cemmap

centre for microdata methods and practice

Instrumental variables estimation for nonparametric models

Whitney K. Newey
James L. Powell

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP07/17

Instrumental Variables Estimation for Nonparametric Models[*]

by

Whitney K. Newey
Princeton University and BellCore
609-258-4034

and

James L. Powell
University of Wisconsin
608-262-2265

December 1988
Revised, October 1989

[*] This research was supported by the National Science Foundation, the Sloan Foundation, and BellCore. An earlier version of this paper was presented at the 1988 North American Winter Meeting of the Econometric Society. Discussions with Angus Deaton inspired this paper. Helpful comments were provided by C. Manski, L. Hansen, J. Heckman, C. Meghir, J. Scheinkman, G. Wahba, and participants in seminars at BellCore and Chicago, Northwestern, Princeton, Yale, and Wisconsin Universities.

1. Introduction

Nonparametric regression in econometric applications provides a way of uncovering the reduced form relationship between a dependent variable and explanatory variables, without imposing functional form restrictions. In particular, nonparametric conditional expectation estimation, which is a well developed topic in the statistics literature, can be used to estimate how the mean of a dependent variable depends on explanatory variables. Recent econometric applications include Deaton (1988) and Pagan and Hong (1990).

In econometrics, there are many occasions where knowledge of the structural relationship among dependent variables is required to answer questions of interest. The purpose of this paper is to develop methods of estimation for nonparametric structural models. As illustrated below, these methods should prove useful in applications such as nonparametric estimation of supply and demand models, marginal rates of substitution in consumption based asset pricing models, and nonparametric regression with errors-in-variables.

The importance of structural estimation is familiar from the literature on econometric policy analysis (e.g. Lucas (1976)), but a nonparametric illustration may help drive the point home. Consider a classical supply and demand example, where the price P and quantity traded Q in some market are assumed to simultaneously solve

$$(1.1) \qquad P = g_D(Q, Y) + U \qquad \text{(inverse demand)},$$

$$(1.2) \qquad Q = g_S(P, W) + V \qquad \text{(supply)}.$$

Here Y and W are exogenous forcing variables (e.g. income and weather respectively), while U and V are unobservable shocks to the supply and demand equations. Suppose a sales tax τ is to be imposed, and it is desired to estimate the expected revenue from this tax. Let $\{(P_t, Q_t, Y_t, W_t), t = 1, \ldots, n\}$ be a sample of observations on this market with $\tau = 0$. If nonparametric estimates $\hat{g}_D(Q,Y)$ and $\hat{g}_S(P,W)$ of the supply and demand equations were available, then the expected revenue, $E[R(\tau)] = E[\tau \cdot Q(\tau)]$, could be estimated by

$$(1.3) \qquad \hat{R}(\tau) = \Sigma_{t=1}^{n} \tau \cdot \hat{Q}_t(\tau)/n,$$

where $\hat{Q}_t(\tau)$ is defined as the solution to

$$(1.4) \qquad \hat{Q}_t(\tau) = \hat{g}_S\big((1-\tau) \cdot (\hat{g}_D(\hat{Q}_t(\tau), Y_t) + \hat{U}_t),\ W_t\big) + \hat{V}_t,$$

and $\hat{U}_t$ and $\hat{V}_t$ are calculated from relations (1.1) and (1.2) with estimated supply and demand functions replacing true. This estimator is not feasible using only the reduced form regression functions $E[Q \mid W, Y]$ and $E[P \mid W, Y]$.

In linear models, zero covariance between instruments and disturbances, along with identification, suffices for consistent estimation. In a nonparametric setting, a stronger restriction that the disturbance has conditional mean zero given instruments is important; a finite number of zero covariance restrictions will not suffice to identify an infinite dimensional function. One object of the present paper is to investigate whether identification with such restrictions permits consistent nonparametric estimation of structural relations. As will be shown below, it does not. The moment restrictions imply that the structural relations solve a particular set of integral equations involving conditional distributions and expectations of observable variables; however, these integral equations are not well-behaved when the true conditional distributions and expectations are replaced by estimates, which complicates the consistency arguments. To obtain consistent estimates, then, we impose further restrictions on the structural relations which, while nonparametric, are stronger than the usual "smoothness" requirements imposed for consistency of nonparametric regression estimators.

We also investigate the extent to which this problem can be circumvented via strengthening the assumptions, by considering a triangular model with independent disturbances and a reduced form for right-hand side endogenous variables that is linear in disturbances. A residual adjusted, additive nonparametric regression estimator is suggested for this model. The consistency of this estimator will not require the smoothness conditions alluded to above.

Unfortunately, the assumptions of the triangular model may be too strong for some important applications. Thus, it would also be of interest to investigate intermediate cases, such as a model with independent disturbances but no restriction on the reduced form. This investigation is currently underway, but as yet there are few results to report.

2. The Conditional Mean Model

The first model we consider takes the form

$$(2.1)\qquad E[\rho(z,\theta_0)|x] = 0$$

where z denotes a data point, θ denotes a vector of functions, $\rho(z,\theta)$ a residual vector, and x a vector of instruments. The difference between this model and familiar moment restriction models is that θ can be infinite dimensional, i.e. be a function rather than a vector of real numbers.

To motivate the results and provide useful illustrations it is helpful to consider a number of examples. The first is a single equation with endogenous right-hand side variables, taking the the form

$$(2.2) \qquad y = \theta_0(z_2) + \varepsilon, \quad E[\varepsilon|x] = 0.$$

This model is a special case of that of equation (2.1) with $z = (y,z_2,x)$ and $\rho(z,\theta) = y-\theta(z_2)$. For instance, if equation (2.2) is interpreted as either the supply or demand equation, the nonparametric supply and demand model has this form if the disturbances have conditional mean zero given Y and W. In that case, $x = (Y,W)$.

Another example is a nonparametric version of the consumption asset-pricing model considered, for example, by Hansen and Singleton (1982). For this case let C and C_+ denote current and next period consumption, and R and R_+ current and next period asset returns. Also, let $\theta(C,C_+)$ denote the intertemporal marginal rate of substitution between the current and next periods. The first order conditions for expected utility maximization include

$$(2.3) \qquad E[\theta_0(C,C_+)R_+|R,C] = 1.$$

This equation is a special case of equation (2.1) where $z = (C,C_+,R,R_+)$, $x = (R,C)$, and $\rho(z,\theta) = \theta(C,C_+)R_+ - 1$. Note that here the unknown function is the marginal rate of substitution. Conditions for its nonparametric identification and estimation will be discussed below.

