}-h'\Sigma h|\leqslant(\hat{h}-h)'\Sigma(\hat{h}-h)+|2(\hat{h}-h)'\Sigma h| \\
&\leqslant C\|\hat{h}-h\|^2+2[(\hat{h}-h)'\Sigma(\hat{h}-h)]^{1/2}[h'\Sigma h]^{1/2} \\
&\leqslant o_p(1)+C\|\hat{h}-h\|\xrightarrow{p}0. \qquad \text{(A.10)}
\end{aligned}$$

Also, let $\tilde{\Sigma} = \sum_i p_i p_i' \varepsilon_i^2/n$. It follows by $\mathrm{E}[\varepsilon_i^4 \mid x_i]$ bounded, and an argument like that for $\|\hat{Q} - I\| \xrightarrow{p} 0$ from Theorem 1, that $\|\tilde{\Sigma} - \Sigma\| \xrightarrow{p} 0$, implying

$$\bar{1}_n |\hat{h}'\tilde{\Sigma}\hat{h} - \hat{h}'\Sigma\hat{h}| = |\hat{h}'(\tilde{\Sigma} - \Sigma)\hat{h}| \leqslant \|\hat{h}\|^2 \|\tilde{\Sigma} - \Sigma\| = \mathrm{O_p}(1)\mathrm{o_p}(1) \xrightarrow{p} 0. \quad \text{(A.11)}$$

Now, let $\Delta_i = g_0(x_i) - \hat{g}(x_i)$. Note that $\zeta_0(K)[(K/n)^{1/2} + K^{-\alpha}] = [\zeta_0(K)^2 K/n]^{1/2}(1 + \sqrt{n}K^{-\alpha}/K^{1/2}) \to 0$, so by Theorem 1, $\max_{i \leqslant n} |\Delta_i| \leqslant |\hat{g} - g_0|_0 = \mathrm{O_p}(\mathrm{o}(1)) \xrightarrow{p} 0$. Also, let $\hat{S} = n^{-1} \sum_i p_i p_i' |\varepsilon_i|$ and $S = \mathrm{E}[p_i p_i' |\varepsilon_i|] = \mathrm{E}[p_i p_i' \mathrm{E}[|\varepsilon_i| x_i]] \leqslant CQ = CI$. By $\mathrm{E}[\varepsilon_i^2 \mid x_i]$ bounded and a similar argument to $\|\hat{Q} - I\| \xrightarrow{p} 0$, it follows that $\|\hat{S} - S\| \xrightarrow{p} 0$. Therefore,

$$\begin{aligned}
\bar{1}_n |F\hat{V}F - \hat{h}'\tilde{\Sigma}\hat{h}| &= |\hat{h}'(\hat{\Sigma} - \tilde{\Sigma})\hat{h}| = \left| n^{-1} \sum_i (\hat{h}' p_i)^2 (2\varepsilon_i \Delta_i + \Delta_i^2) \right| \\
&\leqslant \mathrm{o_p}(1)[\hat{h}'(\hat{S} + \hat{Q})\hat{h}] \\
&\leqslant \mathrm{o_p}(1)[\hat{h}'(\hat{S} + \hat{Q} - S - I)\hat{h}] + \mathrm{o_p}(1)[\hat{h}'(S + I)\hat{h}] \\
&\leqslant \mathrm{o_p}(1)\|\hat{h}\|^2(\|\hat{S} - S\| + \|\hat{Q} - I\|) + \mathrm{o_p}(1)\|\hat{h}\|^2 \xrightarrow{p} 0. \\
&\leqslant 2 \max_{i \leqslant n} |\Delta_i| n^{-1} \sum_i (\hat{h}' p_i)^2 |\varepsilon_i| + \max_{i \leqslant n} |\Delta_i^2| n^{-1} \sum_i (\hat{h}' p_i)^2
\end{aligned} \quad \text{(A.12)}$$

Then by the triangle inequality and Eqs. (A.10)–(A.12), it follows that $\bar{1}_n |F\hat{V}F - 1| \xrightarrow{p} 0$. Then by $\mathrm{Prob}(1_n = 1) \to 1$, $F^2\hat{V} \xrightarrow{p} 1$, implying

$$\sqrt{n}\hat{V}^{-1/2}(\hat{\theta} - \theta_0) = \sqrt{n}F(\hat{\theta} - \theta_0)/(F^2\hat{V})^{1/2} \xrightarrow{d} \mathrm{N}(0, 1),$$

giving the last conclusion.

