INTERNATIONAL ECONOMIC REVIEW
Vol. 41, No. 4, November 2000

EFFICIENT ESTIMATION OF ADDITIVE PARTIALLY LINEAR MODELS*

BY QI LI†[1]

Texas A&M University, U.S.A., and University of Guelph, Canada

I consider the problem of estimating an additive partially linear model using general series estimation methods with polynomial and splines as two leading cases. I show that the finite-dimensional parameter is identified under weak conditions. I establish the root-n-normality result for the finite-dimensional parameter in the linear part of the model and show that it is asymptotically more efficient than a semiparametric estimator that ignores the additive structure. When the error is conditional homoskedastic, my finite-dimensional parameter estimator reaches the semiparametric efficiency bound. Efficient estimation when the error is conditional heteroskedastic is also discussed.

1. INTRODUCTION

Series estimation methods are convenient for imposing certain type of restrictions, such as additive separability (e.g., Stone, 1985; Andrews and Whang 1990) or shape-preserving estimations (e.g., Dechevsky and Penez, 1997). Also, it is computationally convenient because the data are summarized by a relatively few estimated coefficients. For large sample properties of series estimators, see Stone (1985), Cox (1988), Eubank (1988), Andrews and Whang (1990), Eastwood and Gallant (1991), Gallant and Souza (1991), Eubank and Jayasuriya (1993), and Newey (1988, 1994a, 1994b,1995). For using series methods to estimate semiparametric regression models, see Ai and McFadden (1997), Andrews (1991), and Donald and Newey (1994), among others. Newey (1997) established the $\sqrt{n}$-normality result for nonlinear functionals of series estimators; Chen and Shen (1998) consider a more general class of sieve extremum estimates and established the $\sqrt{n}$-normality result for smoothly functional (possibly nonlinearly functional) sieve estimators.

In this article I will consider the problem of estimating an additive partially linear model using series estimation methods. Recently, semiparametric estimation of additive models and additive partially linear models has attracted much attention among econometricians and statisticians (see Linton and Nielsen, 1995; Newey, 1994c;

* Manuscript received April 1998; revised December 1998.

† E-mail: *qi@econ.tamu.edu*.

[1] I would like to thank two referees and a coeditor for their comments that greatly improved the article. I also would like to thank Chunrong Ai, Oliver Linton, and Jianming Yao for very helpful comments. This research was supported by the Social Sciences and Humanities Research Council of Canada, the Natural Sciences and Engineering Research Council of Canada, Ontario Premier's Research Excellence Awards, and Bush Program of Economics and Public Policy.

Tojstheim and Auestad, 1994, to mention only a few). An additive model with two regressors has the following form (e.g., Linton and Nielsen, 1995):

$$Y_i = \gamma_0 + g_1(Z_{1i}) + g_2(Z_{2i}) + u_i \qquad i = 1, \ldots, n \tag{1}$$

where $\{Y_i, Z_{1i}, Z_{2i}\}_{i=1}^n$ are i.i.d. observations, $E(u_i|Z_{1i}, Z_{2i}) = 0$, and $g_1(\cdot)$ and $g_2(\cdot)$ are unknown univariate functions.

Stone (1985, 1986) has shown that the additive components of $g_l(\cdot)$ $(l = 1, 2)$ in (1) can be consistently estimated at the same rate as a fully nonparametric regression model with only one regressor. Hence additive regression models circumvent the "curse of dimensionality" problem that affects the estimation of fully nonparametric regression models.

Linton and Nielsen (1995) proposed to estimate $g_l(z_l)$ $(l = 1, 2)$ by marginally integrating (averaging) a local linear estimator of $g(z_1, z_2) = E(y|z_1, z_2)$, and they showed that their proposed estimator of $g_l(z_l)$ has an asymptotic normal distribution and achieves the one-dimensional optimal convergence rate. Chen et al. (1996) extended the result in Linton and Nielsen (1995) to additive regression models with more than two regressors. Fan et al. (1998) studied more general additive models including additive partially linear models. A typical additive partially linear model has the following form:

$$Y_i = X_i'\gamma + g_1(Z_{1i}) + g_2(Z_{2i}) + \cdots + g_L(Z_{Li}) + u_i \tag{2}$$

The additive partially linear model is particularly convenient when X_i is a vector of discrete variables that takes a large number of different values, i.e., X_i consists of categorical variables. Another interesting aspect of an additive partially linear model is that it allows X_i to be a deterministic (nonadditive) function of $(Z_{1i}, \ldots, Z_{Li})$, thus allowing high-dimension variables to enter the model parametrically; see Section 2 for more discussion on this. Fan and Li (1996) proposed some alternative estimation methods of estimating additive partially linear models, and they showed that, via Monte Carlo simulations, their proposed estimators compare favorably with the estimators proposed by Fan et al. (1998). Both Fan et al. (1998), and Fan and Li (1996) used kernel estimation methods.

In this article I propose to estimate an additive partially linear model using series estimation method. There are at least four advantages to using series method to estimate an additive partially linear model compared with the kernel estimation method.

First, the kernel estimation method is to initially estimate a nonparametric model with high dimension (ignoring the additive structure) and then to use the method of marginal average to obtain an estimator of a function with lower dimension (using the additive structure). In applications, this may cause some *finite sample* efficiency loss due to the fact that the additive structure is not used in the initial estimation stage. In contrast, the series estimation method can easily impose additive structure throughout the estimation procedure.

Second, the kernel marginal integration method can be computationally costly; the computation time of estimating an additive partially linear model is about n (n is the sample size) times the computation time of estimating a nonadditive partially

linear model. Hence, for large samples sizes, the computation burden of estimating an additive partially linear model using the kernel method could be very costly. Using the series estimation method, the computation time of estimating an additive partially linear model is generally less than that of estimating a nonadditive partially linear model.

Third, the kernel estimation methods are two-step methods that either first estimate the additive functions $g_l(z_l)$ $(l = 1, \ldots, L))$ and then estimate the parametric parameter γ (e.g., Fan et al., 1998) or first estimate γ and then to estimate $g_l(z_l)$ (e.g., Fan and Li, 1996). The two-step methods do not lead to efficient estimation of the finite-dimensional parameter γ. Using the series estimation method, one can estimate $\sum_{l=1}^{L} g(z_l)$ and γ simultaneously; I show that my proposed estimator for γ is semiparametric efficient when the error is conditional homoskedastic.[2]

Fourth, series estimators have well-defined meanings even when the model is misspecified; i.e., when the true function of $E(Y|X, Z)$ is not an additive partially linear regression function, the series estimator will converge to an additive partially linear function that best approximates the unknown function $E(Y|X, Z)$ in the mean square error sense, while the two-step kernel estimators do not have this property in misspecified models.[3]

There are also some drawbacks to using the series estimation method compared with the kernel method. For example, using general series estimation methods, it is difficult to establish the asymptotic normality result for the nonparametric additive component estimators under primitive conditions. Therefore, the series estimation method should be viewed as a complement to the kernel method for estimating an additive partially linear model.[4]

A related article is that of Chen and Shen (1998), who tried to estimate general time-series regression models by the sieve method. One regression model considered in Chen and Shen (1998) is a special type of additive partially linear model. Using the notation of Equation (2), Chen and Shen (1998) considered the case that $E(X_i|Z_i)$ is also additive separable in $Z_{i1}, \ldots, Z_{Li}$, say, $E(X_i|Z_i) = \sum_{l=1}^{L} t_l(Z_{li})$ for some (possibly unknown) smooth functionals $t_l(\cdot)$ $(l = 1, \ldots, L)$. In general, $E(X_i|Z_i)$ may not be additive separable in $Z_{1i}, \ldots, Z_{Li}$. It turns out that when $E(X_i|Z_i)$ is not additive separable in $Z_{1i}, \ldots, Z_{Li}$; I have the following interesting results: (1) the parametric coefficient γ is identified even when $E(X_i|Z_i) = X_i$, and (2) my proposed estimator of γ is asymptotically more efficient than a semiparametric estimator of γ that ignores the additive structure of $g(z) = \sum_{l=1}^{L} g_l(z_l)$.

[2] An estimator is said to be semiparametric efficient if the inverse of the asymptotic variance of the estimator equals the semiparametric efficiency bound. The general framework of efficiency bounds is provided by Begun et al. (1983). For applications of the general framework to different econometrics models, see Chamberlain (1986), Cosslett (1987), Hansen et al. (1988), Newey (1990), and Ai (1994), to mention only a few.

[3] For additive specifications without a partially linear component, the kernel marginal integration method can best approximate the unknown regression function (see Linton, 1997; Nielson and Linton, 1998).

[4] A referee pointed out to me that by using the result of Newey (1997), one should be able to establish the asymptotic normality result for the nonparametric (additive) component functions. This subject is left for future research.

2. THE MODEL AND THE MAIN RESULT

Consider the following additive partially linear regression model:

$$Y_i = X_i'\gamma + g_1(Z_{1i}) + g_2(Z_{2i}) + \cdots + g_L(Z_{Li}) + u_i \tag{3}$$

where the prime denotes transpose, X_i is an $r \times 1$ vector of random variables that does not contain a constant term,[5] $\gamma = (\gamma_1, \ldots, \gamma_r)'$ is an $r \times 1$ vector of unknown parameter, and Z_{li} is of dimension q_l ($q_l \geq 1$, $l = 1, \ldots, L$). Denote by Z_i the nonoverlapping variables obtained from $(Z_{1i}, \ldots, Z_{Li})$. Z_i is of dimension q with $L \leq q \leq \sum_{l=1}^{L} q_l$. $E(u_i|X_i, Z_i) = 0$, and $g_1(\cdot), \ldots, g_L(\cdot)$ are unknown smooth functions.

Obviously, the individual functions $g_l(\cdot)$ ($l = 1, \ldots, L$) are not identified without some identification conditions. In the literature of the kernel estimation method, a convenient identification condition is to impose $E[g_l(Z_{li})] = 0$ for all $l = 2, \ldots, L$. However, such conditions are less straightforward to impose using the series estimation method. In this article I will choose

$$g_l(z_l = 0) = g_l(0) = 0 \qquad (l = 2, .., L) \tag{4}$$

as my identification condition. This condition is convenient to impose for series estimators.

Next, I give the definition of the class of additive functions.

