SOME IDENTIFICATION ISSUES IN NONPARAMETRIC LINEAR MODELS WITH ENDOGENOUS REGRESSORS

THOMAS A. SEVERINI
Northwestern University

GAUTAM TRIPATHI
University of Connecticut

In applied work economists often seek to relate a given response variable y to some causal parameter μ^* associated with it. This parameter usually represents a summarization based on some explanatory variables of the distribution of y, such as a regression function, and treating it as a conditional expectation is central to its identification and estimation. However, the interpretation of μ^* as a conditional expectation breaks down if some or all of the explanatory variables are endogenous. This is not a problem when μ^* is modeled as a parametric function of explanatory variables because it is well known how instrumental variables techniques can be used to identify and estimate μ^*. In contrast, handling endogenous regressors in nonparametric models, where μ^* is regarded as fully unknown, presents difficult theoretical and practical challenges. In this paper we consider an endogenous nonparametric model based on a conditional moment restriction. We investigate identification-related properties of this model when the unknown function μ^* belongs to a linear space. We also investigate underidentification of μ^* along with the identification of its linear functionals. Several examples are provided to develop intuition about identification and estimation for endogenous nonparametric regression and related models.

1. INTRODUCTION

Models with endogenous regressors arise frequently in microeconometrics. For example, suppose we want to estimate the cost function of a competitive firm; i.e., we want to estimate the model $y = \mu^*(p,q) + \varepsilon$, where y is the observed cost of production, μ^* the firm's cost function, (p,q) the vector of factor prices and output, and ε an unobserved error term. Because the firm is assumed to be a price taker in its input markets, it is reasonable to assume that the factor prices are exogenously set and are uncorrelated with ε. On the other hand, because an inefficient or high-cost firm will, ceteris paribus, tend to produce less output

than an efficient firm, q may be correlated with ε. Hence, q is endogenous. Similarly, endogenous regressors may also arise in production function estimation. For instance, suppose we want to estimate the model $y = \mu^*(l, k) + \varepsilon$, where y is the firm's output, μ^* the production function, and (l, k) the vector of labor and capital factor inputs. In some cases it may be reasonable to believe that the firm's usage of certain inputs (say, labor) may depend upon the unobserved quality of management. In that case, such factors will be endogenous. Endogeneity can also be encountered in estimating wage equations of the form $y = \mu^*(s, c) + \varepsilon$, where y is log of wage rate, s is the years of schooling, and c denotes agent characteristics such as experience and ethnicity. Because years of schooling are correlated with unobservable factors such as ability and family background, s is endogenous. Another classic example of endogeneity is due to simultaneity. For instance, suppose we want to estimate the market demand for a certain good given by $y = \mu^*(p, d) + \varepsilon$, where y is the quantity demanded in equilibrium, p the equilibrium price, d a vector of demand shifters, and μ^* the market demand function. Because prices and quantities are determined simultaneously in equilibrium, p is endogenous. Several additional examples of regression models with endogenous regressors can be found in econometrics texts; see, e.g., Wooldridge (2002).

These models can be written generically as follows. Let y denote a response variable and x a vector of explanatory variables. Suppose that, corresponding to y, there exist an unknown function $\mu^*(x)$ (we temporarily suppress the dependence of μ^* on y for pedagogical convenience) and an unobservable random variable ε such that $y = \mu^*(x) + \varepsilon$. The parameter of interest in this model is μ^*, and its interpretation in terms of the distribution of (y, x) depends upon the assumptions regarding the joint distribution of x and ε; e.g., if $\mathbb{E}(\varepsilon|x) = 0$ w.p.1 then $\mu^*(x) = \mathbb{E}(y|x)$ w.p.1. In this paper we investigate models defined by more general conditions on the distribution of (x, ε). In particular, we allow some or all of the explanatory variables to be endogenous, i.e., correlated with ε, so that the mean independence of ε and x does not hold.

In the parametric case, i.e., when μ^* is known up to a finite-dimensional parameter, it is well known how to handle endogeneity. Basically, if we have instrumental variables w that suffice to identify μ^*, then we can use two-stage least squares (2SLS), if μ^* is linear, or the more efficient generalized method of moments (GMM) to estimate μ^*. For instance, in the cost function example described previously, the size of the market served by the firm can serve as an instrument for q; in the production function example we could take the wage paid by the firm as an instrument for l if the former is exogenously set; when estimating the wage equation, mother's education can be used to instrument for years of schooling; and, in the market demand example, variables that shift the market supply function but are uncorrelated with ε, such as weather or other exogenous supply shocks, can serve as instruments for p in the demand equation.

Recently there has been a surge of interest in studying nonparametric (i.e., where the functional form of μ^* is completely unknown) models with endog-

enous regressors; see, e.g., Darolles, Florens, and Renault (2002), Ai and Chen (2003), Blundell and Powell (2003), Newey and Powell (2003), and the references therein. In endogenous nonparametric regression models it is typically assumed that μ^* lies in $L_2(x)$, the set of functions of x that are square integrable with respect to the distribution of x, and the instruments w satisfy the conditional moment restriction $\mathbb{E}(\varepsilon|w) = 0$ w.p.1. However, in this paper we allow the parameter space for μ^* to be different from $L_2(x)$ (see Section 2 for the motivation). Hence, our results are applicable to any endogenous nonparametric linear model and not just to the regression models described earlier. Apart from this, the main contributions of our paper are as follows. (i) We develop the properties of the function that maps the reduced form into the structural form in a very general setting under minimal assumptions. For instance, we show that it is a closed map (i.e., its graph is closed) although it may not be continuous. Although lack of continuity of this mapping has been noted in earlier papers, the result that it is closed and further characterization of its continuity properties as done in Lemma 2.4 seem to be new to the literature. (ii) Newey and Powell (2003) characterize identification of μ^* in terms of the completeness of the conditional distribution of x given w. But, in the absence of any parametric assumptions on the conditional distribution of x given w, it is not clear how completeness can be verified in practice. In fact, as Blundell and Powell (2003) point out, the existing literature in this area basically assumes that μ^* is identified and focuses on estimating it. Because failure of identification is not easily detected in nonparametric models (in Section 3 we provide some interesting examples showing that μ^* can be unidentified in relatively simple designs), we investigate what happens if the identification condition for μ^* fails to hold or cannot be easily checked by showing how to determine the "identifiable part" of μ^* by projecting onto an appropriately defined subspace of the parameter space, something that does not seem to have been done earlier in the literature. (iii) In Section 4 we examine the identification of linear functionals of μ^* when μ^* itself may not be identified. We relate identification of μ^* to the identification of its linear functionals by showing that μ^* is identified if and only if all bounded linear functionals of μ^* are identified. To the best of our knowledge, the results in this section are also new to the literature.

We do not focus on estimation in this paper. In addition to the papers mentioned earlier, readers interested in estimating endogenous nonparametric models should see, e.g., Pinkse (2000), Das (2001), Linton, Mammen, Nielsen, and Tanggaard (2001), Carrasco, Florens, and Renault (2002), Florens (2003), Hall and Horowitz (2003), Newey, Powell, and Vella (2003), and the references therein. Additional works related to this literature include Li (1984) and Roehrig (1988). Note that our identification analysis is global in nature because the nonparametric models we consider are linear in μ^*. We hope our results will motivate other researchers to study local properties of nonlinear models of the kind considered by Blundell and Powell (2003) and Newey and Powell (2003).

