Efficient Instrumental Variables Estimation of Nonlinear Models

Author(s): Whitney K. Newey

Source: *Econometrica*, Jul., 1990, Vol. 58, No. 4 (Jul., 1990), pp. 809-837

Published by: The Econometric Society

Stable URL: https://www.jstor.org/stable/2938351

Econometrica, Vol. 58, No. 4 (July, 1990), 809–837

EFFICIENT INSTRUMENTAL VARIABLES ESTIMATION OF NONLINEAR MODELS[1]

BY WHITNEY K. NEWEY

This paper considers asymptotically efficient instrumental variables estimation of nonlinear models in an i.i.d. environment. The class of models includes nonlinear simultaneous equations models and other models of interest. A problem in constructing efficient instrumental variables estimators for such models is that the optimal instruments involve a conditional expectation, calculation of which can require functional form assumptions for the conditional distribution of endogenous variables, as well as integration. Nonparametric methods provide a way of avoiding this difficulty. Here it is shown that nonparametric estimates of the optimal instruments can give asymptotically efficient instrumental variables estimators. Also, ways of choosing the nonparametric estimate in applications are discussed.

Two types of nonparametric estimates of the optimal instruments are considered. Each involves nonparametric regression, one by nearest neighbor and the other by series approximation. The finite sample properties of the estimators are considered in a small sampling experiment involving an endogenous dummy variable model.

KEYWORDS: Instrumental variables, nonlinear models, efficient estimation, nonparametric regression.

1. INTRODUCTION

THE ASYMPTOTIC EFFICIENCY of instrumental variables (IV) estimators of nonlinear models depends on the form of the instruments. Amemiya (1974, 1977) characterized the instruments that minimize the asymptotic covariance matrix of an IV estimator. However, using these optimal instruments presents difficulties. They often involve conditional expectations of nonlinear functions of endogenous variables. Therefore, their calculation often requires specification of the conditional distribution of the endogenous variables as well as integration. This feature presents conceptual and practical problems. One of the attractive properties of IV estimators is that their consistency does not depend on specifying the distribution of the endogenous variables. It would be useful if asymptotic efficiency also did not depend on distributional specification. In addition, even when the distribution is specified, calculation of the conditional expectation can be a formidable task.

Nonparametric estimation of the optimal instruments provides a way of circumventing the need for distributional specification and difficult calculation. The idea is to use nonparametric regression to estimate the conditional expectations that appear in the optimal instruments. Here two types of nonparametric

[1] Several results in this paper were originally included in a paper entitled "Efficient Estimation of Models with Conditional Moment Restrictions," which was presented at the 1986 European Meeting of the Econometric Society. That paper was done independently of Robinson (1987), although the current version makes use of his results. That paper has been split into two papers, one covering heteroskedastic cases and carrying the original title, and the other being this paper. Financial support was provided by NSF Grants SES-8410249 and SES-8606545. Three referees, D. Andrews, G. Chow, G. Chamberlain, L. Hansen, J. Hausman, D. McFadden, J. Powell, P. Robinson, and H. White provided helpful comments.

regression estimators will be considered, nearest neighbor and series approximation. It will be shown that, just as in linear models, estimation of the optimal instruments does not affect asymptotic efficiency, and an efficient IV estimator for nonlinear models can be obtained using nonparametric estimates of the optimal instruments. The estimators considered here represent feasible versions of Amemiya's (1974, 1977) best nonlinear two and three stage least squares.

The idea of using nonparametric estimates of the optimal instruments is related to previous work. Kelejian (1971) suggested using polynomials as instruments, which is closely related to the idea of estimating the instruments nonparametrically by series approximation. Also, Amemiya (1985) suggested searching among different specifications to find one that provides the best fit for the endogenous variables, which might be thought of as a kind of informal, nonparametric procedure. In addition, for other types of models there exists previous work on efficient instrumental variables estimation with nonparametric estimates of optimal instruments. Robinson (1976) considered efficient, feasible, instrumental variables estimation of dynamic linear models. Also, since generalized least squares (GLS) can be thought of as an efficient instrumental variables estimator (e.g. Amemiya (1985, pp. 11–12)) this previous work includes results on feasible GLS for times series regression by Hannan (1963) and for heteroskedasticity by Carroll (1982) and Robinson (1987).

In Section 2 the model is defined and previous results on efficient IV estimation are briefly reviewed. Section 3 discusses nearest neighbor estimation of the optimal instruments, and shows asymptotic efficiency of the resulting estimator under conditions there specified. Section 4 carries out the same exercise for series estimates. Section 5 reports results for a small sampling experiment, involving an endogenous dummy variable model of Heckman (1978). In each of Sections 3, 4, and 5 implementation issues concerning the choice of a nonparametric estimate are dealt with. Both series and nearest neighbor estimates involve a choice of a certain parameter (e.g. the number of series terms). Some suggestions for data-based choices of this parameter are discussed.

2. MODEL

The econometric model to be considered is one where there is a $s \times 1$ residual vector $\rho(z, \beta)$ and instrumental variables x satisfying

$$(2.1) \qquad E[\rho(z_i, \beta_0)|x_i] = 0,$$

where β_0 denotes the true value of a $q \times 1$ vector of parameters and $z_1, \dots, z_n$ are i.i.d. observations on a data vector z_i, which includes x_i among its components. Attention will be restricted to the homoskedastic case, where

$$(2.2) \qquad E[\rho(z_i, \beta_0)\rho(z_i, \beta_0)'|x_i] = \Omega,$$

for a constant matrix Ω. An important example of this model is one where $\rho(z, \beta)$ is a subvector of the residuals of a nonlinear simultaneous equations system.

Instrumental variable (IV) estimation methods for this model were developed by Amemiya (1974, 1977), Kelejian (1971), Jorgenson and Laffont (1974), and Berndt et al. (1974), among others. Such methods are motivated by the conditional moment restriction of equation (2.1). For any $q \times s$ matrix of instruments $A(x)$, consisting of functions of the instrumental variables, the orthogonality condition $E[A(x_i)\rho(z_i, \beta_0)] = 0$ will be satisfied. An estimator of β can be obtained by imposing the sample analog of this population orthogonality condition, i.e. as the solution to

$$\sum_{i=1}^{n} A(x_i)\rho(z_i, \hat{\beta})/n = 0.$$

Often an exact solution is not possible, and $\hat{\beta}$ is chosen to set the sample cross product between instruments and residuals to be close to zero, say

$$(2.3) \qquad \hat{\beta} = \operatorname{argmin}_{\beta \in B} \sum_{i=1}^{n} \rho(z_i, \beta)'A(x_i)' \left[\sum_{i=1}^{n} A(x_i)A(x_i)' \right]^{-1} \times \sum_{i=1}^{n} A(x_i)\rho(z_i, \beta).$$

This is one general definition of nonlinear instrumental variables estimators. The matrix $A(x)$ might consist of a linear combination of $I_s \otimes a(x)$, in which case this estimator uses the instrumental variables $a(x)$ for each residual, as in Jorgenson and Laffont (1974). This formulation also allows there to be different instrumental variables for different residuals, as in Berndt et al. (1974) and Amemiya (1977).

The asymptotic properties of estimators of this form have been worked out by Amemiya (1974, 1977), Burguete, Gallant, and Souza (1982), Hansen (1982), and others. Let

$$(2.4) \qquad D(x_i) = E[\partial\rho(z_i, \beta_0)/\partial\beta | x_i].$$

The asymptotic covariance matrix of $\hat{\beta}$ is

$$(2.5) \qquad \Lambda_A = (E[A(x_i)D(x_i)])^{-1} E[A(x_i)\Omega A(x_i)'](E[D(x_i)'A(x_i)'])^{-1}.$$

Asymptotic efficiency of $\hat{\beta}$ was studied by Amemiya (1974, 1977), Jorgenson and Laffont (1974), and Berndt et al. (1974). Amemiya (1977) showed that the optimal instruments are

$$(2.6) \qquad A^*(x) = F \circ D(x)'\Omega^{-1},$$

where F is any nonsingular constant matrix; see Chamberlain (1987) for a proof for the IV formulation adopted here. For this choice of instruments the asymptotic covariance matrix Λ_A becomes

$$(2.7) \qquad \Lambda^* = \{E[D(x_i)'\Omega^{-1}D(x_i)]\}^{-1}.$$

It should be noted that, although the form of the optimal instruments depends on homoskedasticity, the efficient IV estimator does not fully utilize equation (2.2). As discussed by MaCurdy (1982), one could add additional residuals to the system, consisting of differences of covariance parameters and products of residuals. An efficient method of moments estimator (see Hansen (1982)) that combines $A^*(x)\rho(z,\theta)$ with products of the additional residuals and functions of x may be more efficient than the best IV estimator, at the expense of being inconsistent if the homoskedasticity assumption is violated. An IV estimator with instruments $A^*(x)$ will remain consistent under heteroskedasticity, although the instruments will no longer be optimal; see Chamberlain (1987) for the form of the optimal instruments under conditional heteroskedasticity (given x) and Newey (1986) for efficient estimation with i.i.d. observations.

For an example of the form of the optimal instruments, consider the model

$$(2.8) \qquad y_i=\beta_{10}s_i+f(x_i,\beta_0)+\varepsilon_i, \qquad s_i\in\{0,1\}, \qquad E[\varepsilon_i|x_i]=0,$$

where β_{10} is the first element of β_0 and s_i is correlated with ε_i. This is one of the endogenous dummy variable models of Heckman (1978). It is quite simple, but has a number of important applications, e.g. Heckman and Robb (1985), and can be used to illustrate a number of features of the topic under consideration here. For this model $\partial\rho(z,\beta_0)/\partial\beta=f_\beta(x,\beta_0)+e_1s$, where $f_\beta(x,\beta)\equiv\partial f(x,\beta)/\partial\beta$ and e_1 is the first unit vector. Here the residual is a scalar so that Ω^{-1} can be subsumed in F. Thus, the optimal instruments are any nonsingular linear transformation of $D(x)=f_\beta(x,\beta_0)+e_1\pi(x)$, where $\pi(x)=\text{Prob}(s=1|x)$.

It is generally not feasible to use the optimal instruments $A^*(x)$ to form an efficient IV estimator. The functions $D(x)$, as well as the constants Ω, are unknown. However, if $D(x)$ is known up to some parameters, then estimating the optimal instruments is straightforward. The covariance matrix Ω can be estimated by

$$(2.9) \qquad \hat{\Omega}=\sum_{i=1}^{n}\rho(z_i,\hat{\beta})\rho(z_i,\hat{\beta})'/n,$$

where $\hat{\beta}$ is some initial consistent IV estimate, as could be obtained from equation (2.3) with $A(x)$ known. Since $D(x)$ is a conditional expectation, its parameters could be estimated by, say, least squares. The resulting estimate $\hat{D}(x)$ could be combined with $\hat{\Omega}$ to estimate the optimal instruments as

$$(2.10) \qquad \hat{A}(x)=\hat{D}(x)'\hat{\Omega}^{-1}.$$

If $D(x)$ is a sufficiently smooth function of the unknown parameters and the estimates converge sufficiently fast, then estimation of the optimal instruments will not affect the limiting distribution of the IV estimator. Consequently, it will be asymptotically efficient. For example, it is well known that if $D(x)$ is linear in unknown parameters then consistency of the parameter estimates will suffice. Intuitively, equation (2.1) implies that variation of $A(x)$ around $A^*(x)$ that is asymptotically small in an appropriate sense will have no effect on the asymptotic distribution.

There are many examples where it is useful not to specify the functional form of $D(x)$. Although the structure of the model may result in some elements of

$D(x)$ having known functional form, knowledge of the form of all these functions will often require calculation of the conditional expectation $E[\partial\rho(z,\beta_0)/\partial\beta|x]$. In many models this calculation will require the conditional distribution of z given x. Also, even when the distribution of z is specified, carrying out the required integration could be very difficult. *Nonparametric* methods provide a way of estimating the optimal instruments without relying on auxiliary distributional assumptions or difficult calculations. A nonparametric estimate $\hat{D}(x)$ of the conditional expectation can be used to form a nonparametric estimate of the optimal instruments, as in equation (2.10). By analogy with the parametric case, estimation of the optimal instruments should not affect the limiting distribution. In the next two sections asymptotic efficiency of IV estimators using nearest neighbor and series nonparametric regression estimates of $D(x)$ will be shown.

To illustrate, consider the endogenous dummy example (2.8). The f_β component of the optimal instruments could be estimated by $f_\beta(x,\hat{\beta})$. If the conditional distribution of s given x has a known functional form, then $\pi(x)$ could also be estimated. For example, when $s = 1(v(x,\gamma_0) - \eta > 0)$, where $1(B)$ is the indicator function for a set B, and η is normally distributed, s will follow a probit model. In this case a consistent estimate of $\pi(x)$ is $\Phi(v(x,\hat{\gamma}))$, where $\Phi(\circ)$ is the standard normal cumulative and $\hat{\gamma}$ is the probit maximum likelihood estimate. However, if the normality assumption fails, then the resulting IV estimator will be inefficient, although still consistent (assuming $\Phi(v(x,\hat{\gamma}))$ is asymptotically correlated with s). Asymptotic efficiency loss from distributional misspecification can be avoided by choosing $\hat{\pi}(x)$ to be a nonparametric estimate.