DEFINITION 1. We say that a function $\xi(z)$ belongs to an *additive class* of functions $\mathcal{G}$ ($\xi \in \mathcal{G}$) if (i) $\xi(z) = \sum_{l=1}^{L} \xi_l(z_l)$, $\xi_l(z_l)$ is continuous in its support $\mathcal{S}_l$, where $\mathcal{S}_l$ is a compact subset of R^{q_l} ($l = 1, \ldots, L$), (ii) $\sum_{l=1}^{L} E[\xi_l(Z_l)]^2 < \infty$, and (iii) $\xi_l(0) = 0$ for $l = 2, \ldots, L$.

When $\xi(z)$ is a vector-valued function, we say $\xi \in \mathcal{G}$ if each component of ξ belongs to $\mathcal{G}$.

In vector-matrix notation, I can write (3) as

$$\mathcal{Y} = \mathcal{X}\gamma + g_1 + g_2 + \cdots + g_L + U \equiv \mathcal{X}\gamma + g + U \tag{5}$$

where $\mathcal{Y}$ and U are both $n \times 1$ vectors with ith components given by Y_i and u_i, respectively; $\mathcal{X}$ is $n \times r$ with the ith row given by X_i'; g is $n \times 1$ with the ith component given by $g_i = g(Z_i) \equiv \sum_{l=1}^{L} g_l(Z_{li})$.

I use a linear combination of K_l functions, $p_l^{K_l}(z_l) = [p_{l1}^{K_l}(z_l), \ldots, p_{lK_l}^{K_l}(z_l)]'$, to approximate $g_l(z_l)$ ($l = 1, \ldots, L$). Hence I use a linear combination of $K = \sum_{l=1}^{L} K_l$ functions $[p_1^{K_1}(z_1)', \ldots, p_L^{K_L}(z_L)'] \equiv p^K(z)'$ to approximate $g(z) = g(z_1, \ldots, z_L) = \sum_{l=1}^{L} g_l(z_l)$. The approximation function $p^K(z)$ has the following properties: (1) $p^K(z) \in \mathcal{G}$, and (2) as K_l grows (for all $l = 1, \ldots, L$), there is a linear combination of $p^K(z)$ that can approximate any $g \in \mathcal{G}$ arbitrarily well in the mean square error sense.

[5] Note that this is not a restriction because the possibly nonzero intercept term will be incorporated into $g_1(z_1)$.

I introduce some notation. Define

$$\begin{aligned} p_l &= [p_l^{K_l}(Z_{l1}), \ldots, p_l^{K_l}(Z_{ln})]' \qquad (l = 1, \ldots, L) \\ P &= (p_1, \ldots, p_L) \end{aligned} \tag{6}$$

Note that p_l is of dimension $n \times K_l$ and P is of dimension $n \times K$.

Let $M = P(P'P)^- P'$, where $(\cdot)^-$ denotes any symmetric generalized inverse of $(\cdot)$. For an $n \times 1$ or an $n \times r$ matrix A, Define $\widetilde{A} = MA$. Then premultiplying (5) by M leads to

$$\widetilde{\mathcal{Y}} = \widetilde{\mathcal{X}}\gamma + \tilde{g} + \widetilde{U} \tag{7}$$

Subtracting (7) from (5) gives

$$\mathcal{Y} - \widetilde{\mathcal{Y}} = (\mathcal{X} - \widetilde{\mathcal{X}})\gamma + g - \tilde{g} + U - \widetilde{U} \tag{8}$$

I estimate γ by least squares regression of $\mathcal{Y} - \widetilde{\mathcal{Y}}$ on $\mathcal{X} - \widetilde{\mathcal{X}}$:

$$\hat{\gamma} = [(\mathcal{X} - \widetilde{\mathcal{X}})'(\mathcal{X} - \widetilde{\mathcal{X}})]^-(\mathcal{X} - \widetilde{\mathcal{X}})'(\mathcal{Y} - \widetilde{\mathcal{Y}}) \tag{9}$$

Then $g(z) = \sum_{l=1}^{L} g_l(z_l)$ is estimated by $\hat{g}(z) = p^K(z)'\hat{\beta}$, where $\hat{\beta}$ is given by

$$\hat{\beta} = (P'P)^- P'(\mathcal{Y} - \mathcal{X}\hat{\gamma}) \tag{10}$$

Under the conditions given below, both $(P'P)$ and $(\mathcal{X} - \widetilde{\mathcal{X}})'(\mathcal{X} - \widetilde{\mathcal{X}})$ are asymptotically nonsingular. Hence, all the generalized inverses are in fact inverses when we take the limit of $\min\{K_1, \ldots, K_n\} \to \infty$ (as $n \to \infty$). Note that when both $(\mathcal{X} - \widetilde{\mathcal{X}})'(\mathcal{X} - \widetilde{\mathcal{X}})$ and $(P'P)$ are nonsingular, $\hat{\gamma}$ and $\hat{\beta}$ given in (9) and (10) are numerically identical to the least squares estimator of regressing $\mathcal{Y}$ on $(\mathcal{X}, P)$.[6]

For any scalar or vector function $\mathcal{W}(z)$, I use the notation of $E_A[\mathcal{W}(z)]$ to denote the projection of $\mathcal{W}(\cdot)$ onto the additive functional space $\mathcal{G}$ (under the L_2-norm). That is, $E_A[\mathcal{W}(z)]$ is an element that belongs to $\mathcal{G}$ (has an additive structure), and it is the closest function to $\mathcal{W}(z)$ among all the functions in $\mathcal{G}$. More specifically, I have

$$\begin{aligned} &E\left(\left\{\mathcal{W}(Z_i) - E_A\left[\mathcal{W}(Z_i)\right]\right\}\left\{\mathcal{W}(Z_i) - E_A\left[\mathcal{W}(Z_i)\right]\right\}'\right) \\ &\quad = \inf_{\xi = \sum_l \xi_l \in \mathcal{G}} E\left\{\left[\mathcal{W}(Z_i) - \sum_{l=1}^{L} \xi_l(Z_{li})\right]\left[\mathcal{W}(Z_i) - \sum_{l=1}^{L} \xi_l(Z_{li})\right]'\right\} \end{aligned} \tag{11}$$

where the infimum of (11) is in the sense that

$$\begin{aligned} &E\left(\left\{\mathcal{W}(Z_i) - E_A\left[\mathcal{W}(Z_i)\right]\right\}\left\{\mathcal{W}(Z_i) - E_A\left[\mathcal{W}(Z_i)\right]\right\}'\right) \\ &\quad \le E\left\{\left[\mathcal{W}(Z_i) - \sum_{l=1}^{L} \xi_l(Z_{li})\right]\left[\mathcal{W}(Z_i) - \sum_{l=1}^{L} \xi_l(Z_{li})\right]'\right\} \end{aligned} \tag{12}$$

[6] In finite sample applications, it is possible that $(\mathcal{X} - \widetilde{\mathcal{X}})'(\mathcal{X} - \widetilde{\mathcal{X}})$ and/or $(P'P)$ are singular. However, one can drop the redundant regressors to make these matrices nonsingular.

for all $\xi(z) = \sum_{l=1}^{L} \xi_l(z) \in \mathcal{G}$, where for square matrices A and B, $A \leq B$ means that $A - B$ is negative semidefinite.

Define $\theta(z) = E(X|Z = z)$, and I will use $h(z)$ to denote the projection of $\theta(z)$ onto $\mathcal{G}$, i.e., $h(z) = E_A[\theta(z)]$. By the definition of $E_A(\cdot)$, I know that $h(\cdot)$ is an additive function, i.e., $h(z) = \sum_{l=1}^{L} h_l(z_l) \in \mathcal{G}$, and $h(\cdot)$ is the solution of the following minimization problem:

$$
\begin{aligned}
&E\left\{\left[\theta(Z_i) - h(Z_i)\right]\left[\theta(Z_i) - h(Z_i)\right]'\right\} \\
&\quad = \inf_{\xi = \sum_l \xi_l \in \mathcal{G}} E\left\{\left[\theta(Z_i) - \sum_{l=1}^{L} \xi_l(Z_{li})\right]\left[\theta(Z_i) - \sum_{l=1}^{L} \xi_l(Z_{li})\right]'\right\}
\end{aligned} \tag{13}
$$

Under the L_2-norm, $\mathcal{G}$ is an infinite-dimensional Hilbert space. Therefore, the space $\mathcal{G}$ is not compact, and it is well known that minimization (optimization) over a noncompact set may not have a solution. In the Appendix I prove the existence of the function $h(z) = E_A[\theta(z)] \in \mathcal{G}$ that satisfies (13); i.e., $E_A[\mathcal{W}(z)]$ and $E_A[\theta(z)]$ in (11) and (13) are well-defined functions [the infimum bounds in (11) or (13) are attainable by some function in $\mathcal{G}$].

Note that $h(\cdot)$ is of dimension $r \times 1$. I will use $h_{(s)}(\cdot)$ to denote its sth component ($s = 1, \ldots, r$), i.e., $h(z) = [h_{(1)}(z), h_{(2)}(z), \ldots, h_{(r)}(z)]'$.

From (13) and using $X_i = \theta(Z_i) + v_i$ and $E(v_i|Z_i) = 0$, I immediately get the following equivalent expression of (13) in terms of X_i:

$$
\begin{aligned}
&\inf_{\xi \in \mathcal{G}} E\{[X_i - \xi(Z_i)][X_i - \xi(Z_i)]'\} = \inf_{\xi \in \mathcal{G}} E\{[\theta(Z_i) - \xi(Z_i)][\theta(Z_i) - \xi(Z_i)]'\} \\
&\qquad + E(v_i v_i') = E\{[\theta(Z_i) - h(Z_i)][\theta(Z_i) - h(Z_i)]'\} \\
&\qquad + E(v_i v_i') = E\{[X_i - h(Z_i)][X_i - h(Z_i)]'\}
\end{aligned} \tag{14}
$$

i.e., $h(Z_i)$ is also the projection of X_i onto $\mathcal{G}$ [$h(Z) = E_A(X)$] because $v_i \perp \mathcal{G}$.

The following assumptions are needed to establish the asymptotic distribution of $\hat{\gamma}$ as well as the convergence rates of $\hat{g}(z) = p^K(z)'\hat{\beta}$ to $g(z)$.