2. IDENTIFICATION IN A GENERAL SETTING

The introduction was motivated by looking at endogenous nonparametric regression models of the form $y = \mu^*(x) + \varepsilon$, where $\mu^* \in L_2(x)$ and $\mathbb{E}(\varepsilon|w) = 0$ w.p.1. But in many cases the parameter space for μ^* can be a linear function space different than $L_2(x)$. For instance, suppose that $x = (x_1, x_2)$ and μ^* is additive in the components[1] of x; i.e., $\mu^*(x) = \mu_1^*(x_1) + \mu_2^*(x_2)$, where μ_j^* lies in $L_2(x_j)$ for $j = 1,2$. Notice that once μ^* is identified, we can recover the components up to an additive constant by marginal integration; i.e., $\int_{supp(x_2)} \mu^*(x_1, x_2)\, pdf(x_2)\, dx_2 = \mu_1^*(x_1) + \mathbb{E}\mu_2^*(x_2)$ and a similar operation can be carried out to recover μ_2^*. An alternative model may be based on the assumption that $\mu^*(x) = x_1'\theta^* + \mu_2^*(x_2)$, where θ^* is a finite-dimensional parameter and $\mu_2^* \in L_2(x_2)$. This leads to an endogenous version of the partially linear model proposed by Engle, Granger, Rice, and Weiss (1986) and Robinson (1988). Sometimes we may have information regarding the differentiability of μ^* that we want to incorporate into the model; in this case, we might assume that μ^* is an element of a Sobolev space. We could also allow for μ^* to have certain shape restrictions. In particular, because we assume that μ^* belongs to a linear space, shape restrictions such as homogeneity and symmetry are permissible for μ^*. These variations clearly illustrate the advantage of framing our problem in a general setting. So we now frame our problem in a general Hilbert space setting. The geometric nature of Hilbert spaces allows us to derive a lot of mileage from a few relatively simple concepts.

Let y denote the response variable that is assumed to be an element of $\mathcal{U}$, a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|$. Also, let $\mathcal{M}$ denote a known linear subspace of $\mathcal{U}$ (note that $\mathcal{M}$ is not assumed to be closed). Assume that, corresponding to y, there exists an element $\mu_y^* \in \mathcal{M}$. The vector μ_y^* is a summarization of the distribution of y and may be viewed as the parameter of interest. If $y - \mu_y^*$ is orthogonal to $\mathcal{M}$, then μ_y^* is simply the orthogonal projection of y onto $\mathcal{M}$. Here we assume instead that there exists a known linear subspace of $\mathcal{U}$, denoted by $\mathcal{W}$, such that $\langle y - \mu_y^*, w \rangle = 0$ for all $w \in \mathcal{W}$; i.e., $y - \mu_y^*$ is orthogonal to $\mathcal{W}$, which we write as

$$y - \mu_y^* \perp \mathcal{W}. \tag{2.1}$$

We call $\mathcal{M}$ the "model space" and $\mathcal{W}$ the "instrument space." The symbol $\mathcal{Y}$ denotes the set of all $y \in \mathcal{U}$ for which the model holds; i.e., for each $y \in \mathcal{Y}$ there exists a $\mu_y^* \in \mathcal{M}$ such that (2.1) holds. Because there is a one-to-one correspondence between random variables and distribution functions, $\mathcal{Y}$ can also be interpreted as the set of all distributions for which (2.1) holds. Note that because $\mathcal{Y}$ always includes $\mathcal{M}$, it is nonempty. Also, continuity of the inner product implies that whenever (2.1) holds, $y - \mu_y^*$ is orthogonal to the closure of $\mathcal{W}$. Therefore, $\mathcal{W}$ can be assumed to be closed without loss of generality.

Clearly, the endogenous nonparametric regression models described in the introduction are a special case of (2.1) by letting $\mathcal{M} = L_2(x)$, $\mathcal{W} = L_2(w)$, and

$\mathcal{Y}$ be the set of random variables of the form $y = \mu_y^*(x) + \varepsilon$, where $\mu_y^* \in \mathcal{M}$ and $\mathbb{E}(\varepsilon|w) = 0$ w.p.1.

It is easy to see that μ_y^* is identified, i.e., uniquely defined, if and only if the following condition holds.

Condition (I). If $m \in \mathcal{M}$ satisfies $m \perp \mathcal{W}$, then $m = 0$.

Henceforth, we refer to Condition (I) as the "identification condition." Let $P_{\mathcal{W}}$ denote orthogonal projection from $\mathcal{U}$ onto $\mathcal{W}$ using the inner product $\langle \cdot, \cdot \rangle$. Then the identification condition can be alternatively stated as follows: if $m \in \mathcal{M}$ satisfies $P_{\mathcal{W}}m = 0$, then $m = 0$.

Example 2.1 (Linear regression)

Let $\mathcal{U}$ be the familiar Hilbert space of random variables with finite second moments equipped with the usual inner product $\langle u,v \rangle = \mathbb{E}\{uv\}$. Also, x (the $s \times 1$ vector of explanatory variables) and w (the $d \times 1$ vector of instrumental variables) are random vectors whose coordinates are elements of $\mathcal{U}$. Moreover, $\mathcal{M}$ (resp. $\mathcal{W}$) is the linear space spanned by the coordinates of x (resp. w). Note that in this example $\mathcal{M}$ and $\mathcal{W}$ are both finite-dimensional subspaces of $\mathcal{U}$. By (2.1), for a given $y \in \mathcal{U}$ there exists a linear function $\mu_y^*(x) = x'\theta_y^*$ such that $\langle y - x'\theta_y^*, w'\alpha \rangle = 0$ for all $\alpha \in \mathbb{R}^d$; i.e., $\mathbb{E}\{w(y - x'\theta_y^*)\} = 0$. Condition (I) states that if $\langle x'\theta_y^*, w'\alpha \rangle = 0$ for all $\alpha \in \mathbb{R}^d$, then $\theta_y^* = 0$. Hence, μ_y^* or, equivalently, θ_y^* are uniquely defined if and only if $\mathbb{E}wx'$ has full column rank. Obviously, the order condition $d \geq s$ is necessary for $\mathbb{E}wx'$ to have full column rank.

Example 2.2 (Nonparametric regression)

Again, $\mathcal{U}$ is the Hilbert space of random variables with finite second moments equipped with the usual inner product and (x,w) are random vectors whose components are elements of $\mathcal{U}$. But, unlike the previous example, $\mathcal{M} = L_2(x)$ and $\mathcal{W} = L_2(w)$ are now infinite-dimensional linear subspaces of $\mathcal{U}$ consisting of square integrable functions. By (2.1), for a given y in $\mathcal{U}$ there exists a function μ_y^* in $L_2(x)$ such that $\mathbb{E}\{[y - \mu_y^*(x)]g(w)\} = 0$ holds for all $g \in L_2(w)$. Condition (I) states that if a function $f \in L_2(x)$ satisfies $\mathbb{E}\{f(x)g(w)\} = 0$ for all $g \in L_2(w)$, then $f = 0$ w.p.1;[2] i.e., if $\mathbb{E}\{f(x)|w\} = 0$ w.p.1 for any f in $L_2(x)$, then $f = 0$ w.p.1. But this corresponds to the completeness of $pdf(x|w)$. Therefore, μ_y^* is uniquely defined if and only if the conditional distribution of $x|w$ is complete, a result obtained earlier by Florens, Mouchart, and Rolin (1990, Ch. 5) and Newey and Powell (2003). To get some intuition behind the notion of completeness, observe that if x and w are independent, then completeness fails (of course, if w is independent of the regressors then it is not a good instrument and cannot be expected to help identify μ_y^*). On the other extreme, if x is fully predictable by w then completeness is satisfied trivially, and the endogeneity and identification problems disappear altogether. In fact, we can show the following result.