It will be convenient in what follows to restrict attention to linearized versions of IV estimators based on estimated optimal instruments. The computation of such estimators does not require iterative procedures, and they are technically convenient because the proof of consistency can be bypassed. Let $\hat{\beta}$ be an initial IV estimator. One Newton–Raphson step toward the solution of $\Sigma_{i=1}^{n}\hat{A}(x_i)\rho(z_i,\beta) = 0$ gives

$$(2.11)\qquad \tilde{\beta} = \hat{\beta} - \left[\sum_{i=1}^{n}\hat{A}(x_i)\partial\rho(z_i,\hat{\beta})/\partial\beta\right]^{-1}\sum_{i=1}^{n}\hat{A}(x_i)\rho(z_i,\hat{\beta}).$$

Alternatively, one could replace $\Sigma_{i=1}^{n}\hat{A}(x_i)\partial\rho(z_i,\hat{\beta})/\partial\beta$ by $\Sigma_{i=1}^{n}\hat{A}(x_i)\hat{\Omega}\hat{A}(x_i)'$ or $\Sigma_{i=1}^{n}\hat{A}(x_i)\rho(z_i,\hat{\beta})\rho(z_i,\hat{\beta})'\hat{A}(x_i)'$. Such a replacement will not affect the asymptotic properties of $\tilde{\beta}$, because each matrix, when divided by n, will be a consistent estimate of $(\Lambda^*)^{-1}$, although the finite sample properties of the estimators may be different. The first two estimators exploit homoskedasticity, which might lead to improved finite sample properties.

3. NEAREST NEIGHBOR ESTIMATION OF THE OPTIMAL INSTRUMENTS

The first approach to constructing estimates of the optimal instruments involves nearest neighbor nonparametric regression. Nearest neighbor estimators are considered instead of kernels because the random denominators of kernel estimators creates technical difficulties.

For expositional purposes it is helpful to describe briefly k-nearest neighbor ($k-NN$) estimates of conditional expectations. A more complete exposition can be found in Stone (1977) or McFadden (1985). $k-NN$ estimates of a conditional expectation $g(x)=E[h_j|x_j=x]$ calculated from observations h_j and x_j $(j=1,\dots,n)$, make use of weighted averages of h_j, where all observations but those with x_j among the k closest values to x receive zero weight. Consistency of the resulting estimator of $g(x)$ follows by letting k grow with the sample size at an appropriate rate. "Closeness" for values of x is defined using a scaled version of the Euclidean norm $\|\circ\|$. Let $\hat{\sigma}_l$ be some estimate of the scale of the lth component $(x)_l$ of x, satisfying the conditions given in Stone (1977). If $E(\|x_j\|^2)$ exists, then the sample standard deviation $\hat{\sigma}_l^2$ of the lth component $(x_j)_l$ of x_j will do. The distance between x_j and x is defined by $\{\Sigma_l[(x_j)_l-(x)_l]^2/\hat{\sigma}_l^2\}^{1/2}$.

The weights $\{W_j(x)\}$ for averaging values of $h(z_j)$ are constructed in the following way. For positive integers m and k let $\omega(m, k)$ be constants satisfying

$$(3.1) \qquad \omega(m,k)\geqslant 0, \qquad \omega(m,k)=0, \qquad m>k; \qquad \sum_{m=1}^{k}\omega(m,k)=1.$$

If there are no ties among the distances of x_j from x then for the observation j which has mth smallest distance of x_j from x, let $W_j(x)=\omega(m,k)$. If there are ties, follow the same procedure, but with equal weight given to observations for which x_j is equally distant from x. That is, for the n_1 observations with x_j closest to x let $W_j(x)=\Sigma_{m=1}^{n_1}\omega(m,k)/n_1$, for the n_2 observations with x_j next closest to x let $W_j(x)=\Sigma_{m=n_1+1}^{n_1+n_2}\omega(m,k)/n_2$, and so on. A $k-NN$ estimate of $g(x)$ is

$$(3.2) \qquad \hat{g}(x)=\sum_{j=1}^{n}W_j(x)h(z_j).$$

Examples of weights are the uniform weights $\omega(m, k)=1/k$ and the triangular weights $\omega(m, k)=2(k-m+1)/[k(k+1)]$, $(m\leqslant k)$. The motivation for triangular weights is that $\hat{g}(x)$ will be a smoother function of x than for uniform weights. Further examples of weights can be found in Stone (1977). Both of these examples are such that there is some positive constant W_0 with

$$(3.3) \qquad \omega(m,k)\leqslant W_0/k \qquad (m,k=1,2,\dots).$$

This additional condition will be assumed to hold throughout.

It is conceivable that small sample performance of $k-NN$ estimates can be improved by estimating part of this conditional expectation by a preliminary regression, a procedure referred to as "trend removal" by Stone (1977). The idea is that it may be possible to reduce bias for a given number of nearest neighbors, by removing an important part of the function, which would then allow one to average over a larger number of neighbors to reduce variance. To describe trend removal let $t(x,\gamma)$ be a known function of x and parameter $\gamma, \hat{\gamma}$ be the estimates from nonlinear least squares regression of h_j on $t(x_j,\gamma)$, and $\hat{h}_j = t(x_j, \hat{\gamma})$ the corresponding predicted values. A nonparametric estimate of $g(x)$ can be formed as the sum of $t(x,\hat{\gamma})$ and the $k-NN$ neighbor estimate of the residual $g(x) - t(x,\gamma)$, yielding

$$(3.4) \qquad \hat{g}(x) = t(x,\hat{\gamma}) + \sum_{j=1}^{n} W_j(x)\big(h_j - \hat{h}_j\big).$$

When estimating the optimal instruments it is useful to allow for functional form restrictions, which are sometimes implied by the model. In the endogenous dummy example $D(x) = f_\beta(x,\beta_0) + e_1\pi(x)$, so that the functional form of some components of $D(x)$ is known to be $f_\beta(x,\beta_0)$. Imposing such restrictions will not improve the asymptotic efficiency of the instrumental variables estimate, because the nonparametric estimate of the optimal instruments will already give efficiency. Nevertheless, it is reasonable to expect that imposing such restrictions might improve the finite sample properties of the estimator.

The following assumption allows for both detrending and restrictions on the form of $D(x)$.

Assumption 3.1: *There exists* γ_0 *and known* $T(x,\gamma)$ *such that* $D(x) = T(x,\gamma_0) + G(x)$, *where some elements of* $G(x)$ *may be known to be zero.*

It will be assumed that the elements of γ include β. Also, γ may include reduced form and detrending parameters. In the endogenous dummy example there is a natural such decomposition with $\gamma = \beta$, $T(x,\gamma) = f_\beta(x,\beta)$, and $G(x) = e_1\pi(x)$. The implication of some elements of $G(x)$ known to be zero is that the corresponding elements of $D(x)$ have a known functional form, given by the corresponding element of $T(x,\gamma)$. Alternatively, if an element of $G(x)$ is not necessarily zero, then the corresponding element of $T(x,\gamma)$ can be interpreted as a trend term.

A nearest neighbor estimate of $D(x)$ which allows for functional form restrictions and detrending is constructed in the following way. Let $\rho_\beta(z,\beta)$, $d(x)$, $t(x,\gamma)$, and $g(x)$ denote corresponding components of $\partial\rho(z,\beta)/\partial\beta$, $D(x)$, $T(x,\gamma)$, $G(x)$, and let $h(z,\gamma) \equiv \rho_\beta(z,\beta) - t(x,\gamma)$. Let $\hat{\beta}$ be an initial IV estimator of β_0 and $\hat{\gamma}$ be an estimate of γ_0. To maintain a high level of generality a full description of $\hat{\gamma}$ will not be given here, although natural choices for $\hat{\gamma}$ are available in specific cases. For example, the components of γ that are components of β can be estimated by the corresponding components of $\hat{\beta}$. Also, if $t(x,\gamma)$ is a detrending term, then a natural choice of $\hat{\gamma}$ is the (nonlinear) least squares estimator from a regression of $\rho_\beta(z_i,\hat{\beta})$ on $t(x_i,\gamma)$. With such $\hat{\gamma}$ in hand, in the case where $g(x)$ is known to be zero $d(x_i)$ can be estimated by

$$(3.5) \qquad \hat{d}_i = \hat{t}_i \equiv t(x_i,\hat{\gamma}).$$

If $g(x)$ is not known to be zero, for each i let $W_{ii} = 0$ and for $j \neq i$ let

$W_{ij}=W_j(x_i)$ be the $k-NN$ weight associated with the sample that excludes the ith observation. For $\hat{h}_i=h(z_i,\hat{\gamma})$, adding $\hat{t}_i$ to the $k-NN$ estimate of $g(x)$ calculated from $\hat{h}_i=h(z_i,\hat{\gamma})$ gives

$$\hat{d}_i=\hat{t}_i+\sum_{j=1}^{n} W_{ij}\hat{h}_j=\hat{t}_i+\sum_{j=1}^{n} W_{ij}\left[\rho_\beta\left(z_j,\hat{\beta}\right)-\hat{t}_j\right]. \tag{3.6}$$

The estimates $\hat{d}_i$, each of which corresponds to an element of $D(x_i)$, can be combined in the obvious way to form an estimate $\hat{D}_i$ of $D(x_i)$, $(i=1,\dots,n)$. An estimate of the optimal instruments can then be constructed as in equation (2.10), and a one-step estimator obtained as in equation (2.11).

The device of not using the ith observation in the $k-NN$ estimate of $D(x_i)$ was employed by Robinson (1987) and is technically convenient. An estimate of the instruments could also be constructed by using the ith observation, and the efficiency result will still hold (under stronger regularity conditions), as shown in Newey (1986). The small sample effect of excluding or including the ith observation in the estimation of $D(x_i)$ will be considered in the Monte Carlo example of Section 5.

In the endogenous dummy example, for $T(x,\gamma)=f_\beta(x,\beta)$ and $G(x)=e_1\pi(x)$, the resulting estimate of $D(x_i)$ is

$$\hat{D}_i=f_\beta\left(x_i,\hat{\beta}\right)+e_1\sum_{j=1}^{n} W_{ij}s_j. \tag{3.7}$$

It might also be desirable in this example to allow for a trend term. A natural trend term here is the predicted probability $\hat{\Psi}_i$ from a binary choice model such as the linear probability model, probit, or logit. Such a trend term could be allowed for here by specifying γ to include the binary choice parameters and $T(x,\gamma)=f_\beta(x,\beta)+e_1\Psi(x,\gamma)$, where $\Psi(x,\gamma)$ is the binary choice probability. Here

$$\hat{D}_i=f_\beta\left(x_i,\hat{\beta}\right)+e_1\left[\hat{\Psi}_i+\sum_{j=1}^{n} W_{ij}\left(s_j-\hat{\Psi}_j\right)\right]. \tag{3.8}$$

The small sample properties of the resulting IV estimator will depend on the choice of k. One would expect that the best choice of k will vary with the model and data. Thus, it would be useful to use a data-based method for choosing k. Although the asymptotic distribution of the estimator provides no guide to the choice of k, since its limiting distribution is efficient (i.e. the same) for all suitably chosen of k, there are several reasonable data-based methods.

The parameters of interest are the estimates of β, so that it seems appropriate to use measures of their performance. One possibility would be to minimize a bootstrap estimator of the standard deviation of $\tilde{\beta}$, in a way similar to that considered by Hsieh and Manski (1987) for adaptive estimation of a regression model. Such a method is computationally intensive, so that alternatives are worth considering.

Another possibility is to base the choice of k on some performance measure for the estimate of the optimal instruments. The hope here is that the IV estimator with the optimal instruments has good small sample properties, and that these properties are inherited by instruments that are chosen to approximate the optimal instruments. One such performance measure involves cross-validation, which was recently considered by Robinson (1988) in the context of generalized least squares for time series regression. A cross-validated choice of k would be one which minimizes a sum of squared prediction errors, where the prediction is calculated from different observations than those to be predicted. In the current context, the nearest neighbor estimate already has the ith observation excluded, so that such a cross-validated choice of k could be obtained by minimizing

$$(3.9)\qquad C\hat{V}(k)=\sum_{i=1}^{n}\left(\hat{h}_i-\sum_{j=1}^{n}W_{ij}\hat{h}_j\right)^2.$$

The small properties of this choice of k will be considered in the Monte Carlo example of Section 5.

The asymptotic efficiency result will allow for k to be sample based in a limited way. The specific assumption is as follows:

Assumption 3.2: *k is (possibly) random and with probability approaching one $k\in\{k_1(n),\ldots,k_{\mathscr{L}}(n)\}$, where $\{k_l(n)\}$, $(l=1,\ldots,\mathscr{L})$, are nonrandom sequences satisfying $k_l(n)/\sqrt{n}\rightarrow\infty$ and $k_l(n)/n\rightarrow 0$.*

This assumption allows for a data based choice of k from among a finite number of values satisfying the given growth rate conditions. This assumption severely restricts the choice of k, but even this small degree of data-based flexibility may help avoid a very bad choice of k. It would be useful to extend this result to allow for a larger choice set for k, and to show that particular data-based choices of k automatically satisfy the growth rate conditions. Such results present technical challenges that are beyond the scope of this paper.

The growth rate for k specified here is useful in showing asymptotic efficiency, but is not known to be optimal in any sense. In any case it is not clear the optimal growth rate for estimation of the instruments is also optimai where the estimate of β is concerned. Deriving an optimality result for the estimation of β is difficult. It necessarily involves refinements to asymptotic distribution theory for $\hat{\beta}$. Such a result is beyond the scope of this paper.

Further regularity and identification conditions are helpful for an asymptotic efficiency result. For any matrix B let $\|B\|=[\mathrm{tr}(B'B)]^{1/2}$.