ASSUMPTION 1. (i) $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ are independent and identically distributed as (Y, X, Z); the support of (X, Z) is a compact subset of R^{r+q}; (ii) define $\theta(z) = E(X|Z = z)$. Both $\theta(z)$ and $\mathrm{var}(Y|X = x, Z = z)$ are bounded functions on the support of (X, Z).

ASSUMPTION 2. (i) For every K there is a nonsingular matrix B such that for $P^K(z) = Bp^K(z)$; the smallest eigenvalue of $E[P^K(Z_i)P^K(Z_i)']$ is bounded away from zero uniformly in K; (ii) there is a sequence of constants $\zeta_0(K)$ satisfying $\sup_{z \in \mathcal{S}} \|P^K(z)\| \leq \zeta_0(K)$ and $K = K_n$ such that $(\zeta_0(K))^2 K/n \to 0$ as $n \to \infty$, where $\mathcal{S}$ is the support of Z.

ASSUMPTION 3. (i) For $f = g$ or $f = h_{(s)}$ ($s = 1, \ldots, r$), there exist some $\delta_l\,(> 0)$ ($l = 1, \ldots, L$), $\beta_f = \beta_{fK} = (\beta_{fK_1}', \ldots, \beta_{fK_L}')'$, $\sup_{z \in \mathcal{X}} |f(z) - P^K(z)'\beta_f| = O(\sum_{l=1}^{L} K_l^{-\delta_l})$ as $\min\{K_1, \ldots, K_L\} \to \infty$; (ii) $\sqrt{n}(\sum_{l=1}^{L} K_l^{-\delta_l}) \to 0$ as $n \to \infty$.

Assumption 1 is quite standard in the literature of estimating additive models. Assumption 2 ensures that $(P'P)$ is asymptotically nonsingular. Note that both g and $h_{(s)}$ are additive functions, i.e., $g(z) = \sum_{l=1}^{L} g_l(z_l)$ and $h_{(s)}(z) = \sum_{l=1}^{L} h_{(s),l}(z_l)$ $(s = 1, \ldots, r)$. Hence Assumption 3 is implied by the following: for all $l = 1, \ldots, L$ and for $f_l = g_l$ or $f_l = h_{(s),l}$ $(s = 1, \ldots, r)$, there exists some $\delta_l > 0$, $\beta_{fl} = \beta_{fl,K_l}$, such that $\sup_{z_l \in \mathscr{S}_l} |f_l(z_l) - p_l^{K_l}(z_l)'\beta_{fl}| = O(K_l^{-\delta_l})$ as $K_l \to \infty$, where $\mathscr{S}_l$ is the support of z_l. While Assumptions 2 and 3 are not primitive conditions, it is known that many series functions satisfy these conditions. Newey (1997) gives primitive conditions for power series and splines such that Assumptions 2 and 3 hold (see Assumptions 4 and 5 below).

The following theorem gives the asymptotic distribution of $\hat{\gamma}$.

THEOREM 1. *Define* $\epsilon_i = X_i - h(Z_i) \equiv X_i - E_A(X_i)$, *where* $h(\cdot)$ *is defined by* (13), *and assume that* $\Phi \overset{def}{=} E[\epsilon_i \epsilon_i']$ *is positive definite, then under Assumptions* 1 *to* 3, *we have*

(*i*) $\sqrt{n}(\hat{\gamma} - \gamma) \to N(0, \Sigma)$ *in distribution, where* $\Sigma = \Phi^{-1}\Omega\Phi^{-1}$, $\Omega = E[\sigma_u^2(X_i, Z_i)\epsilon_i\epsilon_i']$ *and* $\sigma_u^2(x, z) = E(u_i^2 | X_i = x, Z_i = z)$.

(*ii*) *A consistent estimator of* Σ *is given by* $\widehat{\Sigma} = \widehat{\Phi}^{-1}\widehat{\Omega}\widehat{\Phi}^{-1}$, *where* $\widehat{\Phi} = n^{-1}\sum_i (X_i - \widetilde{X}_i)(X_i - \widetilde{X}_i)'$, $\widehat{\Omega} = n^{-1}\sum_i \hat{u}_i^2 (X_i - \widetilde{X}_i)(X_i - \widetilde{X}_i)'$, $\widetilde{X}_i'$ *is the ith row of* $\widetilde{\mathscr{X}}$ *and* $\hat{u}_i = y_i - X_i'\hat{\gamma} - \hat{g}(Z_i)$.

The proof of Theorem 1 is given in the Appendix.

The requirement that $\Phi = E(\epsilon_i \epsilon_i')$ is positive definite is an identification condition for γ. Below I show that this condition is weaker than the condition needed to identify γ when one ignores the additive structure of $\sum_{l=1}^{L} g_l(z_l)$, recall that $v_i = X_i - E(X_i|Z_i)$, and define $\eta_i = E(X_i|Z_i) - h(Z_i)$. Then $\epsilon_i = v_i + \eta_i$. Using $E(v_i|Z_i) = 0$, I get

$$\Phi = E[(v_i + \eta_i)(v_i + \eta_i)'] = E(v_i v_i') + E(\eta_i \eta_i') \tag{15}$$

Hence, either $E(v_i v_i')$ is positive definite or $E(\eta_i \eta_i')$ is positive definite will imply that Φ is positive definite. Note that $E(v_i v_i')$ being a positive definite matrix would be the required identification condition (for γ) if one ignores the additive structure of $\sum_{l=1}^{L} g_l(z_l)$ when estimating γ (e.g., Robinson, 1988; and Stock, 1989). Thus, by using the information that the model is additive partially linear, one can weaken the identification condition for γ as Φ being a positive definite matrix.

It is interesting to observe that when $E(X|Z = z) \neq h(z)$, γ is identified even if X is a deterministic function of Z [i.e., $E(X|Z) = X$]. By the definition of $h(z) = \sum_{l=1}^{L} h_l(z_l)$, I know that $h(z)$ is additive separable in $z_1, \ldots, z_l$. Hence, if $\theta(z) = E(X_i|Z_i = z)$ is not additive separable in $z_1, \ldots, z_L$, I will have $\theta(z) \neq h(z)$ [hence $\eta(z) \neq 0$], and consequently, γ is identified even when $E(X|Z) = X$ $(v \equiv 0)$. I will discuss more about this point below.

Next I show that when the error is conditional homoskedastic, i.e., $E(u_i^2|X_i, Z_i) = E(u_i^2) = \sigma_u^2$, my estimator $\hat{\gamma}$ is semiparametric efficient in the sense that the inverse of the asymptotic variance of $\sqrt{n}(\hat{\gamma} - \gamma)$ equals the semiparametric efficiency bound. From the result of Chamberlain (1992:579), I know that the semiparametric efficiency

bound for the inverse of the asymptotic variance of an estimator of γ is

$$J_0 = \inf_{\xi=\sum_l \xi_l \in \mathcal{G}} E\left\{\left[X_i - \sum_{l=1}^{L} \xi_l(z_l)\right]\left[\mathrm{var}(u_i|X_i, Z_i)\right]^{-1}\left[X_i - \sum_{l=1}^{L} \xi_l(z_l)\right]\right\} \tag{16}$$

Under the conditional homoskedastic error assumption, I have $\Sigma = \sigma_u^2\Phi^{-1}$, and (16) becomes

$$\begin{aligned} J_0 &= \frac{1}{\sigma_u^2} \inf_{\xi=\sum_l \xi_l \in \mathcal{G}} E\left\{\left[X_i - \sum_{l=1}^{L} \xi_l(z_l)\right]\left[X_i - \sum_{l=1}^{L} \xi_l(z_l)\right]'\right\} \\ &\equiv \frac{1}{\sigma_u^2}\left(E\left\{\left[X_i - \sum_{l=1}^{L} h(Z_{li})\right]\left[X_i - \sum_{l=1}^{L} h_l(Z_{li})\right]'\right\}'\right) = \frac{1}{\sigma_u^2}\Phi = \Sigma^{-1} \end{aligned} \tag{17}$$

Hence the inverse of the asymptotic variance of $\sqrt{n}(\hat{\gamma} - \gamma)$ reaches the semiparametric efficiency bound when the error is conditional homoskedastic.

I emphasize here that the Chamberlain's efficiency bound of (16) does not impose conditional homoskedasticity.[7] If one were to impose the condition of conditional homoskedasticity, one would get a different (sharper) efficient bound that makes use of the second moment restriction. Therefore, the preceding discussion that my estimator reaches the semiparametric efficiency bound under homoskedasticity should be interpreted as "local" efficiency. A local efficiency semiparametric estimator is one that is efficiency in some semiparametric model when some restrictions are satisfied. By comparison, a globally efficient semiparametric estimator is one that is efficient whether or not those restrictions are satisfied. Obviously, my semiparametric estimator $\hat{\gamma}$ is not efficient when the error is conditional heteroskedastic. Nevertheless, some simple modifications to my estimation procedure can lead to an efficient estimator of γ under the general conditional heteroskedastic errors; see Section 3 for a discussion on this possible extension.

If one is mainly interested in estimating γ [treating $g(z) = \sum_{l=1}^{L} g_l(z_l)$ as nuisance parameter], one can ignore the additive structure of $g(z) = \sum_{l=1}^{L} g_l(z_l)$ and estimate γ $\sqrt{n}$-consistently by using, say, $\bar{p}^K(z) = [\bar{p}_1^K(z), \ldots, \bar{p}_K^K(z)]$ as the approximating function, $K \to \infty$ as $n \to \infty$, where $\bar{p}^K(z)$ is the first K terms of some series function that does not impose additive separable restriction in $z_1, \ldots, z_L$. However, there are at least three drawbacks to using this approach.

First, the unknown function is treated as a function of dimension q with $q > \max\{q_1, \ldots, q_L\}$. Hence it will suffer the "curse of dimensionality" problem.