A third example can be used to illustrate the importance of allowing $\rho(z,\theta)$ to be a vector. Consider a nonparametric errors-in-variables model of the form

$$(2.4) \qquad \begin{aligned} y &= \theta_{10}(w^*) + \zeta, && E[\zeta|x,v] = 0, \\ w &= w^* + \eta, && E[\eta|x,v,\zeta] = 0, \\ w^* &= x + v, && v \text{ independent of } x \text{ with density } \theta_{20}(v), \end{aligned}$$

where w^*, ζ, η, and v are unobserved, but y, w, and x are observed. Here w^* represents an unobserved regressor, w a measurement of w^*, and x a causal variable for w^*. Note that in practice x may depend on unknown parameters; the case with known x is considered here for simplicity. As discussed in Hausman, Ichimura, Newey, and Powell (1985), a fundamental implication of this model is

$$(2.5) \qquad E[w^{\ell-1}y|x] = \int[x+v]^{\ell-1}\theta_{10}(x+v)\theta_{20}(v)dv, \quad \ell = 1,2.$$

This equation has the form of (2.1) for $z = (y,w,x)$, $\theta = (\theta_1,\theta_2)$, and $\rho(z,\theta) = (\rho_1(z,\theta),\rho_2(z,\theta))'$ with $\rho_\ell(z,\theta) = w^{\ell-1}y - \int[x+v]^{\ell-1}\theta_{10}(x+v)\theta_{20}(v)dv$, $(\ell = 1,2)$. Both conditional moment restrictions appear to be important for identification of θ; there are two functions to be estimated, suggesting an "order condition" that at least two conditional moment restrictions are present. Indeed, lack of identification of θ_1 and θ_2 with only one moment restriction is shown in Hausman et. al. (1985). This model is further considered in Hausman, Newey, and Powell (1989).

3. Identification

In the general model of equation (2.1) the fundamental necessary identification condition for the existence of a consistent estimator is

Assumption 3.1: For $\theta \in \Theta$, $E[\rho(z,\theta)|x] = 0$ implies $\theta = \theta_0$.

This identification condition is nonparametric, in the sense that θ is allowed to vary in an infinite dimensional set. For identification it is essential that x not be constant, i.e. that the moment conditions not reduce to unconditional restrictions. In general a finite number of unconditional moment conditions will not identify an infinite-dimensional function. More generally, it is important that there be enough variation in x relative the variation in the argument of the function θ.

When $\rho(z,\theta)$ is linear in θ, it is possible to be more specific about conditions for identification. Consider the single equation model of equation (2.2). The conditional moment restriction $E[\varepsilon|x] = 0$ is equivalent to

$$(3.1) \qquad \pi(x) \equiv E[y|x] = E[\theta_0(z_2)|x] = \int\theta_0(z_2)f(z_2|x)dz_2,$$

where $f(z_2|x)$ denotes the conditional density of z_2 given x. The function $\pi(x)$ is the nonparametric generalization of the reduced form for y; note $y = \pi(x) + v$, with $E[v|x] = 0$. It is known that, under weak regularity conditions, conditional expectations and densities can be estimated consistently. Hence $\pi(x)$ and $f(z_2|x)$ are identified, and can be thought of as known for the purposes of identification of θ_0. The identification of θ_0 thus depends on the existence of a unique solution to the integral equation (3.1).

Existence of a unique solution to the integral equation is equivalent to completeness of the conditional distribution of z_2 given x, a concept we borrow from the literature on minimum variance unbiased estimation. By subtracting equation (3.1) from the same equation with $\tilde{\theta}(z_2)$ substituted for $\theta_0(z_2)$, it is easily seen that identification is equivalent to the nonexistence of any function $\epsilon(z_2) \equiv \tilde{\theta}(z_2) - \theta_0(z_2) \neq 0$ such that $E[\epsilon(z_2)|x] = 0$.

There are important examples where completeness is known to hold. If the conditional density function is a member of an exponential family of the form,

$$(3.2) \qquad f(z_2|x) = a(z_2)b(x)\exp\{h(x)'\tau(z_2)\},$$

where $a(z_2) > 0$ on the support of z_2, $\tau(z_2)$ is one-to-one, and the support of $h(x)$ contains an open set, then completeness of $f(z_2|x)$ follows from well known results on complete, sufficient statistics, e.g. Ferguson (1967, p. 134). Thus, θ_0 will be identified if the conditional distribution of z_2 given x takes this form.

Completeness is a natural generalization to nonparametric models of the familiar conditions for identification in parametric, linear models. Consider a parametric model with θ_2 linear in z_2 with unknown parameters γ, say $\theta_0(z_2) = z_2'\gamma_0$, and the conditional expectation of z_2 given x linear in x, say $E[z_2|x] = \Pi x$. The condition that the integral equation have a unique solution is

$$E[z_2'(\gamma_0-\tilde{\gamma})|x] = x'\Pi'(\gamma_0-\tilde{\gamma}) = 0 \quad \text{implies} \quad z_2'(\gamma_0-\tilde{\gamma}) = 0.$$

If neither of the distributions of x or z_2 concentrate on a hyperplane, this statement is equivalent to

$$\Pi'(\gamma_0-\tilde{\gamma}) = 0 \quad \text{implies} \quad \gamma_0 = \tilde{\gamma};$$

that is, Π' has full column rank. This condition is the familiar rank condition, e.g. see Fisher (1976).

In special cases there are interesting necessary conditions that correspond to the order condition in linear models. For instance, in the exponential family example of equation (3.2), identification requires that $h(x)$ vary over an open set. If $h(x)$ is restricted to be continuously differentiable then this condition implies that there be as many components of x as are in z_2, i.e. as many instruments as right-hand side variables. For another example, suppose that both x and z_2 are discrete with finite support. Then equation (3.1) becomes a set of linear equations in the value of $\theta_0(z_2)$ at each support point of z_2, with coefficients given by the probability of each z_2 point given each x point, and one equation for each support point of x. A necessary condition for existence of a unique solution to such an equation system is that there be as many equations as numbers that have to be solved for, i.e. that x has as many support points as z_2. Indeed, with this discrete example, the necessary and sufficient condition for identification is that the rank of the matrix of conditional probabilities be equal to the number of support points of z_2.

Rhoerig (1988) has previously considered nonparametric identification of the single equation (2.2), but under stronger conditions than those imposed here. He assumes that the equation is a member of a system of equations with continuously distributed instruments and disturbances that are stochastically independent of instruments, while we only impose a conditional mean restriction on the disturbance for a single equation.

The completeness condition is also useful for understanding identification in some more complicated models, such as that of equation (2.3). After differencing this equation for a pair of θ values θ_0 and $\tilde{\theta}$, the identification condition becomes

(3.3) $\quad 0 = \int\delta(C,C_+)R_+f(C_+,R_+|C,R)dC_+dR_+ = \int\delta(C,C_+)\{\int R_+f(C_+,R_+|C,R)dR_+\}dC_+$

implies $\delta(C,C_+) = 0$.

By comparing the expression following the second equality with equation (3.3), fixing C, and multiplying through by $1/E[R_+|C,R]$ we see that this condition is equivalent to completeness, for each C, of the "return adjusted" conditional density $\{\int R_+f(C_+,R_+|C,R)dR_+\}/E[R_+|C,R]$ of C_+ given R. For example, if the original conditional density takes the exponential form

$$f(C_+,R_+|C,R) = b(C,R)a(C_+,R_+)\exp\{h_1(C,R)\tau_1(R_+) + h_2(C,R)\tau_2(C_+)\},$$

then the return adjusted conditional density will have an exponential form with exponent $\exp\{h_2(C,R)\tau_2(C_+)\}$, so that the identification condition will hold if $h_2(C,R)$ varies over an open set for all fixed C in some set with probability approaching one. Of course this condition is restrictive, but we expect that identification will hold more generally.