Assumption 3.3: (a) *$E[A(x_i)\rho(z_i,\beta)]=0$ has a unique solution on B at β_0*; (b) *β_0 is an element of the interior of B, which is compact*; (c) *$\rho(z,\beta)$ is differentiable at each $\beta\in B$ with probability one*; (d) *$E[\|\partial\rho(z,\beta_0)/\partial\beta\|^{\nu}]<\infty$ for*

$\nu > 2$; (e) $E[\|A(x)\|^{\nu}] < \infty$; (f) *$sup_{\beta \in B} \|\rho(z,\beta)\| \leqslant M_0(z)$ for $M_0(z)$ such that* $E[M_0(z)^{2\nu/(\nu-2)}] < \infty$.

The next assumption concerns smoothness conditions that are useful for showing asymptotic efficiency. For a matrix $F(\beta)$ let $\partial^l F(\beta)/\partial\beta^l$ denote a vector of all distinct partial derivatives of order l of all distinct elements of $F(\beta)$, and include β as a subvector of γ.

ASSUMPTION 3.4: (a) $E[D(x)'D(x)]$ *and* Ω *are nonsingular, and* $E[A(x)D(x)]$ *has rank* q; (b) $\sqrt{n}(\hat{\gamma} - \gamma_0)$ *is bounded in probability*; (c) *there exists a neighborhood* N *of* γ_0 *such that* $T(x,\gamma)$ *and* $\rho(z,\beta)$ *have continuous partial derivatives with respect to* γ *on* N *up to order* 2 *and* d *respectively*; (d) *for* $(l = 0,1,2)$, $sup_N \|\partial^l T(x,\gamma)/\partial\gamma^l\| \leqslant M_l^T(z)$ *such that* $E[M_l^T(z)^{\nu}] < \infty$, $l = 0, 1$, *and* $E[M_2^T(z_i)^{2\nu/(\nu+2)}] < \infty$ *for* ν *from Assumption* 3.1; (e) *for* $(l = 1,\dots,d)$, *there exists* ν_l *such that* $M_l(z) \equiv max_N \|\partial^l\rho(z,\beta)/\partial\beta^l\|$ *satisfies* $E[M_l(z)^{\nu_l}] < \infty$.

The order of differentiability d of $\rho(z,\beta)$ will be specified in the hypotheses of the theorems to follow. This assumption is somewhat stronger than needed in the sense that it is possible to replace the assumption of existence and continuity of the d derivative with a Lipschitz condition on the $d-1$ derivative. However, allowing for such extra generality would add little at the expense of much notational complexity.

THEOREM 3.1: *Suppose that Assumptions* 3.1–3.4 *are satisfied with* $d = 3$, $\nu_1 = \nu_2 = \nu$, $\nu_3 = 2\nu/(\nu+2)$. *Then*

$$(3.10) \qquad \sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, \Lambda^*), \qquad \left[\sum_{i=1}^{n} \hat{D}(x_i)'\hat{\Omega}^{-1}\hat{D}(x_i)/n\right]^{-1} \xrightarrow{p} \Lambda^*.$$

All proofs are in the Appendix. Besides stating the asymptotic efficiency of the one-step IV estimator $\tilde{\beta}$, this result also gives a consistent estimator of the asymptotic covariance matrix of $\tilde{\beta}$.

4. SERIES ESTIMATION OF THE OPTIMAL INSTRUMENTS

Another approach to estimation of the optimal instruments is nonparametric regression by series approximation. This method is related to Kelejian's (1971) suggestion of using polynomial instruments. Here an asymptotic efficiency result is presented. Asymptotic efficiency is obtained by letting the number and variety of terms in the series approximation grow at a specified rate with the sample size. Efficiency is a consequence of the arbitrarily good approximation, in the mean square sense, of conditional expectations by enough series terms.

It is helpful to describe briefly series estimates of conditional expectations. Such estimates have a long history in statistics, and have recently received attention in econometrics, e.g. Gallant (1981). Series estimates of a conditional expectation $g(x) = E[h(z)|x]$ make use of the first K terms,

$$(4.1) \qquad P^K(x) = (p_1(x), \dots, p_K(x))',$$

of a sequence of functions $(p_1(x), p_2(x), \dots)$. It is possible to allow $p_m(x)$ to depend on K and n without affecting the following results, but for notational convenience this extra generality will be avoided. A series nonparametric regression estimate is calculated from observations $h(z_j)$ and x_j, $(j = 1, \dots, n)$, as the predicted value obtained from the regression of $h(z_j)$ on $P^K(x_j)$. Let $P = [P^K(x_1), \dots, P^K(x_n)]'$, with the K superscript suppressed for convenience, and let $h = (h(z_1), \dots h(z_n))'$. A series estimate takes the form

$$(4.2) \qquad \hat{g}(x) = P^K(x)'(P'P)^- P'h,$$

where $\hat{\eta} = (P'P)^- P'h$ are the coefficients of a least squares regression of h on P, and B^- denotes a generalized inverse of a matrix B.

The presence of the generalized inverse allows for perfect multicollinearity among the columns of P. One generalized inverse corresponds to the deletion of redundant columns of P and regressing h on the remaining columns, as is done by some regression software. It should be noted that $\hat{g}(x)$ of equation (4.2) may not be invariant to the choice of generalized inverse, although the instrument estimate discussed below will be.

One example of $\hat{g}(x)$ involves a power series. Let the dimension of x be r. Let λ denote an r-dimensional vector of nonnegative integers and let $x^\lambda = (x_1)^{\lambda_1} \cdots (x_r)^{\lambda_r}$ denote a product of powers of the components of x. A basis sequence would take the form

$$(4.3) \qquad p_m(x) \equiv x^{\lambda(m)} \qquad (m = 1, 2, \dots),$$

with distinct $\lambda(m)$. A more robust alternative, which puts less weight on outlying observations in x, can be obtained by weighting by a function $w(x)$ that is small for large values of x and/or replacing each component x_l of x with a one-to-one, bounded function $v(x_l)$, such as $v(x_l) = x_l/(1 + |x_l|)$. For $v(x) = (v(x_1), \dots, v(x_r))'$ the resulting sequence is

$$(4.4) \qquad p_m(x) \equiv w(x)\left[v(x)^{\lambda(m)}\right] \qquad (m = 1, 2, \dots).$$

Trigonometric series are another example. Here x may have to be transformed to lie in $(0, 2\pi)^r$. See Gallant (1981) for trigonometric series formulas.

Trend removal can be carried out in a way that is analogous to that for nearest neighbor estimates. For $t(x, \gamma)$ and $\hat{\gamma}$ as discussed in Section 3, and the predicted trend vector $\hat{t} = (t(x_1, \hat{\gamma}), \dots, t(x_n, \hat{\gamma}))'$, a series estimate with a trend term is

$$(4.5) \qquad \hat{g}(x) = t(x, \hat{\gamma}) + P^K(x)'(P'P)^- P'(h - \hat{t}).$$

It should be noted that such a trend term will be redundant if it is a linear combination of $P^K(x)$. The trend is included here to allow a preliminary estimate that need not be linear in parameters, such as a parametric choice probability in the endogenous dummy example.

As discussed in Section 3, it may be useful to impose restrictions on the form of the optimal instruments. Such restrictions will be allowed for as in Assumption 3.1. A series estimate of $D(x)$ that allows for restrictions is constructed in the following way. Let $\rho_\beta(z,\beta), d(x), t(x,\gamma), g(x), h(z,\gamma) \equiv \rho_\beta(z,\beta) - t(x,\gamma)$, and $\hat{\gamma}$ be as specified in Section 3. If $g(x)$ is known to be zero, let the estimate of $d(x_i)$ be as given in equation (3.5). Otherwise, let

$$\hat{h} = (\hat{h}_1,\dots,\hat{h}_n)', \qquad \hat{h}_i = h(z_i,\hat{\gamma}) = \rho_\beta(z_i,\hat{\beta}) - t(x_i,\hat{\gamma}) \qquad (i=1,\dots,n),$$

and take the estimate of $d(x_i)$ to be

$$\hat{d}_i = t(x_i,\hat{\gamma}) + P^K(x_i)'(P'P)^- P'\hat{h}. \tag{4.6}$$

The resulting vector $(\hat{d}_1,\dots,\hat{d}_n)$ is the sum of the observations on the trend terms added to the predicted values from the least squares regression of $\hat{h}$ on P. The estimates $\hat{d}_i$, each of which corresponds to an element of $D(x_i)$, can be combined in the obvious way to form an estimate $\hat{D}_i$ of $D(x_i)$, $(i=1,\dots,n)$. An estimate of the optimal instruments can then be constructed as in equation (2.10), and a one-step estimator obtained as in equation (2.11).

Series estimates of the optimal instruments depend on the choice of the number of terms K. It is desirable to choose K based on the data. One could use measures of fit for choosing K, as suggested by Amemiya (1985). A cross-validation criteria for choosing K takes a particularly simple form. Let $\hat{\mu}_i = \hat{h}_i - P^K(x_i)'(P'P)^- P'\hat{h}$ be the ith residual from regression of $\hat{h}$ on the series terms and let $I_i = 1 - P^K(x_i)'(P'P)^- P^K(x_i)$. It is easy to check that if the second moment matrix of the series terms is nonsingular for each subsample, then the sum of squared, cross-validated residuals is

$$CV(K) = \sum_{i=1}^{n} (\hat{\mu}_i/I_i)^2. \tag{4.7}$$

The asymptotic efficiency result will allow for K to take on a data-dependent value $\hat{K}$:

Assumption 4.1: *$\hat{K} = \hat{K}(z_1,\dots,z_n,n)$ such that* (a) *$\hat{K} \xrightarrow{p} \infty$, and $\hat{K} = o_p(n^{(\nu-2)/2\nu})$;* (b) *for each $g(x_i)$ either* (*i*) *$\lim_{K\to\infty} E[(g(x_i) - P^K(x_i)'\eta_K)^2] = 0$ for some η_K and there exists $\mathcal{K}(n)$ such that the number of elements of $\mathcal{K}(n)$ is bounded and* $\text{Prob}(\hat{K} \in \mathcal{K}(n)) \to 1$, *or* (*ii*) *there exists $\zeta > 1$ and η_K such that $\lim_{K\to\infty} K^{\zeta}\{E[(g(x_i) - P^K(x_i)'\eta_K)^2]\}^{1/2} = 0$.*

Part (a) can be weakened to $\hat{K} \xrightarrow{p} \infty$ and $\hat{K} = o_p(\sqrt{n})$ if the conditional variance, given x, of each element of $\rho_\beta(z,\beta_0)$ is bounded. Part (b) involves an interaction between flexibility in the choice of K and the rate at which the series can approximate the unknown components of the instruments. If, for each sample size, $\hat{K}$ is chosen from among a finite number of values, then this assumption only requires that the series term can approximate $g(x)$ in mean square. For power series, it is well known that such approximation holds under weak conditions involving the characterization of the distribution of x by its moments; see Freud (1971). Sufficient conditions for weighted, transformed polynomials as in equation (4.4) can be obtained by suitably modifying Theorem 3 of Gallant (1980). Let

$$\mathscr{a}(K) = \max\{l \mid \forall\ \lambda \text{ with } |\lambda| \leqslant l\ \exists\ m \leqslant K \text{ with } \lambda(m) = \lambda\},$$

denote the maximum order such that the $P^K(x)$ includes all power terms up to and including that order.

LEMMA 4.1: *Suppose that* (a) *Assumptions* 3.3 (d) *and* 3.4 (d) *are satisfied*; (b) *there exists* $\tau > 0$ *such that* $E[w(x)^2 \exp\{\tau\|v(x)\|\}]$ *is finite*; (c) $\mathscr{a}(K) \to \infty$. *Then there exists* η_K *such that* $\lim_{K\to\infty}\{E[(g(x_i) - P^K(x_i)'\eta_K)^2]\}^{1/2} = 0$.

The hypotheses of this result are quite weak. Note in particular, that no assumption on the smoothness of $g(x)$ as a function of x is required for this result. Also, for some choices of $v(x)$ and $w(x)$ hypothesis (b) will automatically be satisfied, e.g. if $v(x)$ and $w(x)$ are bounded.

If $g(x)$ can be approximated by the series at a rate $K^{-\zeta}$ for $\zeta > 1$, then Assumption 4.1 allows $\hat{K}$ to be anything satisfying the rate conditions of Assumption 4.1 (a). Because the bounds on $\hat{K}$ are quite wide, it is plausible that certain data-based $\hat{K}$, such as cross-validation, will automatically satisfy these rate conditions under appropriate hypotheses. An automatic $\hat{K}$ is desirable because it frees the investigator from the task of choosing possible values of K that satisfy Assumption 4.1.

The following reasoning suggests that cross-validated $\hat{K}$ might satisfy Assumption 4.1 (a). It can be shown by arguments like those of Lemma A.12 that for nonrandom K,

$$\sum_{i=1}^{n} [\hat{g}(x_i) - g(x_i)]^2/n = O_p(n^{(2-\nu)/\nu}K) + O_p(K^{-2\zeta}), \tag{4.8}$$

where the two terms following the equality essentially represent variance and bias, respectively. Choosing $K_n^* = n^{(\nu-2)/[\nu(2\zeta+1)]}$ balances the two terms, yielding the best convergence rate $n^{-[(\nu-2)2\zeta]/[\nu(2\zeta+1)]}$ for $\sum_{i=1}^{n}[\hat{g}(x_i) - g(x_i)]^2/n$ that is obtainable from equation (4.8). If this rate is close to the best attainable, and cross-validated $\hat{K}$ behaves approximately like the optimal one, as is true for bandwidths for kernel nonparametric regression (Hardle et al. (1988)), then the rate conditions should be automatically satisfied for $\hat{K}$, since $1/(2\zeta + 1) < 1/2$ for $\zeta > 1$. Verification of this conjecture is beyond the scope of this paper.