Second, let $\tilde{\gamma}$ denote the semiparametric estimator of γ without imposing the additive structure on $\bar{p}^K(z)$ [i.e., ignoring that additive structure of $g(z) = \sum_{l=1}^{L} g_l(z_l)$], then under the conditional homoskedastic error assumption, i.e., $E(u_i^2|X_i, Z_i) = E(u_i^2) = \sigma_u^2$, the asymptotic variance of $\sqrt{n}(\tilde{\gamma} - \gamma)$ is $\sigma_u^2(E\{[X_i - E(X_i|Z_i)][X_i - E(X_i|Z_i)]'\})^{-1}$ (e.g., Robinson, 1988; Newey, 1997: Theorem 2), while from Theorem 1, I know that the asymptotic variance of $\sqrt{n}(\hat{\gamma} - \gamma)$ is $\sigma_u^2(E\{[X_i - h(Z_i)][X_i - h(Z_i)]'\})^{-1}$. Hence, when $E(X|Z = z)$ is not

[7] I owe this observation and the discussions of this paragraph to a referee.

additive separable in $(z_1, \ldots, z_L)$, I will have $E(X|Z = z) \neq h(z) \equiv \sum_{l=1}^{L} h_l(z)$. Consequently, $\tilde{\gamma}$ will be asymptotically less efficient than $\hat{\gamma}$.

Third, one needs a stronger condition that $E(v_i v_i')$ is positive definite in order for $\tilde{\gamma}$ to be well defined (for γ to be identified). This rules out the case that X_i is a deterministic function of Z_i. The identification condition of Theorem 1 is weaker than the preceding, and it allows X_i to be a deterministic function of Z_i, say, $X_i = m(Z_i)$, as long as this function is not additive separable in $(Z_{1i}, \ldots, Z_{Li})$. Consider a simple case of $L = 2$ and Z_{1i} and Z_{2i} are scalars. Let $X_i = Z_{1i}Z_{2i}$; then model (3) becomes

$$Y_i = (Z_{1i}Z_{2i})\gamma + g_1(Z_{1i}) + g_2(Z_{2i}) + u_i \tag{18}$$

γ in (18) is identified because $X_i = Z_{1i}Z_{2i}$ is not an additive separable function in Z_{1i} and Z_{2i}. Model (18) has the advantage that it only involves one-dimensional nonparametric regression functions; hence it does not suffer the "curse of dimensionality" problem. Also, it allows an interaction term (enters the model parametrically) so that it is more general than an additive model that does not allow any interaction terms. In practice, one can replace the interaction term $X_i = Z_{1i}Z_{2i}$ by any other known (nonadditive) function of (Z_{1i}, Z_{2i}). Clearly, γ in (18) is not identified if one ignores the additive structure of $g(z) = g_1(z_1) + g_2(z_2)$.

The next theorem gives the convergence rates of $\hat{g}(z) = p^K(z)'\hat{\beta}$ to $g(z) = \sum_{l=1}^{L} g_l(z_l)$.

THEOREM 2. *Under Assumptions* 1 *to* 3, *denote by* $\mathcal{S}$ *the support of* z; *then we have*

(*i*) $\sup_{z\in\mathcal{S}} |\hat{g}(z) - g(z)| = O_p[\zeta_0(K)](\sqrt{K}/\sqrt{n} + \sum_{l=1}^{L} K_l^{-\delta_l})$.
(*ii*) $n^{-1}\sum_{i=1}^{n}[\hat{g}(Z_i) - g(Z_i)]^2 = O_p(K/n + \sum_{l=1}^{L} K_l^{-2\delta_l})$.
(*iii*) $\int[\hat{g}(z) - g(z)]^2 dF(z) = O_p(K/n + \sum_{l=1}^{L} K_l^{-2\delta_l})$, *where* $F(\cdot)$ *is the cumulative distribution function of* Z.

The proof of Theorem 2 follows similar arguments as in the proofs of Theorem 1 of Newey (1997) and Theorem 4.1 of Newey (1995) (also use the result of Theorem 1 above); see Appendix for details.

Theorem 2 basically says that the convergence rates for $\hat{g}(z)$ to $g(z)$ are the same whether γ is known or one uses estimated $\hat{\gamma}$ in constructing $\hat{g}(z)$. This is to be expected because $\hat{\gamma} - \gamma = O_p(n^{-1/2})$, which is faster than the convergence rate of nonparametric (series) estimators such as $\hat{g}(z)$.

I also can estimate $g_l(z_l)$ by $\hat{g}_l(z_l) = p_l^{K_l}(z_l)'\hat{\beta}_l$, where $\hat{\beta}_l$ is a $K_l \times 1$ vector obtained from $\hat{\beta} = (\hat{\beta}_1', \ldots, \hat{\beta}_L')'$. I have the following results:

THEOREM 3. *Under Assumptions* 1 *to* 3, *denote by* $\mathcal{S}_l$ *the support of* Z_l; *then we have, for* $l = 1, \ldots, L$,

(*i*) $\sup_{z_l\in\mathcal{S}_l} |\hat{g}_l(z_l) - g_l(z_l)| = O_p[\zeta_0(K)](\sqrt{K}/\sqrt{n} + K_l^{-\delta_l})$.
(*ii*) $n^{-1}\sum_{i=1}^{n}[\hat{g}_l(Z_{li}) - g_l(Z_{li})]^2 = O_p(K/n + K_l^{-2\delta_l})$.
(*iii*) $\int[\hat{g}_l(z) - g_l(z)]^2 dF_l(z_l) = O_p(K/n + K_l^{-2\delta_l})$, *where* $F_l(\cdot)$ *is the cumulative distribution function of* Z_l.

The proof of Theorem 3 is basically the same as that of Theorem 2 and is omitted here.

Newey (1997) gives primitive conditions for power series and regression spline (B-splines) such that my Assumptions 1 to 3 hold. For your convenience, I restate these primitive conditions below. For the construction of B-spline functions, see Schumaker (1981).

ASSUMPTION 4. (i) The support of Z is a Cartesian product of compact connected intervals on which Z has an absolutely continuous probability density function that is bounded above by a positive constant and bounded away from zero; (ii) for $l = 1, \ldots, L$, $f_l(z_l)$ is continuously differentiable of order c_l on the the support of Z_l, where $f_l(\cdot) = g_l(\cdot)$ or $f_l(\cdot) = h_{(s),l}(\cdot)$ $(s = 1, \ldots, r)$.

ASSUMPTION 5. The support of Z is $[-1, 1]^q$.

When the support of Z is known and Assumption 4(i) is satisfied, Z can always be rescaled so that Assumption 5 holds.

Newey (1997:167) showed that for power series, Assumption 4(i) implies the smallest eigenvalue of $E[P^K(Z_i)P^K(Z_i)']$ is bounded for all K [$P^K(z) = Bp^K(z)$; see Assumption 2] and that $\zeta_0(K) = O(K)$. Also, it follows from Assumption 4(ii) and Lorentz (1966) that Assumption 3 holds with $\delta_l = c_l/r_l$, $l = 1, \ldots, L$. Thus Assumption 4 gives primitive conditions for Assumptions 2 and 3 for power series. Also, Newey (1997) showed that Assumptions 4 and 5 imply that Assumptions 2 and 3 hold for B-splines with $\zeta_0(K) = O(\sqrt{K})$. I summarize the preceding results in two corollaries below:

COROLLARY 1. *For power series, if Assumption* 1 *and Assumption* 4 *are satisfied and* $K^3/n \to 0$ *as* $n \to \infty$, *then*

(*i*) *The conclusion of Theorem* 1 *holds true*.
(*ii*) *The conclusions of Theorems* 2 *and* 3 *hold true with* $\zeta_0(K)$ *replaced by* K.

COROLLARY 2. *For B-splines, if Assumptions* 1, 4, *and* 5 *are satisfied and* $K^2/n \to 0$ *as* $n \to \infty$, *then*

(*i*) *The conclusion of Theorem* 1 *holds true*.
(*ii*) *The conclusions of Theorems* 2 *and* 3 *hold true with* $\zeta_0(K)$ *replaced by* $\sqrt{K}$.

Corollaries 1 and 2 show that the conclusions of Theorems 1 through 3 hold under primitive conditions for power series and splines.

When all the z_l's are all scalars and all the unknown functions [$g_l(\cdot)$ and $h_{(s),l}(\cdot)$] are c-order differentiable, then I can choose all the K_l to have the same order, say, $K_l = K/L$ for all $l = 1, \ldots, L$, and the condition on K becomes $K^3/n \to 0$ and $nK^{-2c} \to 0$ for power series [need $c > (3/2)$]; and $K^2/n \to 0$ and $nK^{-2c} \to 0$ for splines, requiring $c > 1$. I see that the condition on K is weaker for splines than for power series. Also, splines are piecewise polynomials of low order; hence they are less sensitive to outlier observations compared with power series.

3. POSSIBLE EXTENSIONS

In this section I discuss three possible extensions without providing any technical details. First, I propose a more efficient estimation procedure when the error is conditional heteroskedastic. Second, I claim that when the model is misspecified, my series estimator $x'\hat{\gamma} + \hat{g}(z)$ estimates an additive partially linear function that best approximates the unknown regression (in the class of additive partially linear functions) in the mean square sense. Third, I discuss the case of data-driven choice of K.

Theorem 1 holds true even when the error is conditional heteroskedastic, say, $E(u_i^2|X_i, Z_i) = \sigma^2(X_i, Z_i)$. However, $\hat{\gamma}$ is not semiparametric efficient in this case. Assume for the moment that $\sigma^2(x, z)$ is known. Let $\sigma_i = \sqrt{\sigma^2(X_i, Z_i)}$, and further assume that $\sigma^2(x, z)$ is bounded away from zero. Then one just needs to use least-squares regression of regressing Y_i/σ_i on $[X_i/\sigma_i, p^K(Z_i)'/\sigma_i]$. Intuitively, one would expect that the resulting estimator of γ is semiparametric efficient because u_i/σ_i is conditional homoskedastic. This is indeed the case. Let $\hat{\gamma}_{GLS}$ denote the corresponding estimator of γ; then following the similar proof of Theorem 1, one can show that the asymptotic variance of $\sqrt{n}(\hat{\gamma}_{GLS} - \gamma)$ is $J_0^{-1}A_0J_0^{-1} = J_0^{-1}$, where

$$J_0 = \inf_{\xi \in \mathcal{G}} E\{[X_i - \xi(Z_i)][X_i - \xi(Z_i)]'/\sigma^2(X_i, Z_i)\} \tag{19}$$

$$A_0 = \inf_{\xi \in \mathcal{G}} E\{[X_i - \xi(Z_i)][X_i - \xi(Z_i)]'u_i^2/\sigma^4(X_i, Z_i)\}$$

$$= \inf_{\xi \in \mathcal{G}} E\{[X_i - \xi(Z_i)][X_i - \xi(Z_i)]'/\sigma^2(X_i, Z_i)\} = J_0 \tag{20}$$

J_0 given in (19) is the same as (16), the semiparametric efficiency bound derived by Chamberlain (1992) when the error is conditional heteroskedastic. Note that if I let $a(z) \in \mathcal{G}$ denote the solution of the minimization problem of (19), $a(z)$ is usually different from $h(z)$ defined in (13) due to the weighting function $1/\sigma^2(x, z)$.