LEMMA 2.1. *The conditional distribution of $x|w$ is complete if and only if for each function $f(x)$ such that $\mathbb{E}f(x) = 0$ and $\operatorname{var} f(x) > 0$, there exists a function $g(w)$ such that $f(x)$ and $g(w)$ are correlated.*

Hence, in the context of nonparametric regression we can think of completeness as a measure of the correlation between the model space $L_2(x)$ and the instrument space $L_2(w)$.

Let us assume that Condition (I) holds for the remainder of Section 2. Hence, for each $y \in \mathcal{Y}$ there exists a unique $\mu_y^* \in \mathcal{M}$ such that $y - \mu_y^* \perp \mathcal{W}$. It follows that $\mathcal{Y}$ is a linear subspace of $\mathcal{U}$ and the map $y \mapsto \mu_y^*$ is a linear transformation on $\mathcal{Y}$. Therefore, from now on we write Vy for μ_y^* so that $V: \mathcal{Y} \to \mathcal{M}$ denotes a linear map such that $\mu_y^* = Vy$. Employing well-known terminology, V is just the function that maps the reduced form into the structural form. Hence, a clear description of the properties of V is central to understanding the identification and estimation problems in nonparametric linear models with endogenous regressors.

We now study the properties of V. Define $\mathcal{W}_0 = \{w \in \mathcal{W}: w = P_{\mathcal{W}}m$ for some $m \in \mathcal{M}\}$. Because it is straightforward to show that $\mathcal{W}_0$ is the smallest linear subspace of $\mathcal{W}$ satisfying Condition (I), we may view $\mathcal{W}_0$ as the "minimal" instrument space. Let $y \in \mathcal{Y}$. Because $Vy \in \mathcal{M}$, by definition of $\mathcal{W}_0$ we know that $P_{\mathcal{W}}Vy \in \mathcal{W}_0$. But, letting I denote the identity operator, we can write $y = Vy + (I - V)y$, where $Vy \in \mathcal{M}$ and $(I - V)y \perp \mathcal{W}$. Hence, $P_{\mathcal{W}}y = P_{\mathcal{W}}Vy \in \mathcal{W}_0$. Furthermore, because $y - P_{\mathcal{W}}y \perp \mathcal{W}$ and $\mathcal{W}_0 \subseteq \mathcal{W}$, we have $y - P_{\mathcal{W}}y \perp \mathcal{W}_0$. This shows that when applied to elements of $\mathcal{Y}$, the projection $P_{\mathcal{W}}$ has the same properties as orthogonal projection on $\mathcal{W}_0$. Next, let $\bar{P}_{\mathcal{W}}: \mathcal{M} \to \mathcal{W}_0$ denote the restriction of $P_{\mathcal{W}}$ to $\mathcal{M}$. Then $\bar{P}_{\mathcal{W}}$ is a continuous linear mapping from $\mathcal{M}$ to $\mathcal{W}_0$ with inverse[3] $\bar{P}_{\mathcal{W}}^{-1}$. Clearly, $\bar{P}_{\mathcal{W}}^{-1}$ is also a linear map. Therefore, we can characterize V as

$$V = \bar{P}_{\mathcal{W}}^{-1} P_{\mathcal{W}}. \tag{2.2}$$

The next example describes how V looks in some familiar settings.

Example 2.3

In Example 2.1, $\mathcal{W}$ is the linear space spanned by the coordinates of w. Hence, $P_{\mathcal{W}}$ corresponds to the best linear predictor given w; i.e., $(P_{\mathcal{W}}y)(w) = (\mathbb{E}yw')$ $(\mathbb{E}ww')^{-1}w$. It is easy to show that $(\bar{P}_{\mathcal{W}}^{-1}w)(x) = (\mathbb{E}wx')\{(\mathbb{E}xw')(\mathbb{E}ww')^{-1}$ $(\mathbb{E}wx')\}^{-1}x$. Therefore, the map $V: \mathcal{Y} \to \mathcal{M}$ is given by $(Vy)(x) = (\bar{P}_{\mathcal{W}}^{-1}P_{\mathcal{W}}y)$ $(x) = (\mathbb{E}yw')(\mathbb{E}ww')^{-1}(\mathbb{E}wx')\{(\mathbb{E}xw')(\mathbb{E}ww')^{-1}(\mathbb{E}wx')\}^{-1}x$. But because $(Vy)(x)$ can be written as $x'\theta_y^*$, it follows that θ_y^* here is just the population version of the usual 2SLS estimator. By contrast, $\mathcal{W}$ in Example 2.2 is the infinite-dimensional space $L_2(w)$. Hence, $P_{\mathcal{W}}$ is the best prediction operator $(P_{\mathcal{W}}y)(w) = \mathbb{E}(y|w)$. Therefore, in Example 2.2 we have $Vy = \bar{P}_{\mathcal{W}}^{-1}\mathbb{E}(y|w)$.

Before describing additional properties of V, in Lemma 2.2 we propose a series-based approach for determining V. As illustrated by the examples given subsequently, this approach may also be useful as the basis of a practical computational method for estimating V. However, as noted earlier, a full consideration of estimation issues is beyond the scope of the current paper. Instead, the reader is referred to Pinkse (2000), Darolles et al. (2002), Ai and Chen (2003), Hall and Horowitz (2003), and Newey and Powell (2003) for series estimation of endogenous nonparametric models.

LEMMA 2.2. *Let $m_0, m_1, m_2, \ldots$ be a basis for $\mathcal{M}$ such that $\langle m_i, P_{\mathcal{W}} m_j \rangle = 0$ for $i \neq j$. Then,*

$$\bar{P}_{\mathcal{W}}^{-1} w = \sum_{j=0}^{\infty} \frac{\langle m_j, w \rangle}{\langle m_j, P_{\mathcal{W}} m_j \rangle} m_j \quad \text{for any } w \in \mathcal{W}_0 \quad \text{and}$$

$$Vy = \sum_{j=0}^{\infty} \frac{\langle m_j, P_{\mathcal{W}} y \rangle}{\langle m_j, P_{\mathcal{W}} m_j \rangle} m_j \quad \text{for any } y \in \mathcal{Y}.$$

This result is similar in spirit to the eigenvector-based decomposition of Darolles et al. (2002) although we use a different basis in our representation. It demonstrates that if μ_y^* is identified then it can be explicitly characterized in the population by a series representation using a special set of basis vectors (if $\mathcal{M} = \mathcal{W}$ so that endogeneity disappears, then Vy is just the projection onto $\mathcal{M}$ as expected). The basis functions needed in Lemma 2.2 can be constructed from an arbitrary basis by using the well-known Gram–Schmidt procedure as follows. Let $d_0, d_1, d_2, \ldots$ be a basis for $\mathcal{M}$. Define $m_0 = d_0$ and let

$$m_j = d_j - \sum_{k=0}^{j-1} \frac{\langle d_j, P_{\mathcal{W}} m_k \rangle}{\langle m_k, P_{\mathcal{W}} m_k \rangle} m_k \quad \text{for } j = 1, 2, 3, \ldots. \tag{2.3}$$

Then $m_0, m_1, m_2, \ldots$ is a basis for $\mathcal{M}$ satisfying $\langle m_i, P_{\mathcal{W}} m_j \rangle = 0$ for $i \neq j$.