Primitive conditions for the $K^{-\zeta}$ approximation rate of Assumption 4.1(b) will involve smoothness conditions on $g(x)$. In the univariate x case this assumption will hold for power series (with $\zeta = 2$) if $g(x)$ is twice continuously differentiable and the support of x is bounded; see Powell (1981). A literature search has not yet revealed an analogous result for the multivariate case. Nevertheless, if $g(x)$ is restricted to be analytic with geometric order bounds on the magnitude of derivatives, then an elementary Taylor expansion argument can be used to obtain an approximation rate. Let $v = v(x)$ and $g(v) = g(x^{-1}(v))$. Denote the partial derivatives of $g(v)$ on $\mathbb{R}^r$ by

$$D^{\lambda}g(v) = \left(\partial^{\lambda_1}/\partial v_1^{\lambda_1}\right) \cdots \left(\partial^{\lambda_r}/\partial v_r^{\lambda_r}\right)g(v),$$

where $\lambda = (\lambda_1, \ldots, \lambda_r)$ is a r-vector of nonnegative integers. The order of the derivative is $|\lambda| = \Sigma_{l=1}^{r}|\lambda_l|$. Also, let $\mathscr{O}(K) = \max_{m \leqslant K}|\lambda(m)|$ denote the maximum order of the power series terms included in $P^K(x)$.

LEMMA 4.2: *Suppose that $\mathscr{O}(K) = O(\mathscr{a}(K))$ and that there is a set $\mathscr{G}$ such that for all $g(x) \in \mathscr{G}$, $\mathscr{g}(v) \equiv w(x^{-1}(v))^{-1}g(v)$ can be taken to have a compact convex domain $V(g)$ containing the support of $v(x_i)$. Also suppose that there exists bounded $V \supseteq \cup_{g \in \mathscr{G}} V(g)$, that $w(v)$ is bounded, and there exists C such that for all λ, $D^{\lambda}\mathscr{g}(v)$ exists and $\sup_{g \in \mathscr{G}} \sup_{v \in V}|D^{\lambda}\mathscr{g}(v)| \leqslant C^{|\lambda|}$. Then $\lim_{K \to \infty} K^{\zeta} \sup_{g \in \mathscr{G}} \inf_{\eta_K}\{E[(g(x_i) - P^K(x_i)'\eta_K)^2]\}^{1/2} = 0$ for all $\zeta > 0$.*

Primitive conditions for an approximation rate for Fourier series follow from the results of Edmunds and Moscatelli (1977). By their Corollary 1 it will suffice that the support of x be a compact, convex subset of $(0, 2\pi)^r$, and that $g(x)$ be more than r times continuously differentiable, where r is the dimension of x.

The following result gives asymptotic efficiency for an IV estimator with a series estimator of the optimal instruments.

THEOREM 4.1: *Suppose that Assumptions* 3.3, 3.4, *and* 4.1 *are satisfied with $d = 2$, $\nu_1 = \nu$, $\nu_2 = 2$, and that for all K, $E[|p_K(x)|^{\nu}] < \infty$. Then*

$$\sqrt{n}\left(\tilde{\beta} - \beta_0\right) \xrightarrow{d} N(0, \Lambda^*), \qquad \left[\sum_{i=1}^{n} \hat{D}(x_i)'\hat{\Omega}^{-1}\hat{D}(x_i)/n\right]^{-1} \xrightarrow{p} \Lambda^*. \tag{4.9}$$

5. A SAMPLING EXPERIMENT

To see how the estimators might perform in finite samples a small sampling experiment was carried out. The model considered in the experiment was the endogenous dummy example discussed previously, with the following specification:

$$\begin{aligned} &y_i = \beta_{10}s_i + \beta_{20} + \varepsilon_i, \\ &s_i = 1(\alpha_{10} + \alpha_{20}x_i + \eta_i > 0), \end{aligned} \qquad \begin{bmatrix} \varepsilon_i \\ \eta_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}\right), \tag{2.8}$$

$$x_i \sim N(0,1), \qquad \beta_{10} = \beta_{20} = \alpha_{10} = \alpha_{20} = 1,$$

where x_i is distributed independently of ε_i and η_i. Two sample sizes were considered, $n = 100$ and $n = 200$. The number of replications was 400, with computations carried out via Gauss on a microcomputer.

Tables I, II, and III report results for a variety of estimators of β_{10}. Table I gives the performance of several parametric estimators. The estimators of Table I are ordinary least squares (OLS), an IV estimator with instruments $A(x) =$

TABLE I
PARAMETRIC INSTRUMENTAL VARIABLES

n	Estimator	Bias	Std Dev	RMS Ratio
100	OLS	.835	.215	1.97
	Dummy IV	−.032	.594	1.36
	Lin Prob	−.036	.458	1.05
	EIV	−.042	.435	1.00
200	OLS	.850	.142	2.82
	Dummy IV	−.024	.386	1.26
	Lin Prob	−.013	.325	1.06
	EIV	−.016	.306	1.00

$(1, 1(x > 0))'$ (Dummy IV), where the instrument for s is the dummy $1(x > 0)$, an IV estimator with $A(x) = (1, x)'$ (Lin Prob), which corresponds to a linear probability model instrument for s, and the efficient instrumental variables estimator (EIV), with instruments

$$D(x) = -(1, \pi(x))', \qquad \pi(x) = \Phi(1 + x).$$

The performance measures reported in the table are bias, being the Monte Carlo mean of the estimate minus the true value $\beta_{10} = 1$, the Monte Carlo standard deviation (Std Dev), and the ratio of the square root of the Monte Carlo mean square error for the estimate to that of the efficient IV estimator. It is possible that second moments of the IV estimators do not exist, so that the population values of these performance measures are not well defined. However, in this experiment the sample values seemed to be quite stable to changes in the number of replications, suggesting existence of moments may not be a problem.

Interesting features of Table I are the large bias of OLS, the tiny bias of the IV estimators, the relative inefficiency of Dummy IV, and the relative efficiency of the estimator with a linear instrument. One cause of the relative inefficiency of the dummy instrumental variable is that it jumps from zero to one at $x = 0$, which is one standard deviation (in x units) away from the inflection point of $\pi(x) = \Phi(1 + x)$ (which occurs at $x = -1$). The Dummy IV estimator was included in order to illustrate that making a substantial mistake about the form of the optimal instruments can have large efficiency costs.

Table II reports results for IV estimators with nearest neighbor estimates of the optimal instruments. The nearest neighbor estimates involve a moderate size grid for k, the grid being $(10, 15, 20, 25, 30, 35, 40)$ for $n =$ 100 and $(15, 22, 30, 37, 45, 52, 60)$ for $n = 200$. These grids are consistent with Assumption 3.2, in the sense that they are the same size for both sample sizes, and for corresponding grid values k/n falls and $k/\sqrt{n}$ rises as the sample size goes from 100 to 200. Included in Table II is the relative frequency of different values of k (Distribution of k). The first seven rows of results (Uniform) are for uniform weights with fixed values of k; see Section 3 for a description of the nearest neighbor estimate of $\pi(x)$. The other rows of the table present results for a sample based choice of k from the corresponding grid, with k chosen to

TABLE II
NEAREST NEIGHBOR INSTRUMENTS

						Distribution of k						
n	Estimator	Bias	Std Dev	RMS Ratio	$k =$	10	15	20	25	30	35	40
100	Uniform	−.197	.756	1.79		1						
		−.164	.717	1.68			1					
		−.152	.652	1.53				1				
		−.145	.597	1.41					1			
		−.154	.599	1.41						1		
		−.154	.591	1.40							1	
		−.162	.618	1.46								1
	Univ, *CV*	−.085	.550	1.27		.14	.19	.21	.19	.11	.10	.07
	Triang	−.101	.576	1.34		.08	.08	.07	.15	.17	.12	.34
	Uni, $+i$	.105	.429	1.01		.14	.19	.21	.19	.11	.10	.07
	Uni, Detr	−.052	.486	1.12								
					$k =$	15	22	30	37	45	52	60
200	Uni, *CV*	−.027	.326	1.07		.10	.17	.23	.23	.13	.08	.05

minimize $CV(k)$ of equation (3.9). The row labeled Triang gives results for triangular weights rather than uniform, that labeled Uni, $+i$ for inclusion of the ith observation in the uniform nearest neighbor estimate of $\pi(x_i)$, and Uni, Detr for a linear probability trend term. The final row has uniform weights without detrending and with the ith observation excluded, for $n = 200$.

The most striking feature of Table II is the poor performance of the nearest neighbor estimates for $n = 100$. In most cases the bias is more than two standard deviations away from zero, i.e. $|\text{Bias}| \geqslant 2(\text{Std Dev})/(400)^{1/2} = (\text{Std Dev})/10$, suggesting that the estimates are biased. The point estimates of the bias are also quite large. It is interesting to note that letting k be data-based improves the performance of the estimator, lowering both bias and variance by a substantial amount. Also, using triangular weights results in a slight increase in bias and standard deviation. Surprisingly, including the ith observation substantially alters the performance of the estimator. Possibly this is a result of the relatively high correlation of ε_i and η_i. The sensitivity of performance to inclusion of the ith observation is much diminished when $n = 200$, although for brevity the results are not reported here. The performance of the uniform weights with $n = 200$, reported in the last line of Table II, was somewhat more promising. The RMS Ratio is much reduced for the larger sample size.

Table III reports results for IV estimators with series estimates of the optimal instruments. The series estimates use a grid of values for K, the number of series terms, the grid being $(2,3,4,5,6)$ for $n = 100$ and $(3,4,5,6,7)$ for $n = 200$. These grids are consistent with Assumption 4.2, in the sense that the minimum order grows with the sample size and that the number of terms grows less than the square root of the sample size (note that ν can be taken as large as desired since s is bounded). Included in Table III is the relative frequency of different values of K (Distribution of K). The rows labeled Polynomial or Poly are for the series $p_m(x) = x^{m-1}$, $(m = 1,2,\ldots)$, with the first five rows involving fixed values of K; see Section 4 for a description of the series estimate of $\pi(x)$.

TABLE III

POLYNOMIAL INSTRUMENTS

						Distribution of K					
n	Estimator	Bias	Std Dev	RMS Ratio	$K =$	2	3	4	5	6	7
100	Polynomial	−.036	.458	1.05		1					
		−.028	.430	.99			1				
		.002	.416	.95				1			
		.018	.411	.94					1		
		.042	.399	.92						1	
	Poly., *CV*	.003	.425	.97		.13	.56	.14	.13	.05	
	Poly. in v	.014	.418	.96		.13	.65	.11	.05	.07	
200	Poly., *CV*	.004	.302	.99			.44	.20	.25	.06	.05

Because the moment generating function of the normal distribution exists, the hypotheses of Lemma 4.1 are satisfied for this series. The row labeled Poly in v was for the series $p_m(x) = v(x)^{m-1}$, $(m = 1, 2, \ldots)$, where $v(x) = x/(1 + |x|)$. The hypotheses of Lemma 4.2 are satisfied for this series and for $D(x) = \Phi(1 + x)$. The last three rows of Table III involve a sample based choice of K from the corresponding grid, with K chosen to minimize $CV(K)$ of equation (4.7).

The most striking feature of Table III is the outstanding performance of the series estimates. The estimated biases are tiny, and, surprisingly, the RMS ratio is less than one in many cases, although this phenomenon almost disappears for the larger sample size. The performance of the estimator is sensitive to the choice of K, suggesting the need for K to be data-based. The cross-validated K considered here performs well, with small bias, and RMS less than one. Also, it is interesting to note that the transformed power series estimator performs almost the same as the linear power series estimator. This occurs in spite of the fact that the inflection point of the transformation $v(x) = x/(1 + |x|)$ occurs at $x = 0$ rather than at the inflection point of $\pi(x) = \Phi(1 + x)$.

Another interesting feature of Table III is that the bias changes sign as the number of series terms increases. This is undoubtedly an artifact of the particular model considered here, and accounts in part for the tiny biases of the series estimates when a moderate number of terms is included. In general one might expect to find that the bias gets large as the number of terms increases. If $K = n$ and P has full rank, then the series estimate will give an exact fit, and the IV estimator reduces to least squares.

In summary, the series estimates of the optimal instruments perform quite well, but the nearest neighbor estimates are less satisfactory. Both sets of results suggest the usefulness of a sample based choice of the number of nearest neighbors or of series terms.

6. EXTENSIONS

Work on efficient instrumental variables estimation in heteroskedastic cases has been carried out concurrently, e.g. Newey (1986). It is also possible to extend the asymptotic efficiency result for series estimates of the optimal instruments to time series. Such an extension could be carried out via, say, mixing conditions and the type of arguments for efficient estimation by series approximation given here and in Newey (1988). This extension will be investigated in future work.

Department of Economics, Princeton University, Princeton, NJ 08544, U.S.A.

Manuscript received July, 1986; final revision received July, 1989.

APPENDIX

The proofs of the theorems will make use of a number of intermediate lemmas. Throughout, C will denote a generic constant that can take on different values in different appearances, and T, CS, H, and M will refer to the triangle, Cauchy-Schwarz, Hölder, and Markov inequalities, respectively. Also, the n subscript will often be suppressed for notational convenience.