In practice, $\sigma^2(x, z)$ is unknown. One can use the (ordinary) least-squares method to first estimate γ and $g(z)$ by $\hat{\gamma}$ and $\hat{g}(z)$ as given in (9) and (10) and estimate u_i by $\hat{u}_i = Y_i - X_i'\hat{\gamma} - \hat{g}(Z_i)$. Then, based on $\{\hat{u}_i^2, X_i, Z_i\}_{i=1}^n$, one can obtain a consistent estimator of $\sigma^2(x, z)$, say, $\hat{\sigma}^2(x, z)$, using some nonparametric estimation methods (series, kernel, etc.). Denote $\hat{\sigma}_i = \sqrt{\hat{\sigma}^2(X_i, Z_i)}$; then regressing $Y_i/\hat{\sigma}_i$ on $[X_i'/\hat{\sigma}_i, p^K(Z_i)'/\hat{\sigma}_i]$ will result in a semiparametric efficient estimator of γ provided $\hat{\sigma}^2(x, z)$ converges to $\sigma^2(x, z)$ uniformly for all (x, z) in the (compact) support of (X, Z) with certain rates and perhaps with some other (extra) regularity conditions.

Next, I discuss the case that model (3) is misspecified in the sense that $E(Y|X, Z)$ is not an additive partially linear regression function. In this case, one can still estimate (3), but now $E(U_i|X_i, Z_i) \neq 0$. In general, U_i will be a (unknown) function of (X_i, Z_i). Thus, estimating (3) by series methods is asymptotically equivalent to finding $\beta_0 \in R^r$ and $g_0(z) \in \mathcal{G}$ such that $x'\beta_0 + g_0(z)$ best approximates $D(x, z) \equiv E(Y|X = x, Z = z)$ in the mean-square sense, i.e.,

$$E\{[Y - X'\gamma_0 - g_0(Z)]^2\} = \inf_{\gamma \in R^r, \xi \in \mathcal{G}} E\{[Y - X'\gamma - \xi(Z)]^2\} \tag{21}$$

$$\text{with } E(Y|X = x, Z = z) \neq x'\gamma_0 + g_0(z)$$

In fact, in the proof of Theorem 1, I have also shown that $p^K(z)'(P'P)^- P'X$ best approximates $\theta(z) = E(X|Z = z)$ in the mean-square error sense; see the proof of $S_{X-\tilde{X}} = \Phi + o_p(1)$ in the Appendix. Denote $(P'P)^- P'X = \hat{\beta}_x$ and $\hat{h}(z) = p^K(z)'\hat{\beta}_x$. I have shown that the series estimator $\hat{h}(z)$ consistently estimates an unknown function $h(z) \in \mathcal{G}$, where $h(z) \in \mathcal{G}$ that best approximates $\theta(z) = E(X|Z = z)$ [$\theta(z)$ does not belong to $\mathcal{G}$] in the mean-square-error sense; i.e., $h(z)$ is the solution of the following minimization problem:

$$E\{[X - h(Z)][X - h(Z)]'\} = \inf_{\xi \in \mathcal{G}} E\{[X - \xi(Z)][X - \xi(Z)]'\} \quad \text{with } E(X|Z = z) \neq h(z) \tag{22}$$

Following the same arguments as in the proof of Theorem 1, one can easily show that $\hat{\gamma}$ and $\hat{g}(z)$ given by (9) and (10) consistently estimate γ_0 and $g_0(z)$ that satisfy (21) even when model (3) is misspecified.

Finally, I mention that the results of this article can be generalized to the case of data-driven choice of K. For nonparametric estimation of regression functions using series methods, Craven and Wahba (1979), Li (1987), and Eubank and Jayasuriya (1993) showed that various cross-validation and generalized cross-validation methods are optimal for selecting the number of terms K. Andrews and Whang (1990) and Newey (1995) established similar results for additive regression models (without linear components). These data-driven methods of choosing K optimally can be directly applied to my case. The reason is that $\hat{\gamma}$ converges to γ at the rate of $O_p(n^{-1/2})$, which is faster than the convergence rate of the nonparametric series estimator $\hat{g}(z) - g(z)$. Hence the additional linear component $x'\gamma$ will not affect various data-driven optimal choices of K based on an additive regression model without a linear component.

4. CONCLUSION

In this article I propose to use general series methods to estimate an additive partially linear model. I show that γ is identified under weak conditions, and I establish the $\sqrt{n}$-normality result of the finite-dimensional parameter γ. I also show that $\hat{\gamma}$ is asymptotically more efficient than an estimator that ignores the additive structure of the model. When the error is conditional homoskedastic, my proposed estimator $\hat{\gamma}$ obtains the semiparametric efficiency bound of additive partially linear models. I also suggest an efficient estimator when the error is conditional heteroskedastic and show that the proposed estimators have optimal approximation properties in misspecified models. Linton and Härdle (1996) consider the problem of estimating an additive model with known links (with the kernel estimation method). It is possible to generalize the result of this article to the case that $E(Y_i|X_i, Z_i)$ is an additive partially linear function through a known link function using the series estimation method; however, the technical details will be more demanding. Recently, Horowitz (2000) considered a model suggesting that $E(Y_i|Z_i)$ is an additive function through an *unknown* link function (using the kernel method). It would be interesting to know whether the result of this article (using the series method) can be generalized to the case that $E(Y_i|X_i, Z_i)$ has an additive partially linear form through an unknown link function.

APPENDIX

Throughout this Appendix, I use C_1, C to denote generic constants. $\sum_i = \sum_{i=1}^n$. The norm $\|\cdot\|$ for a matrix A is defined by $\|A\| = [tr(A'A)]^{1/2}$.

PROOF OF THE EXISTENCE OF $h(z)$ THAT SATISFIES (13). I will first consider the case that X_i is a scalar. Let $f(z)$ and $m(z)$ be functions $\mathcal{S} \in R^q \to R$, and define the inner product $\langle f, m\rangle$ by $E(fm)$ [$E(f^2)$ and $E(m^2)$ are both finite]. Obviously, the class of function $\mathcal{G}$ defined in Definition 1 is a Hilbert space (a linear space with inner product). Below I show that there exists $h(z) = \sum_{l=1}^L h_l(z_l) \in \mathcal{G}$ that attains the following infimum bound:

$$E\left\{[\theta(Z_i) - \sum_{l=1}^L h_l(Z_{li})]^2\right\} = \inf_{\xi=\sum_l \xi_l \in \mathcal{G}} E\left\{[\theta(Z_i) - \sum_{l=1}^L \xi(Z_{li})]^2\right\} \tag{A.1}$$

where $\theta(Z) = E(X|Z)$. When $\theta(z) \in \mathcal{G}$, I have the simple solution of $h(z) = \theta(z)$. Hence I only need to consider the case that $\theta(z)$ does not belong to $\mathcal{G}$.

For expositional simplicity, I consider the case of $L = 2$ below; the proof of $L > 2$ follows the same argument. Let $\{a_j(z_1)\}_{j=1}^\infty$ be a complete base function that can expand any $g_1(z_1)$ and $\{b_j(z_2)\}_{j=1}^\infty$ and be a complete base function that can expand $g_2(z_2)$, where $g_1(z_1)$ and $g_2(z_2)$ are arbitrary functions with $g_1(z_1) + g_2(z_2) \in \mathcal{G}$. Then, obviously, $\{\phi_j\}_{j=1}^\infty$ is a complete base function that can expand any function $g(z) = g_1(z_2) + g_2(z_2) \in \mathcal{G}$, where the ordering of $\phi_j(z)$ is given by picking up the base functions from $\{a_j(z_1)\}_{j=1}^\infty$ and $\{b_j(z_2)\}_{j=1}^\infty$ alternatively, i.e.,

$$\begin{aligned}(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \ldots)\\ = [a_1(z_1), b_1(z_2), a_2(z_1), b_2(z_2), a_3(z_1), b_3(z_2), \ldots]\end{aligned} \tag{A.2}$$

Without loss of generality, I assume the base function $\{\phi_j(z)\}_{j=1}^\infty$ is orthonormalized. This can always be done using Gram-Schmidt orthonormalization procedure, since $\mathcal{G}$ is a Hilbert space (with inner product defined above).

Define $h(z) = h_1(z_1) + h_2(z_2)$ by

$$h(z) = \sum_{j=0}^\infty \phi_j(z)\beta_{(j)} \tag{A.3}$$

where $\beta_{(j)} = E[\theta(Z_i)\phi_j(Z_i)]$. Define $\eta(z) = \theta(z) - h(z)$. Then using (A.3), I get

$$\theta(z) = \sum_{j=0}^\infty \phi_j(z)\beta_{(j)} + \eta(z) \tag{A.4}$$

Multiply both side of (A.4) by $\phi_{(j)}(z)$ and take expectations; also using the orthonormal property of $\phi_j(\cdot)$ and the fact that $E[\theta(Z)\phi_j(Z)] = \beta_{(j)}$, I obtain

$$E[\phi_j(Z)\eta(Z)] = 0 \qquad (j = 1, 2, 3, \ldots) \tag{A.5}$$

(A.5) shows that η is orthogonal to the base function of $\{\phi_j\}_{j=1}^{\infty}$. Consequently, $\eta(\cdot)$ is orthogonal to any $f \in \mathcal{G}$, i.e.,

$$E[\eta(Z)f(Z)] = 0 \qquad \forall f \in \mathcal{G} \tag{A.6}$$

Using the orthonormal property of $\phi_j(\cdot)$ and the fact that $\eta \perp \phi_j$, I immediately have

$$E\{[\theta(Z)]^2\} = \sum_{j=0}^{\infty} \beta_{(j)}^2 + E\{[\eta(Z)]^2\} \tag{A.7}$$

$\sum_{j=0}^{\infty} \beta_{(j)}^2 < \infty$ follows from $E\{[\theta(Z)]^2\} < \infty$. Hence (A.7) implies that $\sum_{j=0}^{\infty} \phi_j(z)\beta_{(j)}$ will converge to a well-defined function in $\mathcal{G}$.