Some additional insight is provided by a discrete case where the support of consumption and returns is a finite set. Here, equation (3.3) becomes a set of linear equations with coefficients given by return-adjusted conditional probabilities. The necessary and sufficient condition for identification is that the matrix of return adjusted conditional probabilities has rank equal to the number of support points for consumption, for each possible value of lagged consumption, for which it is necessary that there be at least as many return support points as consumption support points. Of course, it should be noted that further lagged variables can aid in identification if the model is not first order Markov.

Identification in models with $\rho(z,\theta)$ nonlinear in θ can be difficult to analyze, just as is identification in parametric nonlinear models. As in parametric models, the only general identification conditions will be local ones.

4. Nonparametric Two-Stage Least Squares

Estimation of θ_0 presents practical and theoretical challenges. It will be helpful to address these challenges by focusing on the example of equation (2.2). For this model, an estimation scheme is suggested by the previous identification analysis. Given estimates $\hat{\pi}(x)$ and $\hat{f}(z_2|x)$ of the reduced form and conditional density respectively, it ought to be possible to "solve" the integral equation

$$(4.1) \qquad \hat{\pi}(x) = \int\theta(z_2)\hat{f}(z_2|x)dz_2$$

to obtain an estimator of θ_0. The practical difficulty with this scheme is solving this functional equation.

An approach to estimating solutions to other integral equations has been considered the statistics literature, e.g. Wahba (1979), Nychka et. al. (1984), and O'Sullivan (1986). This literature has been concerned with a related model, taking the form

$$(4.2) \qquad y = \int K(x,z_2)\theta_0(z_2)dz_2 + v, \quad E[v|x] = 0,$$

where z_2 is not observed, but $K(x,z_2)$ is known, and v represents observational or modeling error. In Nychka et. al. (1984), y is the two-dimensional, cross-section radius of a tumor, z_2 is the three-dimensional radius of a tumor, and the form of $K(x,z_2)$ is that implied by modeling tumors as spheres randomly distributed in tissue. The problem of estimating $\theta_0(z_2)$ from $K(x,z_2)$ and an estimate of $\int K(x,z_2)\theta_0(z_2)dz_2$ is similar to the problem of estimating $\theta_0(z_2)$ from an estimate of $\pi(x)$ and $f(z_2|x)$, although our problem is more complicated because $f(z_2|x)$ must be estimated from observations on z_2 and x, while $K(x_2,z)$ is known.

The approach suggested in this literature, which we will adopt here, is to use a linear in parameters approximation for θ_0. Equation (4.1) becomes linear in parameters for such an approximation, making it relatively straightforward to construct an estimate of θ_0. Let

$$\{p_1(z_2),\ p_2(z_2),\ \ldots\ \}$$

be a sequence such that linear combinations of enough terms can approximate any function of z_2 in a sense to be specified below. A linear in parameters (i.e. series) approximation of $\theta(z_2)$ is given by

$$\theta(z_2,\gamma) = \textstyle\sum_{j=1}^{J}\gamma_j p_j(z_2).$$

Substituting $\theta(z_2,\gamma)$ in equation (4.1) and equating $\int p_j(z_2)\hat{f}(z_2|x)dz_2$ with a conditional expectation estimator $\hat{E}[p_j|x]$ yields

$$(4.3)\qquad \hat{\pi}(x) = \textstyle\sum_{j=1}^{J}\gamma_j\hat{E}[p_j|x].$$

An estimate $\hat{\theta}(z_2) = \sum_{j=1}^{J}\hat{\gamma}_j p_j(z_2)$ of $\theta_0(z_2)$ can be obtained by choosing $\hat{\gamma}_j$ to minimize some measure of distance between observations on the left and right-hand sides of equation (4.3). We will focus on the Euclidean distance measure, with $\hat{\gamma}_j$ obtained as the solution to

$$(4.4)\qquad \min_{\theta(z_2,\gamma)\in\Theta}\textstyle\sum_{t=1}^{n}\{\hat{\pi}(x_t) - \sum_{j=1}^{J}\gamma_j\hat{E}[p_j|x_t]\}^2,$$

where z_t denotes the data observations and the form of the set Θ will be discussed below.

This estimator is a nonparametric generalization of 2SLS. The first stage is calculation of the conditional expectation estimators $\hat{\pi}(x_t)$ and $\hat{E}[p_j|x_t]$, and the second a (constrained) least squares regression of $\hat{\pi}(x_t)$ on these conditional expectations. Nonparametric estimation occurs in both stages. The first stage makes use of estimates of conditional expectations of the right-hand side variables, rather than linear projections. If the model were parametric then such a first stage would result in an efficient instrumental variables estimator, under the conditions of Newey (1990); here the conditional expectations are useful for identification. The second stage is nonparametric in that i~~s~~t makes use of an arbitrarily flexible approximation to θ. Consistency will require that J, the number of approximating terms, goes to infinity with the sample size.

It should be noted that the estimator of equation (4.4) corresponds to a particular interpretation of two-stage least squares, involving a predicted value for the left-hand side variable. It is also possible to replace $\hat{\pi}(x_t)$ by y_t in the objective function without affecting the consistency of the result, although for pedagogical reasons we will continue to focus on equation (4.4).

To operationalize this estimator, the approximating series, the number of included terms, and the estimates of the conditional expectations must each be specified. There are many candidates for the series. The one that we will focus on here is power series in a one-to-one, bounded transformation of z_2. Let $\tau(z_2) = (\tau_1(z_2),\ldots,\tau_q(z_2))'$ denote such a transformation, where q is the dimension of z_2. For example, $\tau_\ell(z_2) = \exp(z_{2\ell})/[1+\exp(z_{2\ell})]$ would do. Such a power series would take the form

(4.5) $p_j(z_2) = \Pi_{\ell=1}^{q}\tau_\ell(z_2)^{\lambda_\ell(j)}$, $(j = 1, 2, \ldots)$,

where $\lambda_\ell(j)$ are nonnegative integers. Typically, such a series would be constructed using low order terms (i.e. small values for $\lambda_\ell(j)$) for low values of j.

There are a number of different estimators of conditional expectations that one could employ in the first stage, including kernel, nearest neighbor, and series. The consistency result will be general enough to allow for several possibilities.

A important practical issue for this estimation procedure, as with other nonparametric methods, is the choice of approximation degree. This choice includes the number of terms J in the second stage as well as the choice of "smoothing parameters" in the first stage. A data based choice of these terms seems essential to making practical the theoretical advantages of nonparametric methods. One possibile data-based method would be cross validation at each stage. The conditional expectation estimates could be chosen by cross-validation (e.g. see Hardle, 1990, for exposition). Then, with the resuting conditional expectations estimates $\hat{\pi}(x_t)$ and $\hat{E}[p_j|x_t]$ in hand, then J could be chosen by cross-validating the second stage, i.e. to minimize

$$\Sigma_{t=1}^{n}\{\hat{\pi}(x_t) - \Sigma_{j=1}^{J}\hat{\gamma}_j^{-t}\hat{E}[p_j|x_t]\}^2,$$

where $\hat{\gamma}_j^{-t}$ is calculated from the second stage for all observations but the t^{th}. As is well known, this formula can be substantially simplified. Of course, these are preliminary suggestions that deserve more careful consideration. In particular, it is not clear that choosing J based on the second stage is appropriate when one is interested in the structural function θ_0; see the discussion following O'Sullivan (1984) for similar remarks.