Note that the estimates of the optimal instruments take the form $\hat{A}_i = \hat{D}_i'\hat{\Omega}^{-1} = (\hat{T}_i + \hat{G}_i)'\hat{\Omega}^{-1}$, where $\hat{T}_i$ is the matrix of trend terms, and $\hat{G}_i$ is a matrix consisting of the nonparametric estimates of the trend residuals and (possibly) zeros. The following lemma gives sufficient conditions for the conclusion of Theorems 3.1 and 4.1, for instruments of this form. Let

$$G_i = G(X_i), \qquad D_i = D(x_i), \qquad A_i = D_i'\Omega^{-1}, \qquad T_i = T(x_i, \gamma_0).$$

LEMMA A.1: *If Assumptions* 3.1, 3.3, *and* 3.4 *are satisfied with* $d = 1$ *and* $\nu_1 = \nu$, *and in addition,*

$$\text{(A.1a)} \qquad \sum_{i=1}^{n} \|\hat{G}_i - G_i\|^2/n = o_p(1),$$

$$\text{(A.1b)} \qquad \sum_{i=1}^{n} (\hat{G}_i - G_i) \otimes \rho(z_i, \beta_0)/\sqrt{n} = o_p(1),$$

then $\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, \Lambda^*)$, *for* $\tilde{\beta}$ *of equation* (2.11), *and*

$$\left(\sum_{i=1}^{n} \hat{D}_i'\hat{\Omega}^{-1}\hat{D}_i/n\right)^{-1} = \Lambda^* + o_p(1).$$

PROOF: The $\sqrt{n}$-consistency of the initial IV estimator under Assumptions 3.3–3.4 follows by a standard argument, so that the proof will be omitted. Consider any $\bar{\beta} = \beta_0 + o_p(1)$, and let $\Delta_i(\beta) = \partial\rho(z_i, \beta)/\partial\beta$. By Assumption 3.4 (e), $E[M_1(z_i)^2] < \infty$. With probability approaching one $\bar{\beta} \in N$, so that by T, CS, and M,

$$\text{(A.2)} \qquad \left\|\sum_{i=1}^{n} (\hat{A}_i - A_i)\Delta_i(\bar{\beta})/n\right\| \leqslant \sum_{i=1}^{n} \|\hat{A}_i - A_i\| \|\Delta_i(\bar{\beta})\|/n \leqslant \sum_{i=1}^{n} \|\hat{A}_i - A_i\| M_1(z_i)/n$$
$$\leqslant \left(\sum_{i=1}^{n} \|\hat{A}_i - A_i\|^2/n\right)^{1/2} \left(\sum_{i=1}^{n} M_1(z_i)^2/n\right)^{1/2}$$
$$= \left(\sum_{i=1}^{n} \|\hat{A}_i - A_i\|^2/n\right)^{1/2} O_p(1).$$

By Assumption 3.4 (d), $E[M_1^T(z_i)^2] < \infty$. Then by $\hat{\gamma} = \gamma_0 + o_p(1)$, a mean value expansion, and CS and M,

$$\text{(A.3)}\qquad \sum_{i=1}^{n} \|\hat{T}_i - T_i\|^2/n \leqslant \sum_{i=1}^{n} \|\partial T(x_i,\tilde{\gamma})/\partial\gamma\|^2 \|\hat{\gamma}-\gamma_0\|^2/n$$

$$\leqslant \left(\sum_{i=1}^{n} M_1^T(z_i)^2/n\right)\|\hat{\gamma}-\gamma_0\|^2 = O_p(1)o_p(1) = o_p(1),$$

where $\tilde{\gamma}$ is the mean-value, which actually differs from element to element of $T(x,\gamma)$. By H, $E[\|D_i\|^2]$ is finite and by a standard application of the i.i.d. uniform law of large numbers and Ω nonsingular, $\hat{\Omega}^{-1} = \Omega^{-1} + o_p(1)$. Then by eq. (A.3) and M,

$$\text{(A.4)}\qquad \sum_{i=1}^{n} \|\hat{A}_i - A_i\|^2/n \leqslant \sum_{i=1}^{n} C\Big(\big\|(\hat{T}_i - T_i)\hat{\Omega}^{-1}\big\|^2 + \big\|(\hat{G}_i - G_i)\hat{\Omega}^{-1}\big\|^2$$

$$+ \big\|D_i(\hat{\Omega}^{-1} - \Omega^{-1})\big\|^2\Big)/n$$

$$\leqslant C\|\hat{\Omega}^{-1}\|^2 \sum_{i=1}^{n} \left(\|\hat{T}_i - T_i\|^2 + \|\hat{G}_i - G_i\|^2\right)/n$$

$$+ C\|\hat{\Omega}^{-1} - \Omega^{-1}\|^2 \sum_{i=1}^{n} \|D_i\|^2/n$$

$$= O_p(1)o_p(1) + o_p(1)O_p(1) = o_p(1).$$

Also, note that $\max_N \|A_i\Delta_i(\beta)\| \leqslant \|D_i\|\,\|\Omega^{-1}\|M_1(z_i)$, so that by the i.i.d. uniform law of large numbers $\sum_{i=1}^{n} A_i\Delta_i(\bar{\beta})/n = (\Lambda^*)^{-1} + o_p(1)$. It then follows from equations (A.2) and (A.4) that

$$\text{(A.5)}\qquad \sum_{i=1}^{n} \hat{A}_i\Delta_i(\bar{\beta})/n = (\Lambda^*)^{-1} + o_p(1).$$

Next, note that $\sum_{i=1}^{n}\|A_i\|^2/n = O_p(1)$ by M. Also, for matrices B_1 and B_2, by T and CS $\|B_1'B_1 - B_2'B_2\| = \|(B_1-B_2)'(B_1-B_2) + B_2'(B_1-B_2) + (B_1-B_2)'B_2\| \leqslant \|B_1-B_2\|^2 + 2\,\|B_1-B_2\|\,\|B_2\|$. By $\operatorname{plim}(\hat{\Omega}) = \Omega$ and Ω positive definite there exists $\hat{F}$ such that $\hat{F}\hat{F}' = \hat{\Omega}$ with probability approaching one and $\hat{F} = O_p(1)$. Let $\hat{R}_i = \hat{A}_i\hat{F}$, and $\tilde{R}_i = A_i\hat{F}$. Then by equation (A.4),

$$\text{(A.6)}\qquad \left\|\sum_{i=1}^{n}\left(\hat{A}_i\hat{\Omega}\hat{A}_i' - A_i\hat{\Omega}A_i'\right)/n\right\| \leqslant \sum_{i=1}^{n} \|\hat{R}_i'\hat{R}_i - \tilde{R}_i'\tilde{R}_i\|/n$$

$$\leqslant \sum_{i=1}^{n} \|\hat{R}_i - \tilde{R}_i\|^2/n + 2\sum_{i=1}^{n} \|\hat{R}_i - \tilde{R}_i\|\,\|\tilde{R}_i\|/n$$

$$\leqslant \|\hat{F}\|^2 \sum_{i=1}^{n} \|\hat{A}_i - A_i\|^2/n + 2\|\hat{F}\|^2\left(\sum_{i=1}^{n}\|\hat{A}_i - A_i\|^2/n\right)^{1/2}$$

$$\times\left(\sum_{i=1}^{n}\|A_i\|^2/n\right)^{1/2}$$

$$= O_p(1)o_p(1) + O_p(1)\left(o_p(1)O_p(1)\right)^{1/2} = o_p(1).$$

Also,

$$\left\|\sum_{i=1}^{n}\left(A_i\hat{\Omega}A_i' - A_i\Omega A_i'\right)/n\right\| \leqslant \sum_{i=1}^{n}\left\|A_i(\hat{\Omega}-\Omega)A_i'\right\|/n \leqslant \|\hat{\Omega}-\Omega\|\sum_{i=1}^{n}\|A_i\|^2/n$$

$$= o_p(1)O_p(1) = o_p(1),$$

and by the law of large numbers $\Sigma_{i=1}^{n} A_i \Omega A_i'/n = (\Lambda^*)^{-1} + o_p(1)$. The second conclusion then follows from equation (A.6) and T.

Next, by a second-order mean value expansion note that for an element $t(x,\gamma)$ of $T(x,\gamma)$, with probability approaching one for all i,

$$t(x_i,\hat{\gamma}) = t(x_i,\gamma_0) + \partial t(x_i,\gamma_0)/\partial\gamma'(\hat{\gamma}-\gamma) + r_{in}, \qquad |r_{in}| \leqslant C M_2^T(z_i)\|\hat{\gamma}-\gamma_0\|^2.$$

Let $\tilde{\rho}_i = \rho(z_i,\beta_0)$ and ρ_i be an element of $\tilde{\rho}_i$. By $E[\rho_i|x_i] = 0$ and Assumption 3.4, $E[\rho_i \partial t(x_i,\gamma_0)/\partial\gamma]$ exists and equals zero, so that by the law of large numbers, $\Sigma_{i=1}^{n}\rho_i\partial t(x_i,\gamma_0)/\partial\gamma/n = o_p(1)$. Also, $E[M_2^T(z_i)\|\rho_i\|] < \infty$ by Assumption 3.4 (d) and H. Then by CS, M and $\hat{\gamma} = \gamma_0 + O_p(1/\sqrt{n})$,

$$\text{(A.7)} \qquad \left|\sum_{i=1}^{n}[t(x_i,\hat{\gamma}) - t(x_i,\gamma_0)]\rho_i/\sqrt{n}\right|$$
$$\leqslant \left\|\sum_{i=1}^{n}\rho_i\partial t(x_i,\gamma_0)/\partial\gamma/n\right\|\sqrt{n}\left\|\hat{\gamma}-\gamma_0\right\| + C\left(\sum_{i=1}^{n} M_2^T(z_i)|\rho_i|/n\right)\sqrt{n}\|\hat{\gamma}-\gamma_0\|^2$$
$$= o_p(1)O_p(1) + O_p(1)O_p\left(1/\sqrt{n}\right) = o_p(1).$$

Since this equation applies to each element of $\tilde{\rho}_i$ and $T(x,\gamma)$, it follows that $\|\Sigma_{i=1}^{n}(\hat{T}_i - T_i)\otimes \rho_i/\sqrt{n}\| = o_p(1)$. Then $\|\Sigma_{i=1}^{n}(\hat{D}_i - D_i)\otimes\rho_i/\sqrt{n}\| = o_p(1)$ by equation (A.1b) and T. Then note that for the lth column B_l of a matrix B,

$$\text{(A.8)} \qquad \left|\sum_{i=1}^{n}(\hat{D}_{il} - D_{il})'\hat{\Omega}^{-1}\tilde{\rho}_i/\sqrt{n}\right| = \left|\operatorname{tr}\left[\hat{\Omega}^{-1}\sum_{i=1}^{n}\tilde{\rho}_i(\hat{D}_{il}-D_{il})'/\sqrt{n}\right]\right|$$
$$\leqslant C\left\|\hat{\Omega}^{-1}\sum_{i=1}^{n}\tilde{\rho}_i(\hat{D}_{il}-D_{il})'/\sqrt{n}\right\|$$
$$\leqslant C\|\hat{\Omega}^{-1}\|\left\|\sum_{i=1}^{n}\tilde{\rho}_i(\hat{D}_{il}-D_{il})'/\sqrt{n}\right\|$$
$$\leqslant O_p(1)\left\|\sum_{i=1}^{n}(\hat{D}_i - D_i)\otimes\tilde{\rho}_i/\sqrt{n}\right\| = O_p(1)o_p(1) = o_p(1).$$

Also, $E[D_i\otimes\tilde{\rho}_i] = 0$ so that by independence and H, $E[\|\Sigma_{i=1}^{n}D_i\otimes\tilde{\rho}_i/\sqrt{n}\|^2] = E[\|\tilde{\rho}_i\|^2\|D_i\|^2] < \infty$. It follows from M that

$$\text{(A.9)} \qquad \left|\sum_{i=1}^{n}D_{il}'(\hat{\Omega}^{-1}-\Omega^{-1})\tilde{\rho}_i/\sqrt{n}\right| = \left|\operatorname{tr}\left[(\hat{\Omega}^{-1}-\Omega^{-1})\sum_{i=1}^{n}\tilde{\rho}_i D_{il}'/\sqrt{n}\right]\right|$$
$$\leqslant C\|\hat{\Omega}^{-1}-\Omega^{-1}\|\left\|\sum_{i=1}^{n}\tilde{\rho}_i D_{il}'/\sqrt{n}\right\|$$
$$\leqslant o_p(1)\left\|\sum_{i=1}^{n}D_{il}\otimes\tilde{\rho}_i/\sqrt{n}\right\| = o_p(1).$$

Since this is true for each l, it follows by the triangle inequality that

$$\text{(A.10)} \qquad \left\|\sum_{i=1}^{n}(\hat{A}_i - A_i)'\tilde{\rho}_i/\sqrt{n}\right\| = o_p(1).$$

The first conclusion now follows from equations (A.5) and (A.10) by standard mean value expansion and central limit arguments, which for brevity are omitted. *Q.E.D.*

It is straightforward to formulate a result that allows data-based choice of the estimate of the optimal instruments from a finite set of sequences.

LEMMA A.2: *If* $\hat{G}_i = \hat{G}_i(\hat{l})$ *for* $\hat{l} \in \mathscr{L}$, $\mathscr{L}$ *is finite, and for each* $l \in \mathscr{L}$, $\hat{G}_i(l)$ *satisfies equation* (A.1) *with* $\hat{G}_i(l)$ *replacing* $\hat{G}_i$, *then the conclusion of Lemma A*.1 *holds.*

PROOF: The conclusion follows immediately from Lemma A.1 upon noting that

$$\sum_{i=1}^{n} \|\hat{G}_i - G_i\|^2/n \leqslant \max_{\mathscr{L}} \sum_{i=1}^{n} \left\|\hat{G}_i(l) - G_i\right\|^2/n,$$

$$\left\|\sum_{i=1}^{n} \left(\hat{G}_i - G_i\right) \otimes \tilde{\rho}_i/\sqrt{n}\right\| \leqslant \max_{\mathscr{L}} \left\|\sum_{i=1}^{n} \left(\hat{G}_i(l) - G_\iota\right) \otimes \tilde{\rho}_\iota/\sqrt{n}\right\|.$$

Let $h(z,\gamma)$ be a function of z and a parameter vector γ and let $\hat{\gamma}$ be a consistent estimate of some value γ_0. The following assumption concerning $h(z,\gamma)$ and γ will be maintained for Lemmas A.3–A.7:

ASSUMPTION A.1: (i) $\hat{\gamma} - \gamma_0 = O_p(1/\sqrt{n})$; (ii) *for* $\nu > 2$, $E[|h(z_i,\gamma_0)|^\nu]$ *is finite*; (iii) $h(z_\iota,\gamma)$ *is continuously differentiable and on a neighborhood* N *of* γ_0; (iv) $sup_{\gamma \in N} \|\partial h(z_i,\gamma)/\partial\gamma\| \leqslant M_1^h(z_\iota)$ *and* $E[M_1^h(z_\iota)^\nu] < \infty$.