The facts that $\theta(\cdot) = h(\cdot) + \eta(\cdot)$, $h \in \mathcal{G}$, and $\eta \perp \mathcal{G}$ imply that $h(z)$ reaches the infimum of (A.1).

Next I discuss the case that $\theta(z)$ is an $r \times 1$ vector ($r > 1$). I use $\theta_{(s)}(z)$ to denote the sth component of $\theta(z)$, i.e., $\theta(z) = [\theta_{(1)}(z), \ldots, \theta_{(r)}(z)]'$. For $s \in \{1, \ldots, r\}$, I define $h_{(s)}(z)$ as in (A.3) with $\beta_{(j)}$ replaced by $\beta_{(s),j} = E[\theta_{(s)}(Z)\phi_j(Z)]$. Also define $h(z) = [h_{(1)}(z), \ldots, h_{(r)}(z)]'$ and $\eta(z) = \theta(z) - h(z)$. Then following the same arguments as above, one can easily show that $\eta \perp \mathcal{G}$ (since $\eta_{(s)} \perp \mathcal{G}$ for all $s = 1, \ldots, r$). Using $\eta = \theta - h$ and suppressing the argument (Z), I get for any $\xi \in \mathcal{G}$:

$$\begin{aligned} E[(\theta - \xi)(\theta - \xi)'] &= E[(\theta - h + h - \xi)(\theta - h + h - \xi)']\} \\ &= E[(\theta - h)(\theta - h)'] + E[(h - \xi)(h - \xi)'] + 0 \\ &= E[(\theta - h)(\theta - h)'] + \text{ a positive definite matrix} \end{aligned} \tag{A.8}$$

where the 0 comes from the facts that $\eta = \theta - h \perp \mathcal{G}$ and $h - \xi \in \mathcal{G}$. (A.8) and $h \in \mathcal{G}$ gives the desired result:

$$E[(\theta - h)(\theta - h)'] = \inf_{\xi \in \mathcal{G}} E[(\theta - \xi)(\theta - \xi)'] \tag{A.9}$$

■

PROOF OF THEOREM 1. Recall that $\theta(Z_i) = E(X_i|Z_i)$, $v_i = X_i - \theta(Z_i)$, $\epsilon_i = X_i - h(Z_i)$, and $\eta_i = \theta(Z_i) - h(Z_i)$ [$h(\cdot)$ is defined in (13)]. I will use the following short-hand notations: $\theta_i = \theta(Z_i)$, $g_i = g(Z_i)$, and $h_i = h(Z_i)$. Hence $v_i = X_i - \theta_i$, $\epsilon_i = \theta_i + v_i - h_i$, and $\eta_i = \theta_i - h_i$.

To avoid introducing too many notations for vector-matrix variables, I will, in the remaining part of this Appendix, use the same notation without subscript to denote a vector or a matrix. For example, θ is the $n \times r$ matrix with the ith row given by $\theta(Z_i)'$. This convention applies to h, g, η, ϵ, v, u, etc.

Recall that I define $\widetilde{A}$ by $\widetilde{A} = (P'P)^{-}P'A$; this definition applies to any $n \times 1$ or $n \times r$ matrices considered in this article. For example, $\tilde{\theta} = (P'P)^{-}P'\theta$. $\tilde{h}$, $\tilde{\eta}$, $\tilde{u}$, and $\tilde{v}$ are similarly defined.

From $X_i = \theta_i + v_i$ and $\theta_i = h_i + \eta_i$, I get $X_i = \eta_i + v_i + h_i$ and $\widetilde{X}_i = \tilde{\eta}_i + \tilde{v}_i + \tilde{h}_i$. Or in vector-matrix notation, $\mathscr{X} = \eta + v + h$ and $\widetilde{\mathscr{X}} = \tilde{\eta} + \tilde{v} + \tilde{h}$. Thus I have

$$\mathscr{X} - \widetilde{\mathscr{X}} = \eta + v + (h - \tilde{h}) - \tilde{v} - \tilde{\eta} \tag{A.10}$$

Equation (A.10) will be used frequently in the proofs below.

For scalars or column vectors A_i and B_i, I define $S_{A,B} = n^{-1}\sum_i A_i B_i'$. Also, $S_{A,A} = S_A$. Note that if $S_{X-\tilde{X}}^{-1}$ exists, then from (8) and (9), I immediately get

$$\sqrt{n}(\hat{\gamma} - \gamma) = S_{X-\tilde{X}}^{-1}\sqrt{n}S_{X-\tilde{X},g-\tilde{g}+u-\tilde{u}} \tag{A.11}$$

Obviously, Theorem 1 will be proved if I can show the following: (1) $S_{X-\tilde{X}} = \Phi + o_p(1)$ (hence $S_{X-\tilde{X}}$ is asymptotically nonsingular), (2) $S_{X-\tilde{X},g-\tilde{g}} = o_p(n^{-1/2})$, (3) $S_{X-\tilde{X},\tilde{u}} = o_p(n^{-1/2})$, and (4) $\sqrt{n}S_{X-\tilde{X},u} \to N(0,\Omega)$ in distribution. These are proved below.

(1) PROOF OF $S_{X-\tilde{X}} = \Phi + o_p(1)$. Using Equation (A.10), I have

$$S_{X-\tilde{X}} = S_{\eta+v+(h-\tilde{h})-\tilde{v}-\tilde{\eta}} = S_{\eta+v} + S_{(h-\tilde{h})-\tilde{v}-\tilde{\eta}} + 2S_{\eta+v,(h-\tilde{h})-\tilde{v}-\tilde{\eta}}.$$

First, $S_{\eta+v} = n^{-1}\sum_i(\eta_i + v_i)(\eta_i + v_i)' \equiv n^{-1}\sum_i \epsilon_i\epsilon_i' = \Phi + o_p(1)$ by virtue of a law of large numbers. Next, $S_{(h-\tilde{h})-\tilde{v}-\tilde{\eta}} \le 3\{S_{h-\tilde{h}} + S_{\tilde{v}} + S_{\tilde{\eta}}\} = o_p(1)$ by Lemma A.4, Lemma A.5(i), and Lemma A.5(iii). Finally, $S_{\eta+v,h-\tilde{h}-\tilde{v}-\tilde{\eta}} \le \{S_{\eta+v}S_{h-\tilde{h}-\tilde{v}-\tilde{\eta}}\}^{1/2} = \{O_p(1)o_p(1)\}^{1/2} = o_p(1)$ by the preceding results.

(2) PROOF OF $S_{X-\tilde{X},g-\tilde{g}} = O_p(\sum_{l=1}^L K_l^{-\delta_l}) = o_p(n^{-1/2})$. Using (A.10), $S_{X-\tilde{X},g-\tilde{g}} = S_{\eta+v+(h-\tilde{h})-\tilde{v}-\tilde{\eta},g-\tilde{g}} = S_{\eta+v,g-\tilde{g}} + S_{h-\tilde{h},g-\tilde{g}} - S_{\tilde{v},g-\tilde{g}} - S_{\tilde{\eta},g-\tilde{g}}$. I consider these four terms separately.

1. $S_{\eta+v,g-\tilde{g}} \le \{S_{\eta+v}S_{g-\tilde{g}}\}^{1/2} = O_p(\sum_{l=1}^L K_l^{-\delta})$ by Lemma A.4 and $S_{\eta+v} = O_p(1)$.
2. $S_{h-\tilde{h},g-\tilde{g}} \le \{S_{h-\tilde{h}}S_{g-\tilde{g}}\}^{1/2} = O_p(\sum_{l=1}^L K_l^{-2\delta_l})$ by Lemma A.4.
3. $S_{\tilde{v},g-\tilde{g}} \le \{S_{\tilde{v}}S_{g-\tilde{g}}\}^{1/2} = o_p(1)O_p(\sum_{l=1}^L K_l^{-\delta_l})$ by Lemma A.4 and Lemma A.5(i).
4. $S_{\tilde{\eta},g-\tilde{g}} \le \{S_{\tilde{\eta}}S_{g-\tilde{g}}\}^{1/2} = o_p(1)O_p(\sum_{l=1}^L K_l^{-\delta_l})$ by Lemma A.4 and Lemma A.5(iii).

(3) PROOF OF $S_{X-\tilde{X},\tilde{u}} = o_p(n^{-1/2})$. Using (A.10), $S_{X-\tilde{X},\tilde{u}} = S_{\eta+v+(h-\tilde{h})-\tilde{v}-\tilde{\eta},u} = S_{\eta,\tilde{u}} + S_{v,\tilde{u}} + S_{h-\tilde{h},\tilde{u}} - S_{\tilde{v},\tilde{u}} - S_{\tilde{\eta},\tilde{u}}$. I consider these five terms separately.

1. $E[\|S_{\eta,\tilde{u}}\|^2|\mathcal{X}] = n^{-2}tr[P(P'P)^-P'\eta\eta'P(P'P)^-P'E(uu'|\mathcal{X})] \le Cn^{-2}tr[\tilde{\eta}\tilde{\eta}'] = Cn^{-1}tr(S_{\tilde{\eta}}) = o_p(n^{-1})$ by Lemma A.5(iii). Hence $S_{\eta,\tilde{u}} = o_p(n^{-1/2})$.
2. $E[\|S_{v,\tilde{u}}\|^2|\mathcal{X},\mathcal{Z}] = n^{-2}tr[P(P'P)^-P'vv'P(P'P)^-P'E(uu'|\mathcal{X},\mathcal{Z})] \le Cn^{-2}tr \times[P(P'P)^{-1}P'vv'P(P'P)^{-1}P'] = Cn^{-2}tr[\tilde{v}\tilde{v}'] = Cn^{-1}tr(S_{\tilde{v}}) = O(K/n^2)$ by Lemma A.5(i). Hence $S_{v,\tilde{u}} = O_p(\sqrt{K}/n)$.
3. $S_{h-\tilde{h},\tilde{u}} \le \{S_{h-\tilde{h}}S_{\tilde{u}}\}^{1/2} = O_p(\sum_{l=1}^L K_l^{-\delta_l})O_p(\sqrt{K}/\sqrt{n})$ by Lemma A.4 and Lemma A.5(ii).
4. $S_{\tilde{v},\tilde{u}} \le \{S_{\tilde{v}}S_{\tilde{u}}\}^{1/2} = O_p(K/n)$ by Lemma A.5(i) and Lemma A.5(ii).
5. By the same argument as in the proof of (1) above, I have $E(\|S_{\tilde{\eta},\tilde{u}}\|^2|\mathcal{Z}) = n^{-2}tr[P(P'P)^-P'\eta'\eta P(P'P)^-P'E(uu'|\mathcal{Z})] \le Cn^{-2}tr[\tilde{\eta}'\tilde{\eta}] = Cn^{-1}tr(S_{\tilde{\eta}}) = Cn^{-1}o_p(1) = o_p(n^{-1})$ by Lemma A.5(iii). Hence $S_{\tilde{\eta},\tilde{u}} = o_p(n^{-1/2})$.