Finally, to show the first conclusion, it suffices to show that $|V_K| \leqslant C\zeta_d(K)^2$, because $\hat{\theta} = \theta_0 + (V_K^{1/2}/\sqrt{n})\sqrt{n}F(\hat{\theta} - \theta_0) = \theta_0 + \mathrm{O_p}(V_K^{1/2}/\sqrt{n})$. Note that for any $\beta$, the Cauchy–Schwartz inequality implies $|p^{K\prime}\beta|_d \leqslant \zeta_d(K)\|\beta\|$, so that $\|A\|^2 = |D(p^{K\prime}A)| \leqslant C|p^{K\prime}A|_d \leqslant C\zeta_d(K)\|A\|$. Dividing by $\|A\|$ then gives $\|A\| \leqslant C\zeta_d(K)$, and hence $|V_K| \leqslant C\|A\|^2 \leqslant C\zeta_d(K)^2$. QED.

*Proof of Theorem 3.* By the Cramer–Wold device and by symmetry of $\hat{V}$ and $V$, it suffices to prove that $c'\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0, c'Vc)$ and $c'\hat{V}c \xrightarrow{p} c'Vc$ for any conformable constant vector $c$ with $\|c\| = 1$. Furthermore, for the functional $c'a(g)$, Assumption 7 is satisfied with $v(x)$ replaced by $c'v(x)$. Therefore, it suffices to prove the result for scalar $v(x)$. Let $v_K = Ap^K(x) = \mathrm{E}[v(x_i)p^K(x_i)']Q^{-1}p^K(x)$ be the mean square projection of $v = v(x)$ on the approximating functions, where the $x$ argument is dropped for notational simplicity. Then $V_K = \mathrm{E}[v_K^2 \sigma^2(x_i)]$ by

Assumption 7, $E[(v - v_K)^2] \leq E[\{v - p^K(x_i)'\beta_K\}^2] \to 0$. Assumptions 1 and the Cauchy–Schwartz inequality then give

$$|V_K - V| \leq E[|v_K^2 - v^2|] \leq E[(v_K - v)^2] + 2E[|v||v_K - v|]$$
$$\leq o(1) + 2(E[v^2])^{1/2}(E[(v - v_K)^2])^{1/2} \to 0.$$

Thus, $V_K \xrightarrow{p} V$. By $\sigma^2(x)$ bounded away from zero and $v(x)$ nonzero, $V > 0$. The proof now follows exactly as in the proof of Theorem 2, except that $F = V_K^{-1/2}$ is bounded by $V_K^{-1/2} \to 1/\sqrt{V}$. Therefore, by the conclusion of Theorem 2, $\sqrt{n}(\hat{\theta} - \theta_0) = V_K^{1/2}\sqrt{n}V_K^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$. Also, as in the proof of Theorem 2, $F\hat{V}^{1/2} \xrightarrow{p} 1$, so that $V^{-1/2}\hat{V}^{1/2} \xrightarrow{p} 1$, and squaring gives $\hat{V} \xrightarrow{p} V$. QED.

*Proof of Theorem 4.* Let $P^K(x)$ be obtained by transforming the $x$'s so that their support is $[-1, 1]^r$ and replacing the individual powers in each term by the Jacobi polynomial of the same order that is orthonormal with respect to the uniform distribution on $[-1, 1]$, which comprises a nonsingular linear transformation of $p^K(x)$. By the density of $x$ bounded below, $CI \leq E[P^K(x_i)P^K(x_i)']$. Also, it follows as in Eqs. (3.13) and (3.14) of Andrews (1991) that $\max_{k \leq K, x \in \mathcal{X}} |P_{kK}(x)| \leq K^{1/2}$, so that $\|P^K(x)\| \leq CK$, and hence $\zeta_0(K)^2 K/n \leq CK^3/n \to 0$. Furthermore, it follows by Theorem 8 of Lorentz (1986) that Assumption 3 is satisfied with $d = 0$ and $\alpha = s/r$, so the conclusion follows by Theorem 1. QED.