The following Assumption will be used for the nearest neighbor case.

ASSUMPTION A.2: $k/\sqrt{n} \to \infty$, $k/n \to 0$.

The following lemmas will be useful in proving Theorem 3.1. Let

$$h_i = h(z_\iota,\gamma_0), \qquad \hat{h}_i = h(z_\iota,\hat{\gamma}), \qquad g_i = E[h_i|x_i],$$
$$\bar{g}_i = \sum_{j=1}^{n} W_{ij} g_j, \qquad \tilde{g}_i = \sum_{j=1}^{n} W_{ij} h_j, \qquad \hat{g}_i = \sum_{j=1}^{n} W_{ij}\hat{h}_j. \tag{A.11}$$

LEMMA A.3: Stone (1977, Proposition 1)): $lim_{n\to\infty} E[|\bar{g}_i - g_i|^\nu] = 0$.

The proofs of the following three Lemmas are nearly identical to the proofs of Lemmas 8, 9, and 5, respectively, of Robinson (1987), and so will be omitted.

LEMMA A.4: $\{E[|\tilde{g}_i - \bar{g}_i|^\nu]\}^{1/\nu} = O(k^{-1/2})$.

LEMMA A.5: $max_{i \leqslant n} |\tilde{g}_i - \bar{g}_i| = O_p(n^{1/\nu}k^{-1/2})$.

LEMMA A.6: $max_{i \leqslant n} |\hat{g}_i - \tilde{g}_i| = O_p(k^{-1/2})$.

Let $X_n = (x_1,\ldots,x_n)$ and $Z_{-i} = (z_1,\ldots,z_{i-1},z_{i+1},z_n)$.

LEMMA A.7: *Suppose that* (*i*) $\rho_n(z,\mathscr{X})$ *is a function such that* $E[\rho_n(z_i,X_n)|x_i,Z_{-i}] = 0$ *and* $E[|\rho_n(z_i,X_n)|^{2\nu/(\nu-2)}] = O(1)$; (*ii*) $h(z_i,\gamma)$ *is twice continuously differentiable in* N; (*iii*) $sup_{\gamma\in N} \|\partial^2 h(z_i,\gamma)/\partial\gamma^2\| \leqslant M_2^h(z_i)$ *and* $E[M_2^h(z_i)^{2\nu/(\nu+2)}] < \infty$. *Then*

$$\sum_{i=1}^{n} (\hat{g}_i - g_i)\rho_n(z_i,X_n)/\sqrt{n} = o_p(1). \tag{A.12}$$

PROOF: Let $\rho_{in} = \rho_n(z_i, X_n)$. By the triangle inequality

$$\left| \sum_{i=1}^{n} (\hat{g}_i - g_i)\rho_{in}/\sqrt{n} \right| \leqslant T_1 + T_2 + T_3, \tag{A.13}$$

$$T_1 = \left| \sum_{i=1}^{n} (\bar{g}_i - g_i)\rho_{in}/\sqrt{n} \right|, \quad T_2 = \left| \sum_{i=1}^{n} (\tilde{g}_i - \bar{g}_i)\rho_{in}/\sqrt{n} \right|, \quad T_3 = \left| \sum_{i=1}^{n} (\hat{g}_i - \tilde{g}_i)\rho_{in}/\sqrt{n} \right|.$$

Note that $E[|\tilde{g}_i|^\nu]$ and $E[|\bar{g}_i|^\nu]$ are finite by $E[|h_i|^\nu]$ finite. Also, by (i) and Holder's inequality all fourth order cross moments of g_i, $\bar{g}_i$, and ρ_{in} exist. Therefore, by $\bar{g}_i$ and g_i functions of only X_n, ρ_{in} uncorrelated with ρ_{sn} conditional on X_n for $s \neq i$, ρ_{in} and $\bar{g}_i - g_i$ identically distributed, Holders inequality, and Lemma A.3, it follows that for $\tau = \nu/(\nu - 2)$,

$$\begin{aligned} E\left(T_1^2\right) &= E\left[\left\{ \sum_{i=1}^{n} (\bar{g}_i - g_i)\rho_{in}/\sqrt{n} \right\}^2\right] = E\left[|\bar{g}_i - g_i|^2 \rho_{in}^2\right] \\ &= \{E[|\bar{g}_i - g_i|^\nu]\}^{2/\nu}\left\{E\left[|\rho_{in}|^{2\tau}\right]\right\}^{1/\tau} = o(1)O(1) = o(1). \end{aligned} \tag{A.14}$$

Thus, by Markov's inequality, $T_1 = o_p(1)$. Next, note that for $i \neq j$,

$$E\left[T_2^2\right] = E\left[|\tilde{g}_i - \bar{g}_i|^2 \rho_{in}^2\right] + (n-1) \circ E\left[(\tilde{g}_i - \bar{g}_i)\rho_{in}\rho_{jn}(\tilde{g}_j - \bar{g}_j)\right]. \tag{A.15}$$

By Lemmas A.1 and A.2 it follows similarly to equation (A.14) that the first term following the equality is $o(1)$. For define $\tilde{g}_{\iota;j} = \tilde{g}_i - W_{ij}h_j$, and note that, conditional on X_n, $\tilde{g}_{i;j}$ is independent of z_i and z_j. Also, by $W_{jj} = 0$, $\tilde{g}_j$ is independent of z_j conditional on X_n. Then

$$\begin{aligned} E\left[\tilde{g}_{i;j}\rho_{in}\rho_{jn}\tilde{g}_j\right] &= E\left[E\left[\tilde{g}_{i;j}\rho_{in}\rho_{jn}\tilde{g}_j | Z_{-j}, x_j\right]\right] \\ &= E\left[\tilde{g}_{i;j}\rho_{in}\tilde{g}_j \circ E\left[\rho_{jn} | Z_{-j}, x_j\right]\right] = 0, \qquad (i \neq j). \end{aligned} \tag{A.16}$$

It follows similarly that

$$0 = E\left[\bar{g}_i\rho_{in}\rho_{jn}\tilde{g}_j\right] = E\left[\tilde{g}_{i;j}\rho_{in}\rho_{jn}\tilde{g}_{j;i}\right] = E\left[\tilde{g}_i\rho_{in}\rho_{jn}\bar{g}_j\right], \qquad (i \neq j). \tag{A.17}$$

Also, the same results hold with j and i interchanged, so that by Holder's inequality, for $\tau = \nu/(\nu - 2)$,

$$\begin{aligned} \left|E\left[(\tilde{g}_i - \bar{g}_i)\rho_{in}\rho_{jn}(\tilde{g}_j - \bar{g}_j)\right]\right| &= \left|E\left[\tilde{g}_i\rho_{in}\rho_{jn}\tilde{g}_j\right]\right| \\ &= \left|E\left[(\tilde{g}_i - \tilde{g}_{i;j})\rho_{in}\rho_{jn}(\tilde{g}_j - \tilde{g}_{j;\iota})\right]\right| \\ &\leqslant E\left[W_{ij}W_{ji}|h_i|\,|h_j|\,|\rho_{in}|\,|\rho_{jn}|\right] \\ &\leqslant (W_0/k)^2\{E[|h_i|^\nu]\}^{2/\nu}\left\{E\left[|\rho_{in}|^{2\tau}\right]\right\}^{1/\tau} \leqslant C/k^2. \end{aligned} \tag{A.18}$$

It then follows by $n/k^2 = o(1)$ that the second term following the equality in eq. (A.15) is $o(1)$, so that $T_2 = o_p(1)$ by Markov's inequality.

Next, note that by a mean value expansion,

$$\hat{g}_i - \tilde{g}_i = \mathring{g}_{i\gamma}(\hat{\gamma} - \gamma_0) = \tilde{g}_{i\gamma}(\hat{\gamma} - \gamma_0) + r_{in} \tag{A.19}$$

where $\mathring{\gamma}_i$ are mean values satisfying $\max_{\iota \leqslant n} |\mathring{\gamma}_\iota - \gamma_0| \leqslant |\hat{\gamma} - \gamma_0|$,

$$\mathring{g}_{i\gamma} = \sum_{j=1}^{n} W_{ij}h_\gamma(z_j, \mathring{\gamma}_i), \qquad \tilde{g}_{i\gamma} = \sum_{j=1}^{n} W_{ij}h_\gamma(z_j, \gamma_0),$$

$$|r_{in}| = \left|(\mathring{g}_{i\gamma} - g_{i\gamma})'(\hat{\gamma} - \gamma_0)\right| \leqslant \|\mathring{g}_{i\gamma} - g_{i\gamma}\|\,\|\hat{\gamma} - \gamma_0\| \leqslant \left[\sum_{j=1}^{n} W_{ij}M_2^h(z_j)\right]\|\hat{\gamma} - \gamma_0\|^2,$$

and the last equality holds for every i with probability approaching one by $\|\mathring{\gamma}_i - \gamma_0\| \leqslant \|\hat{\gamma} - \gamma_0\|$ and

$\operatorname{plim}(\hat{\gamma}) = \gamma_0$. Therefore,

$$T_3 \leqslant T_4\sqrt{n}\,\|\hat{\gamma} - \gamma_0\| + T_5\sqrt{n}\,\|\hat{\gamma} - \gamma_0\|^2 + o_p(1), \tag{A.20}$$

$$T_4 = \left\| \sum_{i=1}^{n} \tilde{g}_{i\gamma}\rho_{in}/n \right\|, \qquad T_5 = \sum_{i=1}^{n} \left[\sum_{j=1}^{n} W_{ij} M_2^h(z_j) \right] |\rho_{in}|/n.$$

By Lemma A.3 applied to $M_2^h(z_i)$ and the Holder inequality, for $\tau = 1/\{1 - [1/\{2\nu/(\nu-2)\}]\} = 2\nu/(\nu+2)$,

$$E(T_5) \leqslant \left\{ E\left[\left| \sum_{j=1}^{n} W_{ij} M_2^h(z_j) \right|^{\tau} \right] \right\}^{1/\tau} \left\{ E\left[\rho_{in}|^{2\nu/(\nu-2)} \right] \right\}^{(\nu-2)/2\nu} = O(1), \tag{A.21}$$

so that $T_5 = O_p(1)$. Also, for $g_{\iota\gamma} = E[h_\gamma(z_i, \gamma_0)|x_i]$, note that $\operatorname{plim}(\sum_{i=1}^{n} g_{i\gamma}\rho_{in}/n) = 0$ by Chebyshev's law of large numbers, so by the triangle, Holder, and Markov inequalities, and Lemmas A.3 and A.4 applied to $h_\gamma(z_i, \gamma_0)$, it follows that for $\tau = 2\nu/(\nu - 2)$,

$$\begin{aligned} T_4 &\leqslant \left\| \sum_{i=1}^{n} (\tilde{g}_{\iota\gamma} - g_{i\gamma})\rho_{in}/n \right\| + \left\| \sum_{i=1}^{n} g_{i\gamma}\rho_{in}/n \right\| \\ &= O_p\left(E\left[\|\tilde{g}_{i\gamma} - g_{i\gamma}\| |\rho_{in}| \right] \right) + o_p(1) \\ &= O_p\left(\left\{ E\left[\|\tilde{g}_{i\gamma} - g_{i\gamma}\|^{2\nu/(\nu+2)} \right] \right\}^{(\nu+2)/2\nu} \{ E[|\rho_{in}|^{\tau}] \}^{1/\tau} \right) + o_p(1) = o_p(1). \end{aligned} \tag{A.22}$$

The conclusion then follows by equation (A.20) and $\sqrt{n}\,\|\hat{\gamma} - \gamma_0\|^2 = o_p(1)$. *Q.E.D.*

Proof of Theorem 3.1: By Lemma A.2 it suffices to show that equation (A.1) is satisfied for a particular sequence $k = k_l(n)$ as in Assumption 3.2. For a nonzero element $g(x)$ of $G(x)$ let γ include β and $h(z,\gamma) = \rho_\beta(z,\beta) - t(x,\gamma)$ for corresponding elements of $\partial\rho(z,\beta)/\partial\beta$ and $T(x,\gamma)$. It follows by Assumption 3.4 and H that Assumption A.1 is satisfied for $\nu = 2$. Then by Lemmas A.3, A.4, and A.6, T, and M,

$$\begin{aligned} \sum_{i=1}^{n} |\hat{g}_i - g_\iota|^2/n &\leqslant C \sum_{i=1}^{n} \left(|\hat{g}_i - \tilde{g}_\iota|^2 + |\tilde{g}_i - \bar{g}_i|^2 + |\bar{g}_i - g_i|^2 \right)/n \\ &\leqslant C \max_{i \leqslant n} |\hat{g}_i - \tilde{g}_i|^2 + O_p\left(E\left[|\tilde{g}_i - \bar{g}_i|^2 \right] \right) + O_p\left(E\left[|\bar{g}_i - g_i|^2 \right] \right) \\ &= O_p(1/k) + O_p(O(1/k)) + O_p(o(1)) = o_p(1). \end{aligned} \tag{A.23}$$

Equation (A.1a) follows since this result holds for each element of $G(x)$.