(4) PROOF OF $\sqrt{n}S_{X-\tilde{X},u} \to N(0, \Omega)$ IN DISTRIBUTION. $S_{X-\tilde{X},u} = S_{\eta+v+(h-\tilde{h})+v-\tilde{v}-\tilde{\eta},u} = S_{\eta+v,u} + S_{h-\tilde{h},u} - S_{\tilde{v},u} - S_{\tilde{\eta},u}$. I consider these four terms separately.

1. $\sqrt{n}S_{\eta+v,u} = n^{-1/2}\sum_i[\eta_i + v_i]u_i \to N(0, \Omega)$ in distribution by Levi-Lindberg central limit theorem.
2. $E(\|S_{h-\tilde{h},u}\|^2|\mathbb{X}) = n^{-2}tr[(h-\tilde{h})(h-\tilde{h})'E(uu'|\mathbb{X})] \le Cn^{-1}tr[(h-\tilde{h})'(h-\tilde{h})/n] = Cn^{-1}tr(S_{h-\tilde{h}}) = o_p(n^{-1})$ by Lemma A.4. Hence $S_{h-\tilde{h},u} = o_p(n^{-1/2})$.
3. By exactly the same arguments as in (2) above, I have $E(\|S_{\tilde{v},u}\|^2|\mathbb{X}) \le Cn^{-1}tr(S_{\tilde{v}}) = o_p(n^{-1})$ by Lemma A.5(i). Thus $S_{\tilde{v},u} = o_p(n^{-1/2})$.
4. $E(\|S_{\tilde{\eta},u}\|^2|\mathbb{X}) \le Cn^{-1}tr(S_{\tilde{\eta}}) = o_p(n^{-1})$ by Lemmas A.5(iii). Thus $S_{\tilde{\eta},u} = o_p(n^{-1/2})$. (1) to (4) above imply that $\sqrt{n}(\hat{\gamma}-\gamma) = \Phi^{-1}N(0,\Omega) + o_p(1) \to N(0, \Phi^{-1}\Omega\Phi^{-1})$ in distribution.

PROOF OF $\widehat{\Sigma} = \Sigma + o_p(1)$. $\widehat{\Sigma} = \widehat{\Phi}^{-1}\widehat{\Omega}\widehat{\Phi}^{-1}$. $\widehat{\Phi} \equiv S_{X-\tilde{X}} = \Phi + o_p(1)$ is proved in the proof of Theorem 1. Below I provide a sketch proof of $\widehat{\Omega} = \Omega + o_p(1)$ since the detailed proof is very similar to the proof of $\widehat{\Phi} = \Phi + o_p(1)$.

Using $\hat{\gamma} - \gamma = O_p(n^{-1/2})$ and $\hat{g}(Z_i) - g(Z_i) = o_p(1)$, it is easy to see that $\hat{u}_i = u_i + o_p(1)$. Also, by Lemma A.4, Lemma A.5(i), and Lemma A.5(iii), I know that $h_i - \tilde{h}_i = o_p(1)$, $\tilde{v}_i = o_p(1)$, and $\tilde{\eta}_i = o_p(1)$. Hence from (A.10) I know that $X_i - \widetilde{X}_i = \eta_i + v_i + (h_i - \tilde{h}_i) - \tilde{v}_i - \eta_i = \eta_i + v_i + o_p(1) \equiv \epsilon_i + o_p(1)$. These results lead to $\widehat{\Omega} = n^{-1}\sum_i \hat{u}_i^2(X_i - \widetilde{X}_i)(X_i - \widetilde{X}_i)' = n^{-1}\sum_{i=1}^n u_i^2\epsilon_i\epsilon_i' + o_p(1) = \Omega + o_p(1)$ by virtue of a law of large numbers. □

PROOF OF THEOREM 2. Basically, Theorem 2 is the same as Theorem 1 of Newey (1997) and Theorem 4.1 of Newey (1995) except that my estimator $\hat{g}(z)$ has an extra term of $x'(\hat{\gamma} - \gamma)$. Hence it suffices to show that the contribution of this extra term is asymptotically negligible. Intuitively, one expects that this is true because $(\hat{\gamma} - \gamma) = O_p(n^{-1/2})$, which has an order smaller than nonparametric series estimation convergence rate.

Let β_g satisfy Assumption 3 (with $f = g$). I consider $\|\hat{\beta} - \beta_g\|$ below:

$$\begin{aligned}
\hat{\beta} &= (P'P)^-P'(\mathcal{Y} - \mathcal{X}\hat{\gamma}) \\
&= (P'P)^-P'[(\mathcal{Y} - \mathcal{X}\gamma) - \mathcal{X}(\hat{\gamma} - \gamma)] \\
&= \beta_g + (P'P)^-P'[(g - P\beta_g) + U] - (P'P)^-P'\mathcal{X}(\hat{\gamma} - \gamma) \\
&\equiv \beta_g + D_{1n} - D_{2n}(\hat{\gamma} - \gamma),
\end{aligned} \tag{A.12}$$

where $D_{1n} = (P'P)^-P'[(g - P\beta_g) + U]$ and $D_{2n} = (P'P)^-P'\mathcal{X}$. $\|D_{1n}\| = O_p(\sum_{l=1}^L K_l^{-\delta_l}) + \sqrt{K}/\sqrt{n})$ was proved by Theorem 4.1 of Newey (1995).

Note that $\mathcal{X} = \theta + v = h + \eta + v$ [see Equation (A.10)]; also note that $h \in \mathcal{G}$ and $\eta \perp \mathcal{G}$ [see Equation (A.6)]; thus I have

$$\begin{aligned}
\|D_{2n}\|^2 &= \|(P'P)^-P'(h + \eta + v)\|^2 = \|\tilde{\beta}_h + \tilde{\beta}_\eta + \tilde{\beta}_v\|^2 \\
&\le C\{\|\tilde{\beta}_h\|^2 + \|\tilde{\beta}_\eta\|^2 + \|\tilde{\beta}_v\|^2\} = n^{-1}O_p(\|\tilde{h}\|^2 + \|\tilde{\eta}\|^2 + \|\tilde{v}\|^2) \\
&= n^{-1}\{O_p(\|h\|^2) + o_p(1)\} = O_p(1)
\end{aligned}$$

by Lemma A.4, Lemma A.5(i) and Lemma A.5(iii). Hence $D_{2n}(\hat{\gamma} - \gamma) = O_p(1/\sqrt{n})$. Thus I see the contribution from the term associated with $(\hat{\gamma} - \gamma)$ has an order smaller than $O_p(\sqrt{K}/\sqrt{n})$. Thus $\|\hat{\beta} - \beta_g\| = O_p(\sqrt{K}/\sqrt{n} + \sum_{l=1}^{L} K_l^{-\delta_l})$, as in Newey (1995).

The rest of the proofs [to prove the conclusions (i) to (iii) of my Theorem 2] follows from the same arguments as in the proofs of Theorem 4.1 of Newey (1995) [for (ii) and (iii)] and Theorem 1 of Newey (1997) [for (i)].

In the remaining part of this Appendix, I give some lemmas that are used in the proof of Theorem 1. Following the same arguments as in Newey (1997), in the proof below I will assume that $B = I$ [hence $p^K(z) = P^K(z)$; see Assumption 2]. This is so because $\hat{g}$, $\tilde{g}$, $\tilde{u}$, etc. are all invariant to nonsingular transformations of $p^K(z)$. Also, I will assume $Q = E[p^K(Z_i)p^K(Z_i)'] = I$. This is so because, for a symmetric square root $Q^{-1/2}$ of Q^{-1}, $Q^{-1/2}p^K(z)$ is a nonsingular transformation of $p^K(z)$ satisfying

$$\tilde{\eta}_0(K) = \sup_{z \in \mathcal{S}} \|Q^{-1/2}p^K(z)\| \leq C\zeta_0(K)$$

Hence this transformation will not change the order of the bound $\zeta_0(K)$, see Newey (1997:161) for more discussion on this. Also, if I change $p^K(z)$ to $\bar{p}^K(z) \equiv Q^{1/2}p^K(z)$ and define $\bar{\beta} = Q^{-1/2}\beta$, Assumption 3 is satisfied because $|g(z) - p^K(z)'\beta| = |g(z) - \bar{p}^K(z)'\bar{\beta}|$. Thus all the assumptions still hold when $p^K(\cdot)$ is changed to $Q^{-1/2}p^K(\cdot)$.

I use $\mathbf{1}_n$ to denote an indicator function that takes value one if $(P'P)$ is invertible and zero otherwise. I will only explicitly use the indicator function $\mathbf{1}_n$ in the proof of Lemma A.2 and omit it in the proofs of Lemmas A.3 to A.5(iii) to simplify the notation. Whenever I have $(P'P)^{-1}$, it should be understood as $\mathbf{1}_n(P'P)^{-1}$, and since $\text{Prob}(\mathbf{1}_n = 1) \to 1$ almost surely, I will often omit the indicator function $\mathbf{1}_n$.