An analogous estimation procedure is available whenever the residuals are linear in θ. For example, consider the model of equation (2.2), with identifying integral equation (3.3). The estimating equation that corresponds to (4.1) is

$$(4.6) \qquad 1 = \int\theta(C,C_+)R_+\hat{f}(C_+,R_+|C,R)dC_+dR_+.$$

Replacing θ by a linear in parameters approximation $\theta(C,C_+,\gamma) = \sum_{j=1}^{J}\gamma_jp_j(C,C_+)$ and equating $\int p_j(C,C_+)R_+\hat{f}(C_+,R_+|C,R)dC_+dR_+$ and $\hat{E}[p_jR_+|C,R]$,

$$(4.7) \qquad 1 = \sum_{j=1}^{J}\gamma_j\hat{E}[p_jR_+|C,R].$$

An estimator $\sum_{j=1}^{J}\hat{\gamma}_jp_j(C,C_+)$ of θ_0 can be obtained by choosing $\hat{\gamma}$ to minimize the Euclidean distance between observations on the right and left hand sides of equation (4.7), i.e. as the solution to

$$(4.8) \qquad \min_{\theta(C,C_+,\gamma)\in\Theta}\sum_{t=1}^{n}\{\ 1 - \sum_{j=1}^{J}\gamma_j\hat{E}[p_jR_+|C_t,R_t]\ \}^2.$$

The estimators for these two models have a common structure. Let $\theta(\gamma)$ denote a parametric approximation to θ. Consider a conditional expectation estimator $\hat{E}[\cdot|x]$ that is linear, in the sense that for constants a_1 and a_2, $\hat{E}[a_1\psi_1 + a_2\psi_2|x] = a_1\hat{E}[\psi_1|x] + a_2\hat{E}[\psi_2|x]$, and that $\hat{E}[1|x] = 1$. Then in each case the estimator is the solution to

$$(4.9) \qquad \min_{\theta(z,\gamma)\in\Theta}\sum_{t=1}^{n}\{\ \hat{E}[\rho(\theta(\gamma))|x_t]\ \}^2,$$

where $\theta(\gamma) = \theta(z,\gamma)$ and $\rho(\theta) = \rho(z,\theta)$ is the model residual. This estimator is the natural generalization of the examples to models that are nonlinear in θ. It is a nonlinear, nonparametric 2SLS estimator, if you will.

For models that are nonlinear in θ, the computation of the solution of

(4.9) may not be particularly easy, but the parametric approximation step at least makes it feasible. For instance, the nonlinear errors-in-variables example has a residual vector that is a quadratic function of θ. Consequently, if a linear in parameters approximation to θ is used, the residual is quadratic in the parameters, unlike the linear in parameters residuals above, and estimation is more difficult. In this example, it turns out that a recursive estimation scheme can be developed when certain polynomial approximations to θ are used; see Hausman et. al. (1985) and Hausman, Newey, and Powell (1989). In other examples, computation may be even more difficult.

In order to estimate models like the nonlinear errors-in-variables example it is useful to be able to use more than one residual. One way to do this is to construct a minimum distance estimator (e.g. Malinvaud, 1980) of the following form. Let $\hat{E}[\rho(\theta)|x]$ denote a vector of estimators of the conditional expectations of the components of $\rho(z,\theta)$ and let $\hat{A}$ be a positive definite matrix. Consider an objective function of the form

$$\hat{Q}(\theta) = \sum_{i=1}^{n}\hat{E}[\rho(\theta)|x_t]'\hat{A}\hat{E}[\rho(\theta)|x_t]/n$$

$$= \mathrm{trace}\{\hat{A}\sum_{t=1}^{n}\hat{E}[\rho(\theta)|x_t]\hat{E}[\rho(\theta)|x_t]'/n\}.$$

Using a parametric approximation $\theta(\gamma)$, an estimator $\hat{\theta} = \theta(\hat{\gamma})$ can be computed by choosing $\hat{\theta}$ as the solution to

$$\min_{\theta(\gamma)\in\Theta}\hat{Q}(\theta).$$

As previously discussed, the single equation estimators considered above are of this form, and the estimator for the errors-in-variables model in Hausman, Newey, and Powell (1989) is also of this form. In the next Section we consider consistency of this estimator, and discuss the nature of the constraints that are implicit in the condition that $\theta(\gamma)\in\Theta$.

5. Consistency

There is a difficulty in showing consistency of nonparametric 2SLS in the first example which is apparently generic. Equation (3.1) is a~~n~~ Fredholm integral equation of the first kind, and hence its solution need not be a continuous (in mean square) function of π. The reason for this is that the integral operator $\int\theta(z_2)f(z_2|x)dx$ need not have a continuous inverse; it does not have closed range. That is, there exists reduced forms that are close in mean-square with corresponding structures that are far apart. This noncontinuity allows the possibility that the estimated structure may be far from the truth, even when the reduced form is close to the truth, causing an obvious problem for a consistency argument. This same feature has previously been noted for the problem of equation (4.2) by Wahba (1979), Nychka et. al. (1984), and O'Sullivan (1986).

We circumvent this problem by restricting the set Θ over which estimation is carried out to be a compact subset of a normed set of functions (and θ_0 to be an element of this set). A well known topological result is that a continuous, one-to-one mapping on compact metric spaces has a continuous inverse. Thus, because the integral operator is continuous, for an identified model the mapping from the reduced form to the structure is continuous on such a compact set. Gallant (1981, 1987) has previously considered estimation on compact function sets (in other contexts), and our consistency results will be based on the same logic as his.

The type of compact set of functions we will consider imposes bounds on higher-order derivatives. These bounds impose constraints on the nonparametric 2SLS coefficients, i.e. they specify the nature of the constraint set Θ imposed in estimation. The practical implication of these constraints is that the estimated function not be "too wiggly." Thus, in applications it will be important to check the shape of the function and impose constraints on the coefficients if it does not appear to be smooth enough. These constraints will often take the form of dampened magnitude of higher order terms in the series approximation, which are typically the source of nonsmooth behavior. See Elbadawi, Gallant and Souza (1983) for further discussion of how such constraints might be imposed in practice.

We will first give a consistency theorem that is applicable to any norm, residual vector, and conditional expectation estimator having certain properties, and then specialize it. For notational simplicity we will assume throughout that the data are stationary. For a matrix $A = [a_{ij}]$ let $\|A\| = [\text{trace}(A'A)]^{1/2}$ denote the Euclidean norm, and for a function θ let $\|\theta\|$ denote a function norm (to be further specified below). The first assumption imposes compactness.

Assumption 5.1: $\theta_0 \in \Theta$, and Θ is compact in the norm $\|\theta\|$.

The second Assumption guarantees that the finite dimensional approximation is rich enough to approximate the truth, no matter what it happens to be. Let J index the degree of a parametric approximation to θ, e.g. the number of terms in the series approximation, and let $\theta_J(\gamma)$ denote a value of this approximation corresponding to a parameter vector γ.

Assumption 5.2: For any $\theta \in \Theta$ there exists $\tilde{\gamma}_J$ such that $\lim_{J\longrightarrow\infty}\|\theta_J(\tilde{\gamma}_J)-\theta\| = 0$.