Next, note that Assumption A.1 is satisfied for ν as in Assumptions 3.3 and 3.4. Let $\rho_n(z, \mathscr{X})$ be an element of $\rho(z_i, \beta_0)$, and note that by Assumption 3.4 the hypotheses of Lemma A.7 are satisfied. Then by the conclusion of Lemma A.7, the corresponding element of $\sum_{i=1}^{n}(\hat{G}_i - G_i) \otimes \tilde{\rho}_i/\sqrt{n}$ is $o_p(1)$. Equation (A.1b) then follows since this is true for each element of $G(x)$ and $\rho(z_i, \beta_0)$. The conclusion of the Theorem then follows by Lemma A.2. *Q.E.D.*

The following lemmas will be useful in proving Theorem 4.1. Let

$$\begin{aligned} &h = (h_1, \ldots, h_n)', \qquad \hat{h} = (\hat{h}_1, \ldots, \hat{h}_n)', \qquad g = (g_1, \ldots, g_n)', \\ &P_i^K = (p_1(x_\iota), \ldots, p_K(x_i))', \qquad P^K = [P_1^K, \ldots, P_n^K]', \qquad P = P^{\hat{K}}, \\ &\tilde{\eta}_K \in \operatorname*{argmin}_{\eta} E\left[\left(g_i - P_i^{K\prime}\eta \right)^2 \right], \qquad \tilde{g}_K = P\tilde{\eta}_K, \qquad \tilde{g} = \tilde{g}_{\hat{K}}. \end{aligned} \tag{A.24}$$

Assumption A.1 and the following assumption will be imposed for Lemmas A.8–A.13.

Assumption A.3: (i) $\hat{K} \xrightarrow{p} \infty$, *and* $\hat{K} = o_p(b)$ *for* $b \to \infty$; (ii) *there exists* $r \geqslant 0$ *and* η_K, $(K = 1, 2, \ldots)$, *such that* $\lim_{K\to\infty} K^{\zeta}\{E[(g_i - P_i^{K\prime}\eta_K)^2]\}^{1/2} = 0$; (iii) *either* $\zeta > 1$, *or* $\zeta = 0$ *and there exists* $\mathscr{K}(n)$ *such that the number of elements of* $\mathscr{K}(n)$ *is bounded and* $Prob(\hat{K} \in \mathscr{K}(n)) \to 1$; (iv) ρ_i $(i = 1, \ldots, n)$ *are i.i.d. random variables satisfying* $E[\rho_i|x_i] = 0$, $E[\rho_i^2|x_i]$ *is bounded, and* $E[|\rho_i|^{2\nu/(\nu-2)}] < \infty$; (iv) $E[|p_K(x_i)|^{\nu}] < \infty$ $(K = 1, 2, \ldots)$.

Henceforth let $\rho = (\rho_1, \ldots, \rho_n)'$.

LEMMA A.8: *There exists nonrandom* K_u *such that* $\hat{K} \leqslant K_u$ *with probability approaching one and* $K_u = o(b)$.

PROOF: By Assumption A.3 (i), $\hat{K}/b = o_p(1)$, so that there exists $\delta \to 0$ such that $\text{Prob}(\hat{K}/b \leqslant \delta) \to 1$. Take $K_u = [\delta b] + 1$, where $[\circ]$ denotes the greatest integer less than. Then $K_u/b = [\delta b]/b + o(1) \leqslant \delta + o(1) = o(1)$. *Q.E.D.*

LEMMA A.9: $\|\hat{h} - h\| = O_p(1)$.

PROOF: By consistency of $\hat{\gamma}$ and Assumption A.1 (iii) and (iv), it follows that with probability approaching one,

$$\|\hat{h} - h\|^2 = \sum_{i=1}^{n} (\hat{h}_i - h_i)^2 \leqslant \sum_{i=1}^{n} M_1^h(z_i)^2 \|\hat{\gamma} - \gamma_0\|^2 = \left(\sum_{i=1}^{n} M_1^h(z_i)^2/n\right)(n\|\hat{\gamma} - \gamma_0\|^2).$$

By M, $\Sigma_{i=1}^n M_1^h(z_i)^2/n = O_p(1)$, so that the conclusion follows from Assumption A.1 (i). *Q.E.D.*

Let $Q = P(P'P)^{-}P'$ denote the matrix of the orthogonal projection onto the space spanned by the columns of P.

LEMMA A.10: $\|Q(h - g)\|^2 = o_p(n^{2/\nu} \circ b)$. *Also, if* $E[h_i^2|x_i]$ *is bounded, then* $\|Q(h - g)\|^2 = o_p(b)$.

PROOF: Take K_u as in the conclusion of Lemma A.8, and let

$$\bar{p}_i = (p_1(x_i), \ldots, p_{K_u}(x_i))', \qquad \bar{P} = [\bar{p}_1, \ldots \bar{p}_n]', \qquad \bar{Q} = \bar{P}(\bar{P}'\bar{P})^{-}\bar{P}'. \tag{A.25}$$

Also, let 1_u be the indicator function for the even $\hat{K} \leqslant K_u$, and note that for $1_u = 1$, $\bar{Q} - Q$ is positive semi-definite. Then for $e = h - g$,

$$1_u\|Qe\|^2 \leqslant 1_u\|\bar{Q}e\|^2 \leqslant \|\bar{Q}e\|^2. \tag{A.26}$$

By Assumption A.1 (ii), and the conditional version of H, $E[|y_i|^{\nu/2}]$ is finite for $y_i = \text{Var}(h_i|x_i)$. Then by M

$$\text{Prob}\left(\max_{i\leqslant n} y_i \geqslant n^{2/\nu}C\right) \leqslant n \circ \text{Prob}\left(|y_i| \geqslant n^{2/\nu}C\right) \leqslant E\left[|y_i|^{\nu/2}\right]/C^{\nu/2}. \tag{A.27}$$

Thus, $\max_{i\leqslant n}\text{Var}(h_i|x_i) = O_p(n^{2/\nu})$. By independence of the observations, the components of e are mutually independent conditional on $X_n = (x_1, \ldots, x_n)$, implying $E[e|X] = 0$ and $\text{Var}(h|X)$ is a diagonal matrix with ith diagonal element $\text{Var}(h_i|x_i)$. Also, Q is idempotent, so that its elements are bounded. Thus,

$$\begin{aligned} E\left[\|\bar{Q}e\|^2|X\right] &= E\left[e'\bar{Q}e|X\right] = \sum_{i,j=1}^{n} \bar{Q}_{ij}E\left[e_ie_j|X\right] \\ &= \sum_{i=1}^{n} \bar{Q}_{ii}\,\text{Var}\,(y_i|x_i) \leqslant \left[\max_{i\leqslant n} \text{Var}\,(y_i|x_i)\right] \text{tr}\,(\bar{Q}) \\ &= O_p(n^{2/\nu})\,\text{rank}\,(\bar{Q}) \leqslant O_p(n^{2/\nu})\,\text{rank}\,(\bar{P}) \leqslant O_p(n^{2/\nu})K_u = o_p(n^{2/\nu} \circ b). \end{aligned} \tag{A.28}$$

It follows by the conditional Markov inequality and bounded convergence that $\|\bar{Q}e\|^2 = o_p(n^{2/\nu} \circ b)$.

Also, since $1 - 1_u = 0$ with probability approaching one $(1 - 1_u)\|\bar{Q}e\|^2 = o_p(n^{2/\nu} \circ b)$. It then follows from equation (A.26) that

$$\|Qe\|^2 = (1 - 1_u)\|Qe\|^2 + 1_u\|Qe\|^2 \leqslant o_p(n^{2/\nu} \circ b) + \|\bar{Q}e\|^2 = o_p(n^{2/\nu} \circ b),$$

giving the first conclusion. The second conclusion follows from the first inequality in equation (A.28) and $\operatorname{tr}(\bar{Q}) = o(b)$. *Q.E.D.*

Lemma A.11: $\|\tilde{g} - g\| = o_p(\sqrt{n})$, $|\rho'(\tilde{g} - g)| = o_p(\sqrt{n})$, $|\rho' Q(\tilde{g} - g)| = o_p(\sqrt{n})$.

Proof: Let K_u be as in the conclusion to Lemma A.8, and let K_l be such that $\operatorname{Prob}(\hat{K} \geqslant K_l) \to 1$ and $K_l \to \infty$. If the number of possible values for $\hat{K}$ is uniformly bounded, take $\mathscr{K} = \mathscr{K}(n)$ as in Assumption 4.1; otherwise, let $\mathscr{K} = \mathscr{K}(n) \equiv \{K_l, K_{l+1}, \ldots, K_u\}$. Note $\operatorname{Prob}(\hat{K} \in \mathscr{K}) \to 1$. Then, for $1_{\mathscr{K}} = 1(\hat{K} \in \mathscr{K})$,

$$E[1_{\mathscr{K}}\|\tilde{g} - g\|] \leqslant E\Big[1_{\mathscr{K}} \circ \max_{\mathscr{K}}\|\tilde{g}_K - g\|\Big] \leqslant \sum_{\mathscr{K}} E[\|\tilde{g}_K - g\|] \leqslant \sum_{\mathscr{K}} \big(E[\|\tilde{g}_K - g\|^2]\big)^{1/2}$$
$$= \sqrt{n} \sum_{\mathscr{K}} \Big(E\big[(\tilde{g}_{Ki} - g_i)^2\big]\Big)^{1/2}. \tag{A.29}$$

Also by Assumption A.3 (iii) and (iv) and H $E[|(\tilde{g}_{Ki} - g_i)(\tilde{g}_{Kj} - g_j)\rho_i\rho_j|]$ is finite for all i, j, K, so that it follows similarly to the proof of Lemma A.10 that

$$E\Big[|\rho'(\tilde{g}_K - g)|^2|X\Big] = (\tilde{g}_K - g)'E[\rho\rho'|X](\tilde{g}_K - g)$$
$$\leqslant \max_{i \leqslant n} E\big[\rho_i^2|x_i\big]\|\tilde{g}_K - g\|^2 \leqslant C\|\tilde{g}_K - g\|^2.$$

Let $P^K = [P^K(x_1) \ldots, P^K(x_n)]'$ and $Q_K = P^K(P^{K\prime}P^K)^- P^{K\prime}$. Since, by Q_K idempotent, the elements of Q_K are bounded, it also follows that

$$E\Big[|\rho' Q_K(\tilde{g}_K - g)|^2|X\Big] = (\tilde{g}_K - g)'Q_K E[\rho\rho'|X]Q_K(\tilde{g}_K - g)$$
$$\leqslant \max_{i \leqslant n} E\big[\rho_i^2|x_i\big]\|Q_K(\tilde{g}_K - g)\|^2 \leqslant C\|\tilde{g}_K - g\|^2.$$

Thus,

$$E\big[1_{\mathscr{K}}|\rho'(\tilde{g} - g)|\big] \leqslant E\Big[1_{\mathscr{K}} \circ \max_{\mathscr{K}}|\rho'(\tilde{g}_K - g)|\Big] \leqslant \sum_{\mathscr{K}} E\big[|\rho'(\tilde{g}_K - g)|\big]$$
$$\leqslant \sum_{\mathscr{K}} \Big(E\big[|\rho'(\tilde{g}_K - g)|^2\big]\Big)^{1/2} = \sum_{\mathscr{K}} \Big(E\Big[E\big[|\rho'(\tilde{g}_K - g)|^2|X\big]\Big]\Big)^{1/2}$$
$$\leqslant C\sum_{\mathscr{K}} E\big[\|\tilde{g}_K - g\|^2\big] = C\sqrt{n} \sum_{\mathscr{K}} \Big(E\big[(\tilde{g}_{Ki} - g_i)^2\big]\Big)^{1/2}, \tag{A.30}$$

$$E\big[1_{\mathscr{K}}|\rho' Q(\tilde{g} - g)|\big] \leqslant \sum_{\mathscr{K}} \Big(E\big[|\rho' Q_K(\tilde{g}_K - g)|^2\big]\Big)^{1/2}$$
$$= \sum_{\mathscr{K}} \Big(E\Big[E\big[|\rho' Q_K(\tilde{g}_K - g)|^2|X\big]\Big]\Big)^{1/2}$$
$$\leqslant C\sqrt{n} \sum_{\mathscr{K}} \Big(E\big[(\tilde{g}_{Ki} - g_i)^2\big]\Big)^{1/2}. \tag{A.31}$$

Note that $L \geqslant K$ implies that $E[(\tilde{g}_{Li} - g_i)^2] \leqslant E[(\tilde{g}_{Ki} - g_i)^2]$, so that for the case where the number of elements of $\mathscr{K}$ is bounded,

$$\sum_{\mathscr{K}}\left(E\left[(\tilde{g}_{Ki}-g_i)^2\right]\right)^{1/2} \leqslant C\max_{\mathscr{K}}\left(E\left[(\tilde{g}_{K\iota}-g_i)^2\right]\right)^{1/2}$$

$$\leqslant C\left(E\left[(\tilde{g}_{K_u i}-g_i)^2\right]\right)^{1/2}=o(1).$$

For the other case, note that $\Sigma_{K=1}^{\infty}(E[(\tilde{g}_{Ki}-g_i)^2])^{1/2} \leqslant C\Sigma_{K=1}^{\infty}K^{-\zeta}<\infty$, so that

$$\sum_{\mathscr{K}}\left(E\left[(\tilde{g}_{Ki}-g_i)^2\right]\right)^{1/2} \leqslant \sum_{K=K_l}^{\infty}\left(E\left[(\tilde{g}_{K\iota}-g_i)^2\right]\right)^{1/2}=o(1).$$

It then follows from equations (A.30) and (A.31) and M that

$$1_{\mathscr{K}}\|\tilde{g}-g\|=o_p(\sqrt{n}), \qquad 1_{\mathscr{K}}|\rho'(\tilde{g}-g)|=o_p(\sqrt{n}), \qquad 1_{\mathscr{K}}|\rho'Q(\tilde{g}-g)|=o_p(\sqrt{n}).$$

The conclusion then follows from the fact that 1-$1_{\mathscr{K}}$= 0 with probability approaching one. *Q.E.D.*

Lemma A.12: $\|Q\hat{h}-g\|=o_p(n^{1/\nu}b^{1/2})+o_p(\sqrt{n})$.