The following example illustrates the usefulness of Lemma 2.2.

Example 2.4

Let x, w, and ε be real-valued random variables such that x and ε are correlated, $\mathbb{E}(\varepsilon | w) = 0$ w.p.1, and (x, w) has a bivariate normal distribution with mean zero and variance $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ where $\rho \in (-1, 1) \backslash \{0\}$. Suppose $y = \mu_y^*(x) + \varepsilon$, where μ_y^* is unknown and $\mathbb{E} \mu_y^{*2}(x) < \infty$. Because $x | w \stackrel{d}{\sim} N(w\rho, 1 - \rho^2)$, the conditional distribution of x given w is complete. Hence, μ_y^* is identified. Now let ϕ be the standard normal probability density function (p.d.f.) and

$$H_0(x) = 1, \qquad H_1(x) = x, \qquad H_2(x) = x^2 - 1, \ldots,$$

$$H_j(x) = (-1)^{j+2} \frac{\phi^{(j)}(x)}{\phi(x)}, \ldots$$

denote Hermite polynomials that are orthogonal with respect to the usual inner product on $L_2(x)$. From Granger and Newbold (1976, p. 202) we know that if $\begin{bmatrix} x \\ w \end{bmatrix} \stackrel{d}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$, then $\mathbb{E}\{H_j(x)|w\} = \rho^j H_j(w)$. This result ensures that the Hermite basis satisfies the requirement in Lemma 2.2.[4] It, plus the facts that $\mathbb{E}H_j^2(x) = j!$ and $\mathbb{E}\{H_j(x)\mathbb{E}[H_j(x)|w]\} = \rho^{2j}j!$, shows that we can write μ_y^* explicitly as

$$\mu_y^*(x) = \sum_{j=0}^{\infty} \frac{\mathbb{E}\{yH_j(w)\}}{j!\rho^j} H_j(x). \tag{2.4}$$

There are some interesting consequences of (2.4). For instance, if $\mathbb{E}(y|w)$ happens to be a polynomial of degree p, then μ_y^* will also be a polynomial of degree p because $\mathbb{E}\{w^p H_j(w)\} = 0$ for all $j > p$. As a particular example, suppose that $\mathbb{E}(y|w) = a + bw + cw^2$. Then it is easily seen that $\mu_y^*(x) = a - c/\rho^2 + bx/\rho + cx^2/\rho^2$. It is also clear from (2.4) that an estimator for μ_y^* can be based on the truncated series for Vy. This is discussed in the next example.

Example 2.5 (Example 2.4 cont.)

As mentioned earlier, an estimator for μ_y^* can be obtained by truncating the series in (2.4). Suppose we have a random sample $(y_1, x_1, w_1), \ldots, (y_n, x_n, w_n)$ from the distribution of (y, x, w). Let $\hat{\gamma}_j$ denote the sample analog of $\gamma_j = \mathbb{E}\{yH_j(w)\}$ based on these observations; i.e., $\hat{\gamma}_j = \sum_{i=1}^{n} y_i H_j(w_i)/n$. By (2.4), an estimator of μ_y^* is given by

$$\hat{\mu}_n(x) = \sum_{j=0}^{k_n} \frac{\hat{\gamma}_j H_j(x)}{j!\rho^j},$$

where k_n is a function of the sample size such that $k_n \uparrow \infty$ as $n \uparrow \infty$. In this example we show that $\hat{\mu}_n$ is mean-square consistent and derive its rate of convergence. Suppose for convenience that $w \mapsto \mathbb{E}(y|w)$ and $w \mapsto \mathrm{var}(y|w)$ are bounded. Then, as shown in the Appendix, for some $\alpha > 0$ the mean integrated squared error (MISE) of $\hat{\mu}_n$ is given by

$$\mathbb{E}\int_{\mathbb{R}} \{\hat{\mu}_n(x) - \mu_y^*(x)\}^2 \phi(x)\,dx = O\left(\frac{\rho^{-2k_n}k_n}{n} + k_n^{-\alpha}\right). \tag{2.5}$$

Although (2.5) holds for a stylized setup, it is very informative; e.g., it is clear that the MISE is asymptotically negligible if $k_n \uparrow \infty$ sufficiently slowly. Hence, $\hat{\mu}_n$ is mean-square consistent for μ_y^*, though its rate of convergence is slow. It converges even more slowly if the instrument is "weak," i.e., if $|\rho|$ is small. In fact, because the MISE converges to zero if and only if $k_n \log \rho^{-2} + \log k_n - \log n \downarrow -\infty$, it follows that k_n must be $O(\log n)$ or smaller. Therefore, even in this simple setting where the joint normality of regressors and instruments is known and imposed in constructing an estimator, the best attainable rate of decrease for the MISE is only $O(\{\log n\}^{-\alpha})$. This suggests that rates of

convergence that are powers of $1/\log n$, rather than $1/n$, are relevant for endogenous nonparametric regression models when the distribution of (x, w) is unknown. Rates better than $O(\{\log n\}^{-\alpha})$ can be obtained by imposing additional restrictions on μ_y^*; e.g., Darolles et al. (2002, Thm. 4.2) and Hall and Horowitz (2003, Thm. 4.1) achieve faster rates by making the eigenvalues of certain integral operators decay to zero at a fast enough rate, thereby further restricting μ_y^* implicitly.

Example 2.6 (Endogenous nonparametric additive regression)

Let $y = \mu_1^*(x) + \mu_2^*(z) + \varepsilon$, where μ_1^* and μ_2^* are unknown functions such that $\mathbb{E}\mu_1^{*2}(x) + \mathbb{E}\mu_2^{*2}(z) < \infty$ and $\mathbb{E}(\varepsilon|w, z) = 0$; i.e., x is the only endogenous regressor. In this example, the model space is $L_2(x) + L_2(z)$, and the instrument space is $L_2(w, z)$. Assume that (x, z, w) is trivariate normal with mean zero and positive definite variance-covariance matrix $\Omega = \begin{bmatrix} 1 & \rho_{xz} & \rho_{xw} \\ \rho_{xz} & 1 & \rho_{zw} \\ \rho_{xw} & \rho_{zw} & 1 \end{bmatrix}$. Because the conditional distribution of x given (w, z) is normal with mean depending on (w, z) and the family of one-dimensional Gaussian distributions with varying mean is complete, $\mu_1^*(x) + \mu_2^*(z)$ is identified. We now use the approach of Lemma 2.2 to recover μ_1^* and μ_2^*. Let $\mu^*(x, z) = \mu_1^*(x) + \mu_2^*(z)$. Note that $\mu^*(x, z) = \sum_{j=0}^{\infty} \alpha_j H_j(x) + \sum_{j=0}^{\infty} \beta_j H_j(z)$ for constants $\{\alpha_j\}_{j=0}^{\infty}$ and $\{\beta_j\}_{j=0}^{\infty}$. But because $\mathbb{E}\{\varepsilon|w\} = 0$ and $\mathbb{E}\{\varepsilon|z\} = 0$, we have $\mathbb{E}\{yH_j(w)\} = \alpha_j \rho_{xw}^j j! + \beta_j j!$ and $\mathbb{E}\{yH_j(z)\} = \alpha_j \rho_{xz}^j j! + \beta_j j!$. Solving these simultaneous equations for each j, it follows that