*Proof of Theorem 5.* It follows as in the proof of Theorem 4 that Assumptions 2 and 3 are satisfied. To show Assumption 5, note that by Assumption 10 there is a sequence of uniformly bounded continuous functions $g_J(x)$ such that $|a(g_J)| > \varepsilon$ for all $J$ and $E[g_J(x_i)^2] \to 0$. Since power series can approximate a continuous function in the supremum norm on a compact set, there is $\beta_{KJ}$ such that $\Delta_{JK} = \sup_{x \in \mathcal{X}} |g_J(x) - \beta_{KJ}' p^K(x)| \to 0$ as $K \to \infty$ for any fixed $J$. Let $K(J)$ be a monotonic increasing sequence such that $\Delta_{JK} < 1/J$ for $K \geq K(J)$, and let $\beta_K = \beta_{KJ}$ for $K(J) \leq K < K(J+1)$, $\beta_K = 0$ for $K < K(1)$, and $g_K(x) = p^K(x)'\beta_K$. Then by the triangle inequality, $|a(g_K)| > \varepsilon$ for all $K > K(J_\varepsilon)$, $J_\varepsilon > 1/\varepsilon$, and $E[g_K(x_i)^2] \to 0$, so Assumption 5 is satisfied. Then, since $d = 0$, and from the proof of Theorem 4, $\zeta_0(K) \leq CK$, so for nonlinear $a(g)$ it follows that $\zeta_0(K)^4 K^2/n \leq CK^6/n \to 0$ and, since Assumption 3 is satisfied with $\alpha = s/r$, $\sqrt{n}K^{-\alpha} = \sqrt{n}K^{-s/r} \to 0$. QED.

*Proof of Theorem 6.* Follows by Theorem 3 similarly to the proof of Theorem 5, with Assumption 7 replacing Assumption 10. QED.

*Proof of Theorem 7.* It follows by Lemma A.16 of Newey (1995), that for $P^K(x)$ equal to the products of normalized $B$-splines (e.g. see Powell, 1981) multiplied by the square root of the number of knots, $\|P^K(x)\| \leq CK^{1/2} = \zeta_0(K)$, and hence

$\zeta_0(K)^2K/n \leqslant CK^2/n \to 0$. Furthermore, it follows by the argument of Burman and Chen (1989, p. 1587), that the rest of Assumption 2 is satisfied. Also, Assumption 3 with $d=0$ and $\alpha=s/r$ follows by Theorem 12.8 of Schumaker (1981), so the conclusion follows from Theorem 1. QED.

*Proof of Theorem 8.* Follows as in the proof of Theorem 5.

*Proof of Theorem 9.* Follows as in the proof of Theorem 6.

## References

Andrews, D.W.K., 1991. Asymptotic normality of series estimators for nonparametric and semiparametric models. Econometrica 59, 307–345.

Andrews, D.W.K., Whang, Y.J., 1990. Additive interactive regression models: Circumvention of the curse of dimensionality. Econometric Theory 6, 466–479.

Burman, P., Chen, K.W., 1989. Nonparametric estimation of a regression function. Annals of Statistics 17, 1567–1596.

Chen, H., 1988. Convergence rates for parametric components in a partly linear model. Annals of Statistics 16, 136–146.

Cox, D.D., 1988. Approximation of least squares regression on nested subspaces. Annals of Statistics 16, 713–732.

Deaton, A., 1988. Rice prices and income distribution in Thailand: A nonparametric analysis. Economic Journal 99 supplement, 1–37.

Eastwood, B.J., Gallant, A.R., 1991. Adaptive rules for seminonparametric estimation that achieve asymptotic normality. Econometric Theory 7, 307–340.

Gallant, A.R., Souza, G., 1991. On the asymptotic normality of Fourier flexible functional form estimates. Journal of Econometrics 50, 329–353.

Hausman, J.A., Newey, W.K., 1995. Nonparametric estimation of exact consumer surplus and deadweight loss. Econometrica, 63, 1445–1476.

Lorentz, G.G., 1986. Approximation of Functions. Chelsea, New York.

Newey, W.K., 1988. Adaptive estimation of regression models via moment restrictions. Journal of Econometrics 38, 301–339.

Newey, W.K., 1994a. The asymptotic variance of semiparametric estimators. Econometrica 62, 1349–1382.

Newey, W.K., 1994b. Series estimation of regression functionals. Econometric Theory 10, 1–28.

Newey, W.K., 1995. Convergence rates for series estimators. In: Maddalla, G.S., Phillips, P.C.B., Srinavasan, T.N. (Eds.), Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C.R. Rao. Blackwell, Cambridge, USA, pp. 254–275.

Powell, M.J.D., 1981. Approximation Theory and Methods. Cambridge University Press, Cambridge, UK.

Schumaker, L.L., 1981. Spline Functions: Basic Theory. Wiley, New York.

Stoker, T.M., 1986. Consistent estimation of scaled coefficients. Econometrica 54, 1461–1481.

Stone, C.J., 1982. Optimal global rates of convergence for nonparametric regression. Annals of Statistics 10, 1040–1053.

Stone, C.J., 1985. Additive regression and other nonparametric models. Annals of Statistics 13, 689–705.

Vander Vaart, A., 1991. "On Differentiable Functionals", Annals of Statistics 13, 689–705.

White, H., 1980. Using least squares to approximate unknown regression functions. International Economic Review 21, 149–170.