It should be noted the parametric approximation is important as a computational device. Under the following Assumptions, $\hat{Q}(\theta)$ is a continuous function on the compact set Θ, so that there exists an estimator that minimizes $\hat{Q}(\theta)$ over all of Θ. However, this estimator is difficult to calculate. Restricting θ to a finite dimensional family for any particular realization of the data simplifies this task.

The next assumption imposes a dominated Lipschitz condition on the residual, in terms of the norm $\|\theta\|$.

Assumption 5.3: There exists ϵ, $M(z) > 0$ such that $E[\|\rho(z,\theta_0)\|^{2+\epsilon}] < \infty$ and for all θ, $\tilde{\theta} \in \Theta$, $\|\rho(z,\tilde{\theta})-\rho(z,\theta)\| \leq M(z)\|\tilde{\theta}-\theta\|^{\epsilon}$ and $E[M(z)^{2+\epsilon}] < \infty$.

Assumptions 5.1 - 5.3 are sufficient for the compactness, denseness, and continuity hypotheses for consistent estimation with finite dimensional approximations to compact parameter sets, as exposited in Gallant (1987). The next Assumption is useful for guaranteeing uniform convergence in probability of the objective function. Let $\psi(z)$ denote a function of a data observation.

Assumption 5.4: For $\epsilon > 0$ from Assumption 5.2, i) $\hat{A} \xrightarrow{p} A$, A is positive definite, and if $E[|\psi(z_t)|^{1+\epsilon/2}] < \infty$ then $\sum_{t=1}^{n}\psi(z_t)/n \xrightarrow{p} E[\psi(z_t)]$; ii) if $E[\psi(z)^{2+\epsilon}]$ is finite, $\sum_{t=1}^{n}\|\hat{E}[\psi(z)|x_t] - E[\psi(z)|x_t]\|^2/n \xrightarrow{p} 0$; iii) either a) $\hat{E}[\psi(z)|z_t] = \sum_{s=1}^{n}w_{ts}\psi(z_s)$, $w_{ts} \geq 0$, $\sum_{s=1}^{n}w_{st} = 1$, $(s,t=1,\ldots,n)$, and and if $E[|\psi(z)|^{1+\epsilon/2}] < \infty$, $\sum_{t=1}^{n}\hat{E}[\psi(z)|z_t]/n = O_p(1)$; or b) $\hat{E}[\psi(z)|z_t] = P_t'(\sum_{s=1}^{n}P_sP_s')^{-}\sum_{s=1}^{n}P_s\psi(z_s)$.

Assumption 5.4 can easily be checked in some circumstances, and is general enough to allow verification via future results. For instance, if the data are i.i.d. then it is easy to use known results to show that Assumption 5.4 holds for nearest neighbor and series estimators. For K-nearest neighbor estimators with $K \rightarrow \infty$, $K/n \rightarrow 0$, ii) follows by Lemma 8 of Robinson (1987) and Proposition 1 of Stone (1977), while iii) a) holds by construction and

Stone's (1977) L^p convergence result for nearest neighbor estimators. For series estimator of the form given in iii) b), with P_t containing K elements such that any function with finite mean square can be approximated arbitrarily well in mean-square for large enough K, ii) follows as in Lemmas A.10 and Lemma A.11 of Newey (1990), as long as $K \longrightarrow \infty$ and $K/n^{\epsilon/(\epsilon+2)} \longrightarrow 0$. Assumption 5.4 should also be "plug-compatible" with future results on nonparametric condtional expectation estimators, such as time series properties needed for the fully primitive treatment of the second example.

This Assumption allows for some forms of data-based smoothing for conditional expectations estimators. Convergence results that hold for all features of a problem always allow trivially for a choice of these features from among a finite number of all possible features. Thus, the result below automatically allows for the conditional expectations estimators to be chosen from a finite number of alternatives; e.g. for each sample size the K for nearest neighbor or series estimators could be chosen from among the elements, correspdonding to that sample size, of a finite number of sequences.

For series estimators, Assumption 5.4 will be satisfied for general forms of data dependence under stronger conditions; see Newey (1990). However, Assumption 5.4 imposes a strong restriction on the form of such data dependence. Implicitly, the form of the weights w_{st} in Assumption 5.4 and the approximating functions P_t are restricted to not depend on ψ. Thus, while they could be chosen based on some fixed ψ (e.g. a linear combination of $\rho(z,\bar{\theta})$ for some preliminary estimator $\bar{\theta}$), they are not allowed to vary with ψ (i.e. with θ in $\rho(z,\theta)$).

The next Assumption specifies the behavior of the approximation degree.

Assumption 5.5: $\hat{\theta} = \text{argmin}_{\theta\in\hat{\Theta}}\hat{Q}(\theta)$ where $\hat{\Theta} = \{\theta_{\hat{J}}(\gamma) \in \Theta\}$ and $\hat{J} \xrightarrow{p} \infty$.

As promised earlier, this assumption allows for the degree of approximation J

to be data-based, in a very general way. However, it should be noted that it is not restrictions on the growth rate of J that are used to obtain consistency, but rather the compactness restriction. Thus, while rapid growth rates of J are allowed by this assumption, it is plausible that the compactness restriction would have more "bite" for large values of J, imposing strong constraints on the coefficients of higher order terms.

The following is the general consistency result.

Theorem 5.1: If Assumptions 3.1, 5.1 - 5.5 are satisfied then $\|\hat{\theta}-\theta_0\| \xrightarrow{p} 0$.

We specified the hypotheses of this result to be very general in the hopes that they would be satisfied in a variety of environments. However, for the purpose of checking them in a particular example, primitive conditions are more useful. Assumptions 5.3 and 5.5 are already in a primitive form, and primitive conditions for Assumption 5.4 were discussed above, but we have not given conditions for Assumptions 5.1 and 5.2.

The primitive conditions for these Assumptions will be smoothness restrictions of the form alluded to earlier. For the moment, consider the case where θ is a single function of an argument $v \in \mathbb{R}^p$. Denote the partial derivatives of $\theta(v)$ by

$$D^{\lambda}\theta(v) = (\partial/\partial v_1)^{\lambda_1}\cdots(\partial/\partial v_p)^{\lambda_p}\theta(v),$$

where $\lambda = (\lambda_1,\cdots,\lambda_p)$ is a p-vector of nonnegative integers. The order of the derivative is $|\lambda| = \sum_{\ell=1}^{p}|\lambda_\ell|$. Let V denote a subset of $\mathbb{R}^p$.

Assumption 5.1′: V is a bounded, open, convex set, $\Theta = \{\theta : \theta(v)$ is continuously differentiable to order $m+2$ on V, $\max_{|\lambda|\le m+1}\sup_{v\in V}\|D^{\lambda}\theta(v)\| \le b$, $|\lambda| \le m+1\}$, and $\|\theta\| = \max_{|\lambda|\le m}\sup_{v\in V}|D^{\lambda}\theta(v)|$.

Compactness of Θ in the norm $\|\theta\|$ is shown in Elbadawi, Gallant, and Souza

(1983).