$$\mu^*(x, z) = \alpha_0 + \beta_0 + \sum_{j=1}^{\infty} \frac{\mathbb{E}\{yH_j(w)\} - \rho_{zw}^j \mathbb{E}\{yH_j(z)\}}{j!(\rho_{xw}^j - \rho_{xz}^j \rho_{zw}^j)} H_j(x)$$

$$+ \sum_{j=1}^{\infty} \frac{\rho_{xw}^j \mathbb{E}\{yH_j(z)\} - \rho_{xz}^j \mathbb{E}\{yH_j(w)\}}{j!(\rho_{xw}^j - \rho_{xz}^j \rho_{zw}^j)} H_j(z).$$

Therefore, using the fact that $\mathbb{E}H_j(x) = 0$ and $\mathbb{E}H_j(z) = 0$ for $j \geq 1$,

$$\mu_1^*(x) = \mathbb{E}y + \sum_{j=1}^{\infty} \frac{\mathbb{E}\{yH_j(w)\} - \rho_{zw}^j \mathbb{E}\{yH_j(z)\}}{j!(\rho_{xw}^j - \rho_{xz}^j \rho_{zw}^j)} H_j(x)$$

and

$$\mu_2^*(z) = \mathbb{E}y + \sum_{j=1}^{\infty} \frac{\rho_{xw}^j \mathbb{E}\{yH_j(z)\} - \rho_{xz}^j \mathbb{E}\{yH_j(w)\}}{j!(\rho_{xw}^j - \rho_{xz}^j \rho_{zw}^j)} H_j(z).$$

Hence, μ_1^* and μ_2^* are identified.

Next, we consider an iterative scheme for determining V.[5] The advantage of this approach is that we do not have to explicitly calculate the inverse operator

$\bar{P}_{\mathcal{W}}^{-1}$. We only need $P_{\mathcal{W}}$ and $P_{\mathcal{M}}$, where the latter denotes orthogonal projection onto the closure of $\mathcal{M}$. In contrast, the series approach of Lemma 2.2 did not require any knowledge of $P_{\mathcal{M}}$.

LEMMA 2.3. *Fix* $w \in \mathcal{W}_0$ *and consider the equation* $\bar{P}_{\mathcal{W}} m = w$ *for* $m \in \mathcal{M}$. *Let* m_0 *denote its solution and define* $m_1 = P_{\mathcal{M}} w$. *If there exist a constant* $a \neq 0$ *and an* $m^* \in \mathcal{M}$ *such that* $m_{n+1} = (I - a P_{\mathcal{M}} \bar{P}_{\mathcal{W}}) m_n + a P_{\mathcal{M}} w$ *converges to* m^* *as* $n \uparrow \infty$, *then* $m^* = m_0$.

This result, which is related to the Landweber–Fridman procedure described in Kress (1999, Ch. 15) and Carrasco et al. (2002), shows that if the sequence m_n converges, then it converges to m_0. Therefore, given y, we can obtain Vy by applying this procedure to $w = P_{\mathcal{W}} y$. Because $m_n = a \sum_{j=0}^{n} (I - a P_{\mathcal{M}} \bar{P}_{\mathcal{W}})^j P_{\mathcal{M}} w$, convergence in Lemma 2.3 is ensured if there exists a nonzero constant a such that the partial sum $\sum_{j=0}^{n} (I - a P_{\mathcal{M}} \bar{P}_{\mathcal{W}})^j m$ converges pointwise for each $m \in \mathcal{M}$. A well-known sufficient condition for this to happen is that $\sup_{\{m \in \mathcal{M} : \|m\|=1\}} \|(I - a P_{\mathcal{M}} \bar{P}_{\mathcal{W}}) m\| < 1$. Of course, if $\mathcal{M} = \mathcal{W}$ so that there is no endogeneity, then $m_1 = P_{\mathcal{M}} y$, and there are no further adjustments to m_1.

Example 2.7

The iterative procedure of Lemma 2.3 also works for Example 2.4. To see this, let $a = 1$ and note that $\mathbb{E}(y|w) = \sum_{j=0}^{\infty} \beta_j H_j(w)$, where $\beta_j = \mathbb{E}\{y H_j(w)\}/j!$. Hence, $m_n(x) = \sum_{j=0}^{\infty} \beta_j \rho^j \{1 + (1 - \rho^{2j}) + (1 - \rho^{2j})^2 + \cdots + (1 - \rho^{2j})^{n-1}\} H_j(x)$, and by (2.4) it follows that $\int_{\mathbb{R}} \{m_n(x) - \sum_{j=0}^{\infty} \beta_j H_j(x)/\rho^j\}^2 \phi(x) \, dx \to 0$ as $n \uparrow \infty$; i.e., m_n converges in mean-square to μ_y^*.

Before ending this section, we comment briefly on the pervasiveness of "ill-posed" endogenous nonparametric models. Recall that Condition (I) guarantees that for each $y \in \mathcal{Y}$ the vector μ_y^* is uniquely defined, i.e., $V: y \mapsto \mu_y^*$ is a function from $\mathcal{Y}$ into $\mathcal{M}$. But Condition (I) is not strong enough to ensure that this function is continuous;[6] i.e., the identification condition by itself is not strong enough to ensure that the problem is well-posed. However, it can be shown that V is a closed linear operator. To see this, let $y_1, y_2, \ldots$ denote a sequence in $\mathcal{Y}$ such that $y_n \to y \in \mathcal{U}$ as $n \uparrow \infty$ and suppose that $V y_n \to m \in \mathcal{M}$ as $n \uparrow \infty$. To show that V is closed, it suffices to show that $y \in \mathcal{Y}$ and $m = Vy$. Note that, for each $n = 1, 2, \ldots, y_n - V y_n \perp \mathcal{W}$ and, because $y_n - V y_n \to y - m$ as $n \uparrow \infty$, $y - m \perp \mathcal{W}$. Hence, by definition of $\mathcal{Y}$ and V, $y \in \mathcal{Y}$ and $m = Vy$. The next result characterizes the continuity of V.

LEMMA 2.4. *The following statements are equivalent: (i)* V *is continuous on* $\mathcal{Y}$; *(ii)* $\mathcal{Y}$ *is closed; (iii) if* $m_1, m_2, \ldots$ *is a sequence in* $\mathcal{M}$ *such that* $P_{\mathcal{W}} m_n \to 0$ *as* $n \uparrow \infty$, *then* $m_n \to 0$ *as* $n \uparrow \infty$; *(iv)* $\mathcal{W}_0$ *is closed; (v) there exists a closed linear subspace* $\mathcal{Y}_0$ *of* $\mathcal{Y}$ *such that* $\mathcal{W}_0 \subseteq \mathcal{Y}_0$.