LEMMA A.1. *$\widehat{Q} - I = O_p(\zeta_0(K)\sqrt{K}/\sqrt{n})$, where $\widehat{Q} = (P'P/n)$.*

PROOF. See the proof of Theorem 1 of Newey (1997:161–162). □

LEMMA A.2. *$\|\tilde{\beta}_f - \beta_f\| = O_p(\sum_{l=1}^{L} K_l^{-\delta_l})$, where $\tilde{\beta}_f = (P'P)^{-}P'f$, β_f satisfies Assumption 3, $f = g$ or $f = h_{(s)}$, $s = 1, \ldots, r$.*

PROOF.

$$\begin{aligned}
\mathbf{1}_n\|\tilde{\beta}_f - \beta_f\| &= \mathbf{1}_n\|(P'P)^{-1}P'(f - P\beta_f)\| \\
&= \mathbf{1}_n\{(f - P\beta_f)'P(P'P)^{-1}(P'P/n)^{-1}P'(f - P\beta_f)/n\}^{1/2} \\
&= \mathbf{1}_n O_p(1)\{(f - P\beta_f)'P(P'P)^{-1}P'(f - P\beta_f)/n\}^{1/2} \\
&\leq O_p(1)\{(f - P\beta_f)'(f - P\beta_f)/n\}^{1/2} = O_p(\sum_{l=1}^{L} K_l^{-\delta_l})
\end{aligned}$$

by Lemma A.1, Assumption 3, and the fact that $P(P'P)^{-1}P'$ is idempotent. Finally, $\|\tilde{\beta}_f - \beta_f\| = O_p(\sum_{l=1}^{L} K_l^{-\delta_l})$ since $\text{Prob}(\mathbf{1}_n = 1) \to 1$. □

LEMMA A.3. $(P'\eta/n) = O_p(\zeta_0(K)/\sqrt{n}) = o_p(1)$.

PROOF. Note that $E[P_i\eta_i] = 0$ because $p^K(\cdot) \in \mathcal{G}$ and $\eta(\cdot) \perp \mathcal{G}$; thus I have

$$\begin{aligned} E\|P'\eta/n\|^2 &= n^{-2}\left\{\sum_i \sum_j E(P_i'P_j\eta_i\eta_j)\right\} \\ &= n^{-2}\left\{\sum_i E(P_iP_i'\eta_i^2) + \sum_i \sum_{j\neq i} E(P_i'\eta_i)E(P_j\eta_j)\right\} \\ &= n^{-2}\sum_i E(P_i'P_i\eta_i^2) \leq Cn^{-1}E(P_i'P_i) = O[(\zeta_0(K))^2/n] \end{aligned}$$

Hence $(P'\eta/n) = O_p\big[\zeta_0(K)/\sqrt{n}\big]$. □

LEMMA A.4. $S_{f-\tilde{f}} = O_p(\sum_{l=1}^L K_l^{-2\delta_l}) = o_p(n^{-1/2})$, *where* $f = g$ *or* $f = h$.

PROOF. Note that $\tilde{f} \equiv P\tilde{\beta}_f$; thus I have

$$\begin{aligned} S_{f-\tilde{f}} &= 2n^{-1}\|f - \tilde{f}\|^2 \leq n^{-1}\{\|f - P\beta_f\|^2 + \|P(\beta_f - \tilde{\beta}_f)\|^2\} \\ &= O\left(\sum_{l=1}^L K_l^{-2\delta_l}\right) + (\beta_f - \tilde{\beta}_f)'(P'P/n)(\beta_f - \tilde{\beta}_f) \\ &= O\left(\sum_{l=1}^L K_l^{-2\delta_l}\right) + O_p(1)\|\beta_f - \tilde{\beta}_f\|^2 = O_p\left(\sum_{l=1}^L K_l^{-2\delta_l}\right) \end{aligned}$$

by Assumption 3, Lemma A.1, and Lemma A.2. □

LEMMA A.5. (*i*) $S_{\tilde{v}} = O_p(K/n)$, (*ii*) $S_{\tilde{u}} = O_p(K/n)$, (*iii*) $S_{\tilde{\eta}} = o_p(1)$.

PROOF. (i) Similar to the proof of Theorem 1 of Newey (1997), I have

$$\begin{aligned} E[S_{\tilde{v}}|\mathbb{X}] &= n^{-1}E\{v'P(P'P)^{-1}P'v|\mathbb{X}\} = n^{-1}tr[P(P'P)^{-1}P'E(vv'|\mathbb{X})] \\ &\leq Cn^{-1}tr[P(P'P)^{-1}P'] = O(K/n) \end{aligned}$$

which implies $S_{\tilde{v}} = O_p(K/n)$.

(ii) Follow the same proof as in the proof of Lemma A.5(i).

(iii) $S_{\tilde{\eta}} = n^{-1}\tilde{\eta}'\tilde{\eta} = (\eta'P/n)(P'P/n)^{-1}(P'\eta/n) = O_p(\zeta_0^2(K)/n) = o_p(1)$ by Lemma A.1 and Lemma A.3. □

REFERENCES

AI, C., "A Semiparametric Efficiency Bound of a Disequilibrium Model Without Observed Regime," *Journal of Econometrics* 62 (1994), 143–63.

——— AND D. MCFADDEN, "Estimation of Partially Specified Nonlinear Models," *Journal of Econometrics* 76 (1997), 1–37.

ANDREWS, D. W. K., "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica* 59 (1991), 307–45.

——— AND Y. J. WHANG, "Additive Interactive Regression Models: Circumvention of the Curse of Dimensionality," *Econometric Theory* 6 (1990), 466–79.

BEGUN, J. M., W. HALL, W. M. HUANG, AND J. A. WELLNER, "Information and Asymptotic Efficiency in Parametric-Semiparametric Models," *Annals of Statistics* 11 (1983), 432–52.

CHAMBERLAIN, G. "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics* 32 (1986), 189–218.

———, "Efficiency Bounds for Semiparametric Regression," *Econometrica* 60 (1992), 567–96.

CHEN, R., W. HÄRDLE, O. LINTON, AND E. SEVERANCE-LOSSIN, "Estimation and Variable Selection in Additive Nonparametric Regression Models," in W. Härdle and M. Schimek, eds., *Proceedings of the COMPSTAT Conference* (Heidelberg: Phyisika Varlag, 1996).

CHEN, X., AND X. SHEN, "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica* 66 (1998), 289–314.

COSSLETT, S. R., "Efficiency Bounds for Distribution-Free Estimators of the Binary Choice And The Censored Regression Models," *Econometrica* 55 (1987), 559–85.

COX, D. D., "Approximation of Least Squares Regression on Nested Subspaces, *Annals of Statistics* 16 (1988), 713–32.

CRAVEN, P., AND G. WHABA, "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by Generalized Cross-Validation. *Numerical Mathematics* 31 (1979), 377–403.

DECHEVSKY, L., AND S. PENEZ, "On Shape-Preserving Probabilistic Wavelet Approximators," *Stochastic Analysis and Applications* 15:2 (1997), 187–215.

DONALD, S. G. AND W. K. NEWEY, "Series Estimation of Semilinear Regression," *Journal of Multivariate Analysis* 50 (1994), 30–40.

EASTWOOD, B. J., AND A. R. GALLANT, "Additive rules for Semiparametric Estimation that Achieve Asymptotic Normality," *Econometric Theory* 7 (1991), 307–40.

EUBANK, R. L. , *Spline Smoothing and Nonparametric Regression* (New York: Marcel-Dekker, 1988).

——— AND B. R. JAYASURIYA, "The Asymptotic Average Squared Error for Polynomial Regression," *Statistics* 24 (1993), 311–9.

FAN, J., W. HÄRDLE, AND E. MAMMEN, "Direct Estimation of Low Dimensional Components in Additive Models," *Annals of Statistics* 26 (1998), 943–71.

FAN, Y., AND Q. LI, "On Estimating Additive Partially Linear Models," mimeo, University of Guelph, 1996.

GALLANT, A. R., AND G. SOUZA, "On the Asymptotic Normality of Fourier Flexible Functional Form Estimates," *Journal of Econometrics* 50 (1991), 329–53.

HANSEN, L. P., J. C. HEATON, AND M. OGAKI, "Efficiency Bounds Implied by Multiperiod Conditional Moment Restrictions," *Journal of American Statistical Association* 83 (1988), 863–71.

HOROWITZ, J., "Nonparametric Estimation of a Generalized Additive Model with an Unknown Link Function," (2000), forthcoming in *Econometrica*.

LI, K. C., "Asymptotic Optimality for C_p, C_L, Cross-Validation, and Generalized Cross-Validation: Discrete Index," *Annals of Statistics* 15 (1987), 958–75.

LINTON, O. B., "Efficient Estimation of Additive Nonparametric Regression Models," *Biometrika* 84 (1997), 469–73.

——— AND W. HÄRDLE, "Estimating Additive Regression with Known Links," *Biometrika* 83 (1996), 529–40.

——— AND J. P. NIELSEN, "A Kernel Method of Estimating Structured Nonparametric Regression based on Marginal Integration," *Biometrika* 82 (1995), 91–100.

LORENTZ, G. G., *Approximation of Functions* (New York: Chelsea, 1966).
NIELSEN, J. P., AND O. B. LINTON, "An Optimization Interpretation of Integration and Back-Fitting Estimators for separable nonparametric models," *Journal of Royal Statistical Society B* 60 (1998), 217–22.
NEWEY, W. K., "Adaptative Estimation of Regression Models via Moment Restriction," *Journal of Econometrics* 38 (1988), 301–39.
———, "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics* 5 (1990), 99–135.
———, "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62 (1994a), 1349–82.
———, "Series Estimation of Regression Functionals," *Econometric Theory* 10 (1994b), 1–28.
———, "Kernel Estimation of Partial Means in a General Variance Estimator," *Econometric Theory* 10 (1994c), 233–53.
———, "Convergence Rates for Series Estimators," in G. S. Maddala, P. C. B. Phillips, and T. N. Srinavsan, eds., *Statistical Methods of Economics and Quantitative Economics*: Essays in Honor of C. R. Rao (Cambridge, MA: Blackwell, 1995).
———, "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics* 79 (1997), 147–68.
ROBINSON, P., "Root-*N* Consistent Semiparametric Regression," *Econometrica* 56 (1988), 931–54.
SCHUMAKER, L. L., *Spline Functions: Basic Theory* (New York: Wiley, 1981).
STOCK, J. H., "Nonparametric Policy Analysis," *Journal of American Statistical Association*, 84 (1989), 567–75.
STONE, C. J., "Additive Regression and Other Nonparametric Models," *The Annals of Statistics* 13 (1985), 685–705.
———, "The Dimensionality Reduction Principle for Generalized Additive Models," *The Annals of Statistics* 14 (1986), 592–606.
TOJSTHEIM, D., AND B. H. AUESTAD, "Nonparametric Identification of Nonlinear Time Series: Projections," *Journal of American Statistical Association* 89 (1994), 1398–1409.