A bounded domain for θ is a strong restriction, that can be circumvented in some cases. For instance, it may be possible to transform the argument of θ so that this assumption is satisfied. Let $\tilde{v} \in \mathbb{R}^p$ denote the argument of θ, and let $\tau(\tilde{v})$ denote a bounded tansformation with infinitely continuously differentiable inverse, e.g.

$$\tau(\tilde{v}) = (e^{\tilde{v}_1}/[1+e^{\tilde{v}_1}],\dots,e^{\tilde{v}_p}/[1+e^{\tilde{v}_p}])'.$$

For $v = \tau(\tilde{v})$, θ is a function of v of the form $\theta(v) = \theta(\tau^{-1}(v))$. If the range of $\tau(\tilde{v})$ is an open, convex set, $\theta(v)$ is continuously differentiable to order $m+2$, and and there is a fixed b bounding the derivatives of $\theta(v)$ up to order $m+1$, then Assumption 5.1′ will be satisfied.

To provide primitive conditions for Assumption 5.2 it is necessary to specify the approximating functions. Elbadawi, Souza, and Gallant (1983) show that for Θ and $\|\cdot\|$ in Assumption 5.1′, Fourier series satisfy Assumption 5.2. Here we consider weighted, transformed polynomials. For a p-vector λ of nonnegative integers let $v^\lambda = \Pi_{i=1}^p (v_i)^{\lambda_i}$.

Assumption 5.2′: There are $\omega(v) > 0$ and one-to-one $\tau(v)$ that are $m+2$ times continuously differentiable such that $\theta_J(\gamma) = \omega(v)\sum_{j=1}^J \gamma_j \tau(v)^{\lambda(j)}$, where $(\lambda(1),\lambda(2),\dots)$ includes all p-tuples of nonnegative integers.

Note that typically one would use lower order polynomial terms first, meaning that $\lambda(j)$ with smallest $|\lambda(j)|$ come first.

It is now possible to state a result that imposes more primitive conditions. The following theorem allows for θ to be a vector, each element of which satisfies Assumptions 5.1 and 5.2.

Theorem 5.2: Suppose that $\theta = (\theta_1(v_1),\ldots,\theta_s(v_s))$, *with corresponding* Θ_i *and* $\|\theta_i\|_i$, *satisfying Assumption 5.1′. Also suppose that* $\theta_J(\gamma) = (\theta_{J1}(\gamma_1),\ldots,\theta_{Js}(\gamma_s))$, *with corresponding* $\theta_{Ji}(\gamma_i)$ *satisfying Assumption 5.2′. If, in addition, Assumptions 3.1 and 5.3 - 5.5 hold for* $\Theta = \Theta_1 x\cdots x\Theta_s$ *and* $\|\theta\| = \|\theta_1\|_1 + \cdots + \|\theta_s\|_s$, *then* $\|\hat{\theta}-\theta_0\| \xrightarrow{p} 0$.

6. Work in Progress

The noncontinuity problem encountered in Section 5 suggests that it may be worthwhile to consider models that strenghten the conditional mean restriction, with the goal of relaxing the compactness restriction and obtaining estimators that should be more efficient. One such model is a triangular special case of the first example, taking the form

$$\begin{array}{l} y = \theta_0(z_2) + \varepsilon \\ z_2 = \Pi(x) + V \end{array}, \qquad (\varepsilon,\ V') \text{ independent of } x. \tag{6.1}$$

where $\Pi(z)$ denotes an unknown reduced form function and V a vector of residuals. This model strengthens the assumptions by imposing independence of the disturbances and instruments rather than conditional mean zero. An even more restrictive feature of this model is a reduced form for the right-hand side variables that is linear in disturbances.

A useful implication of this model is the following: By independence of x and (ε, V'),

$$(6.2) \qquad E[y|z_2,V] = \theta_0(z_2) + E[\varepsilon|z_2,V] = \theta_0(z_2) + E[\varepsilon|\Pi(x),V]$$

$$= \theta_0(z_2) + E[\varepsilon|V] = \theta_0(z_2) + \lambda(V).$$

That is, conditional expectation of y given x and V is an additive nonparametric regression model. If V were known, one could estimate (references). Although V is not known, V can be estimated as the residuals $\hat{V}$ from a nonparametric regression of z_2 on x. Then $\theta_0(z_2)$ can be estimated as the z_2 component from an additive nonparametric estimator for the regression of y on z_2 and $\hat{V}$. Preliminary results for this estimator indicate that it is (mean-square) consistent, without the need for any smoothness restrictions. Furthermore, the estimator of the function θ_0 (as well as its derivatives) is asymptotically normal, including an adjustment for the presence of $\hat{V}$.

The restrictive nature of the reduced form for right-hand side endogenous variables suggests that it may be fruitful to investigate other models intermediate between the conditional mean model and the triangular model. One such model is where the disturbance is independent of instruments but the model is not triangular. We are currently investigating this case is, but are not yet ready to report our results.

Other work in progress includes empirical work on the examples that were mentioned. We intend report on all of this work.

Appendix

Throughout the appendix C will denote a generic constant that can be different in different uses. The following restates Corollary 2.1 of Newey (1989), where it is proved.

Lemma A.1: Suppose that i) Θ is a compact set; ii) $\hat{Q}(\theta)$ is continuous and $Q_n(\theta)$ is nonrandom and equicontinuous; iii) for each $\theta \in \Theta$, $\hat{Q}(\theta) - Q_n(\theta) = o_p(1)$; iv) there exists $\epsilon > 0$ and $B_n = O_p(1)$ such that for θ, $\tilde{\theta} \in \Theta$, $|\hat{Q}(\theta)-Q(\theta)| \le B_n\|\tilde{\theta}-\theta\|^{\epsilon}$. Then $\max_{\theta\in\Theta}|\hat{Q}(\theta)-Q(\theta)| \xrightarrow{p} 0$.

The following result is a convergence in probability version of Gallant (1987).

Lemma A.2: Suppose i) $Q(\theta)$ *has a unique minimum on* Θ *at* θ_0; ii) $\hat{Q}(\theta)$ *and* $Q(\theta)$ *are continous,* Θ *is compact, and* $\max_{\theta\in\Theta}|\hat{Q}(\theta)-Q(\theta)| \xrightarrow{p} 0$; iii) $\hat{\Theta}$ *are subsets of* Θ *such that for any* $\theta \in \Theta$ *there exists* $\tilde{\theta} \in \hat{\Theta}$ *such that* $\tilde{\theta} \xrightarrow{p} \theta$. *Then* $\hat{\theta} = \text{argmin}_{\theta\in\hat{\Theta}}\hat{Q}(\theta) \xrightarrow{p} \theta_0$.

Proof: Consider any neighborhood N of θ_0. By compactness, continuity of $Q(\theta)$, and i) (identification),

$$\epsilon \equiv [\min_{\theta\in\Theta\cap N^c} Q(\theta)] - Q(\theta_0) > 0.$$

Using iii), let $\tilde{\theta} \in \hat{\Theta}$ be such that $\tilde{\theta} \xrightarrow{p} \theta_0$. By the definition of $\hat{\theta}$, $\hat{Q}(\hat{\theta}) \le \hat{Q}(\tilde{\theta})$, so that by uniform convergence (see ii)), $Q(\hat{\theta}) < Q(\tilde{\theta}) + \epsilon/2$ with probability approaching one (w.p.a.1). Furthermore, by the definition of $\tilde{\theta}$ and continuity of $Q(\theta)$, $Q(\tilde{\theta}) < Q(\theta_0) + \epsilon/2$ w.p.a.1. Then by the triangle inequality, $Q(\hat{\theta}) < Q(\theta_0) + \epsilon$ w.p.a.1. By the defintion of ϵ, this event can only happen when $\hat{\theta} \in N$, so that $\hat{\theta} \in N$ holds with probability approaching one. The conclusion follows by the arbitrary choice of N. ∎

Proof of Theorem 5.1: The proof will proceed by verifying the hypotheses of Lemma A.2. Consider hypothesis i) first. Define $Q(\theta) \equiv \mathrm{trace}(AE[E[\rho(\theta)|x]E[\rho(\theta)|x]']) = E[E[\rho(\theta)|x]'AE[\rho(\theta)|x]]$, where the z index is dropped from $\rho(z,\theta)$. Note that $Q(\theta_0) = 0$. By Assumption 3.1 and positive definiteness of A, if $\theta \neq \theta_0$, $E[\rho(\theta)|x]'AE[\rho(\theta)|x] > 0$ with positive probability, and hence $Q(\theta) > 0$.