The restrictive nature of this lemma reveals that well-posed endogenous non-parametric models are an exception rather than the rule; e.g., even the simple Gaussian setting of Example 2.4 is not sufficient to make the problem there well-posed. To see this, let $f_n(x) = H_n(x)/\sqrt{n!}$ denote the normalized nth Hermite polynomial. It is then easy to verify that $\mathbb{E}\{f_n(x)|w\}$ converges to zero in mean-square whereas f_n does not. Therefore, (iii) does not hold, and, hence, V is not continuous. Of course, if $\mathcal{M}$ is finite-dimensional (as in parametric models, or, in nonparametric models where the regressors are discrete random variables with finite support),[7] then (iii) holds and V is continuous. Similarly, if $\mathcal{W}$ is finite-dimensional then $\mathcal{W}_0$ will be finite-dimensional and, hence, closed, implying that V is continuous. But these are clearly very special cases. A practical consequence of ill-posedness is that some type of "regularization" is needed in estimation procedures to produce estimators with good asymptotic properties. For instance, a truncation-based regularization ensures convergence of the estimator described in Example 2.5. For more about the different regularization schemes used in the literature, see, e.g., Wahba (1990, Ch. 8), Kress (1999, Ch. 15), Carrasco et al. (2002), Loubes and Vanhems (2003), and the references therein.

3. UNDERIDENTIFICATION

In this section, we investigate the case where μ_y^* in (2.1) fails to be uniquely defined. As mentioned earlier in Example 2.2, Newey and Powell (2003) and others have characterized identification of the endogenous nonparametric regression model in terms of completeness of the conditional distribution of x given w. They also point out that it is sufficient to restrict $pdf(x|w)$ to the class of full rank exponential densities for it to be complete. However, Examples 3.2 and 3.3 illustrate that this sufficient condition can fail to hold in relatively simple cases. Furthermore, if the distribution of $x|w$ is not assumed to be parametric, completeness can be very hard to verify. Hence, it is important to know what happens when completeness fails or cannot be checked. We now focus on this issue.

Let $\mathcal{M}_0 = \{m \in \mathcal{M} : m \perp \mathcal{W}\}$ be the set of all "identification-destroying" perturbations of μ_y^*. From Condition (I) it follows that μ_y^* is identified if and only if $\mathcal{M}_0 = \{0\}$. Note that $\mathcal{M}_0$ is a closed linear subspace of $\mathcal{M}$. The properties of $\mathcal{M}_0$ play an important role in the identification of μ_y^*.

Example 3.1 (Underidentification in linear regression)

We maintain the setup of Example 2.1. For linear instrumental variables regression it is easily seen that $\mathcal{M}_0 = \{x'\theta : (\mathbb{E}wx')\theta = 0 \text{ for some } \theta \in \mathbb{R}^s\}$. Hence, the identification condition fails to hold, i.e., $\mathcal{M}_0 \neq \{0\}$, if $\mathbb{E}wx'$ is not of full column rank.

Example 3.2 (Underidentification in nonparametric regression)

Let $y = \mu_y^*(x) + \varepsilon$, where $\mu_y^* \in L_2(x)$ is unknown. The regressor is endogenous, but we have an instrument w satisfying $\mathbb{E}(\varepsilon|w) = 0$ w.p.1. Suppose that $x = w + v$, where $w, v \overset{iid}{\sim} \text{Uniform}[-\frac{1}{2}, \frac{1}{2}]$. Hence, $\mathcal{M}_0 = \{f \in L_2(x): \mathbb{E}[f(x)|w] = 0$ for a.a. $w \in [-\frac{1}{2}, \frac{1}{2}]\}$. Because $\mathbb{E}\{f(x)|w\} = \int_{w-1/2}^{w+1/2} f(u)\, du$, it is straightforward to show that $\mathbb{E}\{f(x)|w\} = 0$ holds for a.a. $w \in [-\frac{1}{2}, \frac{1}{2}]$ if and only if f is periodic in the sense that $f(x) = f(1 + x)$ for a.a. $x \in [-1, 0]$ and $\int_{-1}^{0} f(x)\, dx = 0$. Thus, $\mathcal{M}_0$ can be explicitly characterized as $\mathcal{M}_0 = \{f \in L_2(x): f(x) = f(1 + x)$ for a.a. $x \in [-1, 0]$ and $\int_{-1}^{0} f(x)\, dx = 0\}$. Because $\mathcal{M}_0$ is clearly not equal to $\{0\}$, Condition (I) does not hold. Therefore, μ_y^* is not uniquely defined and, hence, cannot be estimated even for the simple design given in this example.

Example 3.3 (Underidentification in nonparametric additive regression)

Let $y = \mu_1^*(x) + \mu_2^*(z) + \varepsilon$, where μ_1^* and μ_2^* are unknown functions $L_2(x)$ and $L_2(z)$, respectively, and $\mathbb{E}(\varepsilon|w) = 0$; i.e., both x and z are endogenous, but we only have one instrument w. Obviously, here the model space is $L_2(x) + L_2(z)$, but the instrument space is $L_2(w)$. As in Example 2.6, assume that (x, z, w) are jointly normal with mean zero and variance Ω. Because the conditional distribution of $x, z|w$ is not complete, it follows that $\mu_1^*(x) + \mu_2^*(z)$ is not identified. In fact, it can be shown that $\mathcal{M}_0 = \mathcal{Q} \neq \{0\}$, where $\mathcal{Q} = \text{span}\{Q_0(x, z), Q_1(x, z), \ldots, Q_j(x, z), \ldots\}$ and $Q_j(x, z) = \rho_{zw}^j H_j(x) - \rho_{xw}^j H_j(z)$.[8]

Suppose that a μ_y^* satisfying (2.1) is not uniquely defined. Loosely speaking, this means that the model space is "too large"; i.e., it contains more than one element satisfying (2.1). Hence, to obtain identifiability, we may choose a smaller model space. This approach is analogous to eliminating redundant regressors in an underidentified linear regression model. We now formalize this intuition. For a given $y \in \mathcal{Y}$, define $\mathcal{M}_y = \{m \in \mathcal{M} : y - m \perp \mathcal{W}\}$. Identification holds when $\mathcal{M}_y$ consists of a single element. Otherwise, $\mathcal{M}_y$ is a collection of elements that cannot be distinguished based on (2.1). A nice property of $\mathcal{M}_y$ is that each of its elements has the same projection onto $\mathcal{M}_0^\perp$, the orthogonal complement of $\mathcal{M}_0$. Hence, $\mathcal{M}_0^\perp$ is a natural choice for the reduced model space;[9] i.e., if μ_y^{**} is the orthogonal projection of an arbitrarily chosen element of $\mathcal{M}_y$ onto $\mathcal{M}_0^\perp$, then μ_y^{**} can be regarded as the "identifiable part" of μ_y^*. In technical terms, when Condition (I) does not hold, the "true parameter" of the model is, in effect, an equivalence class of elements of $\mathcal{M}$, which we have denoted by the symbol $\mathcal{M}_y$. This class of true parameters may be described in terms of their common features as follows. Because $\mathcal{M}_0^\perp = (I - P_{\mathcal{M}_0})\mathcal{M}$, each $m_y \in \mathcal{M}_y$ may be decomposed into two components $P_{\mathcal{M}_0} m_y$ and $P_{\mathcal{M}_0^\perp} m_y$. But because $P_{\mathcal{M}_0^\perp} m_y$ is the same for all $m_y \in \mathcal{M}_y$, each equivalence class $\mathcal{M}_y$ may be described by a single element μ_y^{**}, which we refer to as the identifiable part of μ_y^*. We may take this canonical element to be $\mu_y^{**} = P_{\mathcal{M}_0^\perp} \mathcal{M}_y$. It is easy to

show that μ_y^{**} is an element of $\mathcal{M}_y$.[10] The remaining elements of $\mathcal{M}_y$ are those $m \in \mathcal{M}$ such that $P_{\mathcal{M}_0^{\perp}} m = P_{\mathcal{M}_0^{\perp}} \mu_y^{**}$; i.e., all $m \in \mathcal{M}$ of the form $\mu_y^{**} + \mathcal{M}_0$.