Proof: By Q idempotent, $\tilde{g}=Q\tilde{g}$, and by Lemmas A.9, A.10, and A.11,

$$\|Q\hat{h}-g\| \leqslant \|Q(\hat{h}-h)\|+\|Q(h-g)\|+\|Q(g-\tilde{g})\|+\|\tilde{g}-g\|$$

$$\leqslant \|\hat{h}-h\|+o_p(n^{1/\nu}b^{1/2})+2\|\tilde{g}-g\|=O_p(1)+o_p(n^{1/\nu}b^{1/2})+o_p(\sqrt{n})$$

$$=o_p(n^{1/\nu}b^{1/2})+o_p(\sqrt{n}). \qquad Q.E.D.$$

Lemma A.13: $|\rho'(Q\hat{h}-g)|=o_p(n^{1/\nu}b)+o_p(\sqrt{n})$.

By Lemma A.10 and $\mathrm{Var}(\rho_i|x_i)$ bounded, $\|Q\rho\|=o_p(b^{1/2})$. Then by Q idempotent, $\tilde{g}=Q\tilde{g}$, and by Lemmas A.9, A.10, and A.11,

$$|\rho'(Q\hat{h}-g)| \leqslant |\rho'Q(\hat{h}-h)|+|\rho'Q(h-g)|+|\rho'Q(g-\tilde{g})|+|\rho'(\tilde{g}-g)|$$

$$\leqslant \|Q\rho\|\,\|\hat{h}-h\|+\|Q\rho\|\,\|Q(h-g)\|+o_p(\sqrt{n}),$$

$$=o_p(b^{1/2})O_p(1)+O_p(b^{1/2})O_p(b^{1/2}n^{1/\nu})+o_p(\sqrt{n}),$$

$$=o_p(bn^{1/\nu})+o_p(\sqrt{n}). \qquad Q.E.D.$$

Proof of Theorem 4.1: For a nonzero element $g(x)$ of $G(x)$, let γ include β and $h(z,\gamma)=\rho_\beta(z,\beta)-t(x,\gamma)$ for corresponding elements of $\partial\rho(z,\beta)/\partial\beta$ and $T(x,\gamma)$. It follows by Assumption 3.4 and H that Assumption A.1 is satisfied. Also, note that by Assumptions 4.1, Assumption A.3 is satisfied for $b=n^{(\nu-2)/2\nu}$ and ρ_i any element of $\rho(z_\iota,\beta_0)$. Then by Lemma A.12,

$$\text{(A.32)} \qquad \sum_{t=1}^{n}|\hat{g}_i-g_i|^2/n=\left(\|Q\hat{h}-g\|/\sqrt{n}\right)^2=\left[o_p(n^{[(2-\nu)/2\nu]+(\nu-2)/4\nu})+o_p(1)\right]^2$$

$$=\left[o_p(o(1))+o_p(1)\right]^2=o_p(1).$$

Equation (A.1a) follows since this result holds for each element of $G(x)$. Also, by Lemma A.13,

$$\text{(A.33)} \qquad \sum_{i=1}^{n}(\hat{g}_i-g_\iota)\rho_i/\sqrt{n}=\rho'(Q\hat{h}-g)/\sqrt{n}=o_p(n^{(2-\nu)/2\nu+(\nu-2)/2\nu})+o_p(1)$$

$$=o_p(1).$$

Thus, the corresponding element of $\Sigma_{i=1}^{n}(\hat{G}_i-G_i)\otimes\rho_i/\sqrt{n}$ is $o_p(1)$. Equation (A.1b) then follows since this is true for each element of $G(x)$ and $\rho(z_i, \beta_0)$, and the conclusion of Theorem 4.1 follows by Lemma A.1. *Q.E.D.*

Proof of Lemma 4.1: Since $v(x)$ is a one-to-one function of x, $g(x)$ and $w(x)$ can be regarded as functions of $v = v(x)$. Let F_v denote the probability measure for v, and define a new probability measure by $F(S) = \int_S w(v)^2 dF_v / \int w(v)^2 dF_v$. Note that by Assumption 3.3

$$\text{(A.34)} \qquad \int [g(v)/w(v)]^2 dF = C\int g(v)^2 dF_v < \infty,$$

$$\text{(A.35)} \qquad \int \exp[\tau \|v\|]^2 dF = C\int w(v)^2 \exp[\tau\|v\|]^2 dF_v < \infty.$$

Then by Theorem 3 of Gallant (1980) and $\mathscr{a}(K) \to \infty$ it follows that there exists η_K such that $C\int [g(v)/w(v) - \sum_{m=1}^K \eta_{mK} v^{\lambda(m)}]^2 dF = o(1)$, implying

$$\text{(A.36)} \qquad E\left[\{g(x) - P^K(x)'\eta_K\}^2\right] = \int\left[g(v)/w(v) - \sum_{m=1}^{K} \eta_{mK} v^{\lambda(m)}\right]^2 w(v)^2 dF_v$$

$$= C\int\left[g(v)/w(v) - \sum_{m=1}^{K} \eta_{mK} v^{\lambda(m)}\right]^2 dF = o(1). \qquad Q.E.D.$$

Proof of Lemma 4.2: Let $P(\mathscr{g}, J, v)$ denote the Taylor series up to order J for an expansion around a point $\bar{v} = \bar{v}(\mathscr{g}) \in V(\mathscr{g})$,

$$\text{(A.37)} \qquad P(\mathscr{g}, J, v) = \mathscr{g}(\bar{v}) + \sum_{l=1}^{J} (1/l!) \sum_{m_1,\dots,m_l=1}^{r} \frac{\partial^l \mathscr{g}(\bar{v})}{\partial v_{m_1} \cdots \partial v_{m_l}} (v_{m_1} - \bar{v}_{m_1}) \cdots (v_{m_l} - \bar{v}_{m_l}).$$

By the mean value form of the remainder,

$$\text{(A.38)} \qquad |\mathscr{g}(v) - P(\mathscr{g}, J, v)|$$

$$= \left|[1/(J+1)!] \sum_{m_1,\dots,m_{J+1}=1}^{r} \frac{\partial^{J+1}\mathscr{g}(\hat{v})}{\partial v_{m_1} \cdots \partial v_{m_{J+1}}} (v_{m_1} - \bar{v}_{m_1}) \cdots (v_{m_{J+1}} - \bar{v}_{m_{J+1}})\right|$$

$$\leqslant [1/(J+1)!] \cdot r^{J+1} \sup_{v \in V(\mathscr{g})} \sup_{|\lambda| = J+1} |D^\lambda \mathscr{g}(\mathring{v})| \left(\sup_j \sup_{v \in V} |v_j - \bar{v}_j|\right)^{J+1}$$

$$\leqslant [1/(J+1)!] r^{J+1} C^{J+1} \leqslant C^{J+1}/(J+1)!,$$

where $\hat{v}$ lies on the line joining v and $\bar{v}$ and C following the last inequality does not depend on v or $\mathscr{g}$.

Next, note that there exists C such that $\mathscr{O}(K) \leqslant C\mathscr{a}(K)$. Also, for each positive integer l, the set $\{v^\lambda \mid |\lambda| \leqslant l\}$ is a subset of the set of elements of $(1, \dots, (v_1)^l) \otimes \cdots \otimes (1, \dots, (v_r)^l)$. Therefore,

$$\text{(A.39)} \qquad K \leqslant r^{\mathscr{O}(K)} \leqslant r^{C\mathscr{a}(K)}.$$

Note that all polynomials of order less than or equal to $\mathscr{a}(K)$ can be formed from linear combinations of $P^K(x) = (v^{\lambda(1)}, \dots, v^{\lambda(K)})$. Therefore, since $P(\mathscr{g}, \mathscr{a}(K), v)$ is a polynomial of order $\mathscr{a}(K)$, there exists $\eta_K(\mathscr{g})$ such that $w(v)^{-1}P^K(v)'\eta_K = P(g, \mathscr{a}(K), v)$. Then for any $\zeta > 0$, by equations (A.38) and (A.39)

$$K^\zeta \sup_{\mathscr{g} \in \mathscr{G}} \sup_{v \in V(\mathscr{g})} |\mathscr{g}(v) - w(v)^{-1} P^K(v)'\eta_K(\mathscr{g})|$$

$$\leqslant K^\zeta C^{\mathscr{a}(K)}/\mathscr{a}(K)! \leqslant (r^{C\mathscr{a}(K)})^\zeta C^{\mathscr{a}(K)}/\mathscr{a}(K)! \leqslant C^{\mathscr{a}(K)}/\mathscr{a}(K)! = o(1).$$

The conclusion now follows from the fact that $V(\mathscr{g})$ includes the support of $v(x_i)$, implying

$$\begin{aligned}&\left\{E\left[\left(g(x_i)-P^K(x_i)'\eta_K(\mathscr{g})\right)^2\right]\right\}^{1/2}\\&\quad\leqslant\left\{E\left[\left(\mathscr{g}(v_i)-w(v_i)^{-1}P^K(v_i)'\eta_K(\mathscr{g})\right)^2w(v_i)^2\right]\right\}^{1/2}\\&\quad\leqslant C\sup_{v\in V(\mathscr{g})}\left|\mathscr{g}(v)-w(v)^{-1}P^K(v)'\eta_K(\mathscr{g})\right|.\end{aligned}$$

Q.E.D.

REFERENCES

Amemiya, T. (1974): "The Non-linear Two-stage Least-Squares Estimator," *Journal of Econometrics*, 2, 105–110.

——— (1977): "The Maximum Likelihood and Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equations Model," *Econometrica*, 45, 955–968.

——— (1985): *Advanced Econometrics*. Cambridge, Massachusetts: Harvard University Press.

Berndt, E. R., B. H. Hall, R. E. Hall, and J. A. Hausman (1974): "Estimation and Inference in Nonlinear Structural Models," *Analysis of Economic and Social Measurement*, 3, 653–666.

Burguete, J. F., A. R. Gallant, and G. Souza (1982): "On the Unification of the Asymptotic Theory of Nonlinear Econometric Models," *Econometric Reviews*, 1, 151–190.

Carroll, R. J. (1982): "Adapting for Heteroskedasticity in Linear Models," *Annals of Statistics*, 10, 1224–1233.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Edmunds, D. E., and V. B. Moscatelli (1977): "Fourier Approximation and Embedding in Sobolev Space," *Dissertationes Mathematicae*, 145, 1–46.

Freud, G. (1971): *Orthogonal Polynomials*. Oxford: Pergammon Press.

Gallant, A. R. (1980): "Explicit Estimators of Parametric Functions in Nonlinear Regression," *Journal of the American Statistical Association*, 75, 182–193.

——— (1981): "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form," *Journal of Econometrics*, 15, 211–245.

Hannan, E. J. (1963): "Regression for Time Series," in *Time Series Analysis*, ed. by M. Rosenblatt. New York: Wiley, pp. 17–37.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hardle, W., P. Hall, and J. S. Marron (1988): "How Far Are Automatically Chosen Regression Parameters From Their Optimum," *Journal of the American Statistical Association*, 83, 86–101.

Heckman, J. J. (1978): "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica*, 46, 931–959.

Heckman, J. J., and R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. Singer. Cambridge, England: Cambridge University Press.

Hsieh, D., and C. Manski (1987): "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression," *Annals of Statistics*, 15, 541–551.

Jorgenson, D. W., and J. Laffont (1974): "Efficient Estimation of Nonlinear Simultaneous Equations with Additive Disturbances," *Annals of Economic and Social Measurement*, 3, 615–640.

Kelejian, H. (1971): "Two-Stage Least Squares and Econometric Systems Linear in Parameters and Nonlinear in Endogenous Variables," *Journal of the American Statistical Association*, 66, 373–374.

MaCurdy, T. E. (1982): "Using Information on the Moments of the Disturbance to Increase the Efficiency of Estimation," mimeo, Stanford University.

McFadden, D. (1985): "Specification of Econometric Models," Presidential Address, 1985 Meeting of the Econometric Society.

Newey, W. K. (1986): "Efficient Estimation of Models with Conditional Moment Restrictions," mimeo, Princeton University.

——— (1988): "Adaptive Estimation of Regression Models Via Moment Restrictions," *Journal of Econometrics*, forthcoming.

POWELL, M. J. D. (1981): *Approximation Theory and Methods*. New York: Cambridge University Press.

ROBINSON, P. (1976): "Instrumental Variables Estimation of Differential Equations," *Econometrica*, 44, 756–776.

——— (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875–891.

——— (1988): "Automatic Generalized Least Squares," mimeo, London School of Economics.

STONE, C. J. (1977): "Consistent Nonparametric Regression" (with discussion), *Annals of Statistics*, 5, 595–645.