Before proceeding it is useful to note by Θ compact and Assumption 3.1

$$\text{(A.1)} \qquad \|\rho(\theta)\| \le \|\rho(\theta_0)\| + M(z)\|\theta-\theta_0\| \le C\{\|\rho(\theta_0)\| + M(z)\} \equiv \tilde{M}(z),$$

$$E[\tilde{M}(z)^{2+\epsilon}] < \infty.$$

Continuity of $Q(\theta)$ then follows by Assumption 5.3, and the Cauchy-Schwarz and triangle inequalities, which for θ, $\tilde{\theta} \in \Theta$ give

$$\text{(A.2)} \qquad |Q(\tilde{\theta})-Q(\theta)| \le E[|E[\rho(\tilde{\theta})|x]'AE[\rho(\tilde{\theta})-\rho(\theta)|x]|]$$

$$+ E[|E[\rho(\theta)|x]'AE[\rho(\tilde{\theta})-\rho(\theta)|x]|]$$

$$\le \|A\|E[(E[\|\rho(\tilde{\theta})\||x] + E[\|\rho(\theta)\||x])E[\|\rho(\tilde{\theta})-\rho(\theta)\||x]]$$

$$\le C2\{(E[\|\tilde{M}(z)\|^2])^{1/2}\}(E[\|\rho(\tilde{\theta})-\rho(\theta)\|^2])^{1/2} \le C\|\tilde{\theta}-\theta\|.$$

Next, we give some important inequalities on the sample second moment of the conditional expectation estimates. In case a) of Assumption 5.5 iii) it follows by Cauchy Schwarz and (A.1) that

(A.3) $$\sum_{t=1}^{n}\|\hat{E}[\rho(\theta)|x_t]\|^2/n = \sum_{t=1}^{n}\|\sum_{s=1}^{n}w_{st}\rho(z_s,\theta)\|^2/n$$

$$\le \sum_{t=1}^{n}\sum_{s=1}^{n}w_{st}\|\rho(z_s,\theta)\|^2/n \le \sum_{t=1}^{n}\sum_{s=1}^{n}w_{st}\tilde{M}(z)^2/n$$

$$= \sum_{t=1}^{n}\hat{E}[\tilde{M}^2|x_t]/n \equiv \hat{B}_1 = O_p(1),$$

where the last equality follows by $E[(\tilde{M}^2)^{1+\epsilon/2}] = E[\tilde{M}^{2+\epsilon}] < \infty$. Also, it follows similarly by Assumption 5.3 that for θ, $\tilde{\theta} \in \Theta$ that

(A.4) $$\sum_{t=1}^{n}\|\hat{E}[\rho(\tilde{\theta})-\rho(\theta)|x_t]\|^2/n \le (\sum_{t=1}^{n}\hat{E}[M^2|x_t]/n)\|\tilde{\theta}-\theta\|^{2\epsilon}$$

$$\equiv \hat{B}_2\|\tilde{\theta}-\theta\|^{2\epsilon}, \quad \hat{B}_2 = O_p(1).$$

In case b), for $\underset{\sim}{\rho}(\theta) = (\rho(z_1,\theta),\ldots,\rho(z_n,\theta))'$ and $Q = [P_1,\ldots,P_n]'(\sum_{s=1}^{n}P_sP_s')^{-}[P_1,\ldots,P_n]$, by Q idempotent and Markov,

(A.5) $$\sum_{t=1}^{n}\|\hat{E}[\rho(\theta)|x_t]\|^2/n = \|\underset{\sim}{\rho}(\theta)'QQ\underset{\sim}{\rho}(\theta)\|^2/n \le \|\underset{\sim}{\rho}(\theta)'\underset{\sim}{\rho}(\theta)\|^2/n$$

$$\le \sum_{t=1}^{n}\tilde{M}(z_t)^2/n \equiv \hat{B}_1 = O_p(1).$$

Similarly, it follows by Assumption 5.3 that

(A.6) $$\sum_{t=1}^{n}\|\hat{E}[\rho(\tilde{\theta})-\rho(\theta)|x_t]\|^2/n \le \sum_{t=1}^{n}\|\rho(\tilde{\theta})-\rho(\theta)\|^2/n$$

$$\le (\sum_{t=1}^{n}M(z_t)^2/n)\|\tilde{\theta}-\theta\|^{2\epsilon} \equiv \hat{B}_2\|\tilde{\theta}-\theta\|^{2\epsilon}, \quad \hat{B}_2 = O_p(1).$$

Next, note that under Assumption 5.4, $\hat{E}[\psi(z)|z_t]$ is a linear function of $\psi(z_s)$ for any s, so that for θ, $\tilde{\theta} \in \Theta$ it follows by Cauchy Schwarz, and eqs. (A.3) and (A.4) in case a) or eqs. (A.5) and (A.6) in case b) that, (A.7) $$|\hat{Q}(\tilde{\theta}) - \hat{Q}(\theta)|$$

$$\le \|\hat{A}\|\textstyle\sum_{t=1}^{n}(\|\hat{E}[\rho(\tilde{\theta})|x_t]\| + \|\hat{E}[\rho(\theta)|x_t]\|)\hat{E}[\|\rho(\tilde{\theta})-\rho(\theta)\||x_t]/n$$

$$\le \|\hat{A}\|\{(\textstyle\sum_{t=1}^{n}\|\hat{E}[\rho(\tilde{\theta})|x_t]\|^2/n)^{1/2} + (\sum_{t=1}^{n}\|\hat{E}[\rho(\theta)|x_t]\|^2/n)^{1/2}\}\cdot$$

$$(\textstyle\sum_{t=1}^{n}\|\hat{E}[\rho(\tilde{\theta})-\rho(\theta)|x_t]\|^2/n)^{1/2}$$

$$\le \|\hat{A}\|\cdot 2\cdot(\hat{B}_1\hat{B}_2)^{1/2}\|\tilde{\theta}-\theta\|^{\epsilon}, \quad \|\hat{A}\|\cdot 2\cdot(\hat{B}_1\hat{B}_2)^{1/2} = O_p(1).$$

It follows that both hypotheses ii) and iv) of Lemma A.1 are satisfied.