Example 3.4 (Example 3.1 cont.)

Suppose that $\mathbb{E}wx'$ is not of full column rank so the identification condition fails to hold. Here, $\mathcal{M}_y = \{x'\theta : (\mathbb{E}wx')\theta = \mathbb{E}wy\}$. Because $P_{\mathcal{M}_0}(x'\theta) = x'\theta - x'A\theta$, where $A = (\mathbb{E}xx')^{-1}(\mathbb{E}xw')\{(\mathbb{E}wx')(\mathbb{E}xx')^{-1}(\mathbb{E}xw')\}^{-1}(\mathbb{E}wx')$, we have $P_{\mathcal{M}_0^{\perp}}(x'\theta) = x'A\theta$. Hence, we can only identify linear functions of the form $x'A\theta$. Of course, if the identification condition holds, i.e., $\mathbb{E}wx'$ is of full column rank, then A reduces to the identity matrix and $\mathcal{M}_y = \{x'\theta^*\}$ with θ^* as defined in Example 2.3.

Example 3.5 (Example 3.2 cont.)

The identifiable part of μ_y^* is given by projecting $\mathcal{M}_y$ onto $\mathcal{M}_0^{\perp}$, where $\mathcal{M}_y = \{f \in L_2(x) : \int_{w-1/2}^{w+1/2} f(u) \, du = \mathbb{E}(y|w)$ for a.a. $w \in [-\frac{1}{2}, \frac{1}{2}]\}$. Recall that x has the triangular distribution on $[-1,1]$; i.e., the p.d.f. of x is given by $h(x) = 1 + x$ for $-1 \le x \le 0$ and $h(x) = 1 - x$ for $0 \le x \le 1$. It can be shown that $\mathcal{M}_0^{\perp} = \mathcal{B}$, where[11]

$$\mathcal{B} = \{f \in L_2(x) : f(x)h(x) = -f(x+1)h(x+1) + c$$

$$\text{for a.a. } x \in [-1,0] \text{ and some constant } c\}.$$

Hence, the identifiable part of μ_y^* is a function f satisfying $f(x)h(x) = -f(x+1)h(x+1) + c$ for a.a. $x \in [-1,0]$ and some constant c.

Example 3.6 (Example 3.3 cont.)

Now let us determine the identifiable part of the underidentified model in Example 3.3. It can be shown that $\mathcal{M}_0^{\perp} = \mathcal{A}$, where $\mathcal{A} = \text{span}\{1, A_1(x,z), \ldots, A_j(x,z), \ldots\}$ and $A_j(x,z) = (\rho_{xw}^j - \rho_{xz}^j \rho_{zw}^j)H_j(x) + (\rho_{zw}^j - \rho_{xz}^j \rho_{xw}^j)H_j(z)$.[12] Hence, we can identify only those additive functions whose Hermite representation is of the form $c + \sum_{j=1}^{\infty} \gamma_j A_j(x,z)$, where c denotes a constant. For example, suppose that $\rho_{xw} = \rho_{zw}$; i.e., the instrument has the same correlation with each regressor. Then $\mathcal{M}_0^{\perp}$ consists of functions of the form $c + \sum_{j=0}^{\infty} \gamma_j H_j(x) + \sum_{j=0}^{\infty} \gamma_j H_j(z)$; i.e., only elements of $L_2(x) + L_2(z)$ of the form $c + f(x) + f(z)$ are identified.

As mentioned earlier, underidentification may be viewed as a consequence of the fact that the model space $\mathcal{M}$ is too big. Hence, to obtain identifiability, we may choose a smaller model space. A natural choice for this reduced model space is $\mathcal{M}_0^{\perp}$. But to use $\mathcal{M}_0^{\perp}$ in place of $\mathcal{M}$, we must verify that it satisfies two conditions. The first is that for each $y \in \mathcal{Y}$ there exists a $\mu_y^{**} \in \mathcal{M}_0^{\perp}$ such that $y - \mu_y^{**} \perp \mathcal{W}$. The second is that $\mathcal{M}_0^{\perp}$ satisfies Condition (I). It is easy to see that both these conditions are satisfied: Fix $y \in \mathcal{Y}$. Then by (2.1), there exists $\mu_y \in \mathcal{M}$ such that $y - \mu_y \perp \mathcal{W}$. Let $\mu_y^{**} = (I - P_{\mathcal{M}_0})\mu_y = P_{\mathcal{M}_0^{\perp}} \mu_y$.

Then $y - \mu_y^{**} = y - \mu_y + P_{\mathcal{M}_0}\mu_y$. Because $P_{\mathcal{M}_0}\mu_y$ is an element of $\mathcal{M}_0$ and, hence, orthogonal to $\mathcal{W}$, it follows that $y - \mu_y^{**} \perp \mathcal{W}$. This shows that $\mathcal{M}_0^{\perp}$ satisfies the first requirement. Next, let m denote an element of $\mathcal{M}_0^{\perp}$ that is orthogonal to $\mathcal{W}$. By definition, there exists $m_1 \in \mathcal{M}$ such that $m = (I - P_{\mathcal{M}_0})m_1$. Because $m \perp \mathcal{W}$, for any $w \in \mathcal{W}$ we have $0 = \langle (I - P_{\mathcal{M}_0})m_1, w \rangle = \langle m_1, (I - P_{\mathcal{M}_0})w \rangle = \langle m_1, w \rangle$. It follows that $m_1 \perp \mathcal{W}$ and, hence, that $m = (I - P_{\mathcal{M}_0})m_1 = 0$. Therefore, for $m \in \mathcal{M}_0^{\perp}$, $m \perp \mathcal{W}$ implies that $m = 0$, proving that $\mathcal{M}_0^{\perp}$ also satisfies the second requirement.

Note that to describe μ_y^{**}, we can use $\mathcal{M}_0^{\perp}$ in place of $\mathcal{M}$ in the theory developed in Section 2. Because Condition (I) is satisfied by $\mathcal{M}_0^{\perp}$, all of the previous results hold with respect to this choice and $\mu_y^{**} = Vy$, where V is now based on $\mathcal{M}_0^{\perp}$.