Next, to check hypothesis iii) of Lemma A.1, note that for any $\theta \in \Theta$,

$$|\hat{Q}(\tilde{\theta}) - \textstyle\sum_{t=1}^{n}E[\rho(\theta)|x_t]'\hat{A}E[\rho(\theta)|x_t]/n|$$

$$\le \|\hat{A}\|\textstyle\sum_{t=1}^{n}(\|\hat{E}[\rho(\theta)|x_t]\| + \|E[\rho(\theta)|x_t]\|)\|\hat{E}[\rho(\theta)|x_t]-E[\rho(\theta)|x_t]\|/n$$

$$\le \|\hat{A}\|\{(\textstyle\sum_{t=1}^{n}\|\hat{E}[\rho(\theta)|x_t]\|^2/n)^{1/2} + (\sum_{t=1}^{n}E[\|\rho(\theta)\|^2|x_t]/n)^{1/2}\}\cdot$$

$$(\textstyle\sum_{t=1}^{n}\|\hat{E}[\rho(\theta)|x_t]-E[\rho(\theta)|x_t]\|^2/n)^{1/2}$$

$$\le O_p(1)\{O_p(1) + (\textstyle\sum_{t=1}^{n}E[\tilde{M}^2|x_t]/n)^{1/2}\}o_p(1) = o_p(1).$$

It follows similarly by Assumption 5.2 i) that

$$|\textstyle\sum_{t=1}^{n}E[\rho(\theta)|x_t]'\hat{A}E[\rho(\theta)|x_t]/n - E[E[\rho(\theta)|x_t]'AE[\rho(\theta)|x_t]]| = o_p(1),$$

so that iii) of Lemma A.1 follows by the triangle inequality. Then ii) of Lemma A.2 follows by the conclusion of Lemma A.1.

To check iii) of Lemma A.2, consider $\theta \in \Theta$, and choose γ_J such that $\|\theta_J(\gamma_J)-\theta\| \longrightarrow 0$ as $J \longrightarrow \infty$. By $J \xrightarrow{p} \infty$ it follows that $\|\theta_{\hat{J}}(\gamma_{\hat{J}})-\theta\| \xrightarrow{p} 0$. Also, by the specification of $\hat{\Theta}$ in Assumption 5.5, $\theta_{\hat{J}}(\gamma_{\hat{J}}) \in \hat{\Theta}$, giving iii).

Since we have verified hypotheses i) - iii) of Lemma A.2, the proof follows from the conclusion of Lemma A.2. ■

Proof of Theorem 5.2: The proof proceeds by verifying the hypotheses of Theorem 5.1 for $\Theta = \Pi_{i=1}^{s}\Theta_i$ and $\|\theta\| = \Sigma_{i=1}^{s}\|\theta_i\|_i$. Assumption 5.1 holds for each i, $(i=1,\ldots,s)$, by Lemma 2 of Elbadawi, Gallant and Souza (1983), upon noting that their proof only uses the lower bound on the derivative to show a continuity property that is part of neither the hypotheses or the conclusion here. To check Assumption 5.2, we note that by nonnegativity of $\omega(v)$, $\tau(v)$ one-to-one, and the smoothness assumptions on these functions, it suffices to show the result with $\omega(v) = 1$ and $\tau(v) = v$. By the Corollary of Theorem 1 of Gallant (1981), the set consisting of a finite linear combination of Fourier terms in v is $\|\cdot\|_i$ dense in Θ_i. Therefore, it suffices to show that the $\|\cdot\|_i$ closure of the set of functions in Assumption 5.2′ includes all Fourier terms.

Note that a Fourier term is continuously differentiable to all orders. Also, the abloslute value of partial derivatives are uniformly bounded by $C^{|\lambda|}$, for some constant C and the order of the derivative $|\lambda|$. By convexity of V, it follows that the Taylor expansion of any derivative of a Fourier term around any fixed point will converge uniformly to the derivative of the Fourier term. Furthermore, the Taylor expansion of the deriviative equals the derivative of the Taylor expansion, implying convergence of the Taylor expansion of a Fourier term in the Sobolev norm of Assumption 5.1′. Since the Taylor expansion is a multivariate polynomial, the conclusion follows. ■

References

Deaton, A.S. (1988): "Rice Prices and Income Distribution in Thailand: A Nonparametric Analysis," preprint, Woodrow Wilson School, Princeton University.

Elbadawi, I., A.R. Gallant, and G. Souza (1983): "An Elasticity Can be Estimated Consistently without a Priori Knowledge of Functional Form," *Econometrica*, 51, 1731-1751.

Ferguson, T.S (1967): *Mathematical Statistics: A Decision Theoretic Approach,* New York, Academic Press.

Fisher, F.M. (1976): *The Identification Problem in Econometrics*, Huntington, New York, Krieger Publishing Company.

Gallant, A.R. (1981): "On the Bias in Flexible Functional forms and an Essentially Unbiased Form: The Fourier Flexible Form," *Journal of Econometrics,* 15, 211-245.

Gallant, A.R. (1987): "Identification and Consistency in Nonparametric Regression," in T.F. Bewley, ed., *Advances in Econometrics: Fifth World Congress*, Cambridge: Cambridge University Press, 145-169.

Hansen, L.P. and K.J. Singleton (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50, 1269-1286.

Hardle, W. (1990): *Applied Nonparametric Regression*, in press, Cambridge: Cambridge University Press.

Hausman, J.A., H. Ichimura, W.K. Newey, J.L. Powell (1985): "Semiparametric Identification and Estimation of Polynomial Errors-in-Variables Models," preprint, Department of Economics, University of Wisconsin.

Hausman, J.A., W.K. Newey, and J.L. Powell (1989): "Nonparametric and Semiparametric Estimation of Errors-in-Variables Models," in preparation, Department of Economics, Princeton University.

Lucas, R.E. (1976): "Econometric Policy Evaluation: A Critique," in Brunner, K. and A.H.Meltzer, *The Phillips Curve and Labor Markets,* Carnegie-Rochester Conference Series on Public Policy, vol. 1. Amsterdam: North Holland, 1976.

Malinvaud, E. (1980): *Statistical Methods of Econometrics*, New York: North-Holland.

Newey, W.K. (1989): "Uniform Convergence in Probability and Stochastic Equicontinuity," Econometrics Research Program Working paper, Princeton University.

Newey, W.K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica,* forthcoming.

Nychka, D., G. Wahba, S. Goldfarb, and T. Pugh (1984): "Cross-validated Spline Methods for the Estimation of Three-Dimensional Tumor Size Distributions from Observations on Two-Dimensional Cross-Sections," *Journal of the American Statistical Association*, 79, 832-846.

O'Sullivan, F. (1986): "Ill Posed Inverse Problems (with discussion)," *Statistical Science*, 4, 503-527.

Pagan, A. and Y. Hong (1990): "Nonparametric Estimation and the Risk Premium," forthcoming in Barnett, W., J. Powell, and G. Tauchen eds., *Nonparametric and Semiparametric Methods in Statistics and Econometrics*, Cambridge: Cambridge University Press.

Robinson, P.M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875-891.

Roehrig, C.S. (1988): "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 56, 433-447.

Stone, C.J. (1982): "Consistent Nonparametric Regression" (with discussion), *Annals of Statistics*, 5, 595-645.

Wahba, G. (1979): "Smoothing and Ill-Posed Inverse Problems," in M. Goldberg, ed. *Solution Methods for Integral Equations with Applications*, New York: Plenum, 183-194.