4. IDENTIFICATION OF BOUNDED LINEAR FUNCTIONALS

Economists are often interested in estimating real-valued functions of conditional expectations. For example, letting y denote the market demand for a certain good and x the price, Newey and McFadden (1994) consider estimating $\int_D \mathbb{E}(y|x)\, dx$, the approximate change in consumer surplus for a given price change on interval D. In this section we consider an endogenous version of their problem by characterizing the identification of bounded linear functionals of μ_y^* when the latter itself may not be identified (obviously, if μ_y^* is uniquely defined then so is $\rho(\mu_y^*)$). The results of Ai and Chen (2003) can be used to estimate linear functionals of μ_y^* when the latter is identified.

Let $\rho : \mathcal{M} \to \mathbb{R}$ denote a continuous linear functional on $\mathcal{M}$, where a possibly nonunique $\mu_y^* \in \mathcal{M}$ satisfies $y - \mu_y^* \perp \mathcal{W}$; i.e., we let (2.1) hold though we do not assume that Condition (I) necessarily holds. We now introduce the condition under which $\rho(\mu_y^*)$ is uniquely defined.

Condition (I-F). If $m \in \mathcal{M}$ satisfies $m \perp \mathcal{W}$, then $\rho(m) = 0$.

As shown subsequently, Condition (I-F) is necessary and sufficient for $\rho(\mu_y^*)$ to be identified.

THEOREM 4.1. $\rho(\mu_y^*)$ *is identified if and only if Condition (I-F) holds.*

The next example illustrates the usefulness of this result.

Example 4.1 (Identification of expectation functionals)

Let $y = \mu_y^*(x) + \varepsilon$, where $\mu_y^* \in L_2(x)$ is unknown. The regressors are endogenous, but we have instruments satisfying $\mathbb{E}(\varepsilon|w) = 0$ w.p.1. Assume that the conditional distribution of x given w is not complete. Hence, μ_y^* is not identified. Now consider the expectation functional $\rho(\mu_y^*) = \mathbb{E}\{\mu_y^*(x)\psi(x)\}$, where ψ is a known weight function satisfying $\mathbb{E}\psi^2(x) < \infty$. Theorem 4.1 reveals that

$$\rho(\mu_y^*) = \mathbb{E}\{\mu_y^*(x)\psi(x)\} \quad \text{is identified if and only if } \psi \in \mathcal{M}_0^\perp. \tag{4.1}$$

The case $\psi(x) = 1$ is a special case of (4.1) because $\mathcal{M}_0^\perp$ contains all constant functions (in fact, because $\mathbb{E}\mu_y^*(x) = \mathbb{E}y$, it is obvious that $\mu_y^* \mapsto \mathbb{E}\mu_y^*(x)$ is identified irrespective of whether μ_y^* is identified or not). From (4.1) we can immediately see that in applications where μ_y^* is not identified certain expectation functionals of μ_y^* may still be identified. Of course, if μ_y^* is identified to begin with, then $\mathcal{M}_0 = \{0\}$ and $\mathcal{M}_0^\perp = L_2(x)$; hence, $\rho(\mu_y^*)$ is identified for all square integrable weight functions. We can also use (4.1) to characterize the identification of bounded linear functionals of the form $\mu_y^* \mapsto \int_{\mathbb{R}^s} \mu_y^*(x)\psi(x)\,dx$. In particular, it is easily seen that $\mu_y^* \mapsto \int_{\mathbb{R}^s} \mu_y^*(x)\psi(x)\,dx$ is identified if and only if ψ/h lies in $\mathcal{M}_0^\perp$, where h denotes the unknown Lebesgue density of x. Note that for $\mu_y^* \mapsto \int_{\mathbb{R}^s} \mu_y^*(x)\psi(x)\,dx$ to be a bounded linear functional on $L_2(x)$ it is implicitly understood that the random vector x is continuously distributed and $\int_{\mathbb{R}^s} \psi^2(x)/h(x)\,dx < \infty$. Of course, $\mu_y^* \mapsto \mathbb{E}\{\mu_y^*(x)\psi(x)\}$ is bounded on $L_2(x)$ even when some components of x are discrete.

Finally, we show that Condition (I) holds if and only if Condition (I-F) holds for all bounded linear functionals of μ_y^*. Hence, identification of μ_y^* can also be characterized as follows.

THEOREM 4.2. *μ_y^* is identified if and only if all bounded linear functionals of μ_y^* are identified.*

For endogenous nonparametric regression, this result provides a direct link between identification of μ_y^* and its expectation functionals by revealing that μ_y^* is identified if and only if all its expectation functionals are identified; i.e., μ_y^* is identified if and only if $\mathbb{E}\{\mu_y^*(x)\psi(x)\}$ is identified for all $\psi \in L_2(x)$.

5. LINEAR MOMENT CONDITIONS AND INSTRUMENTAL VARIABLES

We now formulate (2.1) in terms of moment conditions generated by linear operators and also provide an example to illustrate the usefulness of this characterization. Although this formulation may seem different from the manner in which (2.1) is stated, we show that the two representations are in fact logically equivalent. Let y denote an element of $\mathcal{U}$ and let $\mathcal{M}$ be a known linear subspace of $\mathcal{U}$. Suppose that corresponding to y is an element of $\mathcal{M}$, denoted by μ_y^*, defined as follows: "There exists a linear subspace of $\mathcal{U}$, denoted by $\mathcal{V}$, and a continuous linear operator $T:\mathcal{U} \to \mathcal{V}$ such that $T(y - \mu_y^*) = 0$." Let $\mathcal{Y}$ denote the set of $y \in \mathcal{U}$ for which this model holds; i.e., for each $y \in \mathcal{Y}$ there exists $\mu_y^* \in \mathcal{M}$ such that $T(y - \mu_y^*) = 0$. Note that because $\mathcal{M} \subseteq \mathcal{Y}$, the domain of T may be taken to be $\mathcal{Y}$.

Condition (I-M). If $m \in \mathcal{M}$ satisfies $Tm = 0$, then $m = 0$.

Condition (I-M) is necessary and sufficient for μ_y^* to be uniquely defined (the proof is straightforward and hence is omitted). We say that $(\mathcal{Y}, \mathcal{M})$ is a "moment-condition" model if there is a linear subspace $\mathcal{V}$ of $\mathcal{U}$ and a continuous linear function $T: \mathcal{Y} \to \mathcal{V}$ such that for each $y \in \mathcal{Y}$ there exists $m \in \mathcal{M}$ satisfying $T(y - m) = 0$ and Condition (I-M) holds. We call T the "identification function." Similarly, we say that $(\mathcal{Y}, \mathcal{M})$ is an "instrumental variables" model if there is a closed linear subspace $\mathcal{W}$ of $\mathcal{U}$ such that for each $y \in \mathcal{Y}$ there exists $m \in \mathcal{M}$ satisfying $y - m \perp \mathcal{W}$ and Condition (I) holds. In fact, it can be easily shown that $(\mathcal{Y}, \mathcal{M})$ is a moment-condition model if and only if it is an instrumental variables model.

The following example illustrates a situation where the nature of the available information makes it easier to write an endogenous nonparametric regression model as a moment-condition model.

Example 5.1

Let $y = \mu_y^*(x) + \varepsilon$, where $x \in \mathbb{R}^s$ for $s > 1$ and $\mu_y^* \in L_2(x)$ is unknown. The regressors are correlated with the error term such that the conditional distribution of ε given x satisfies the index restriction $\mathbb{E}\{\varepsilon|x\} = \mathbb{E}\{\varepsilon|h(x)\}$ w.p.1 for some known function h with $dim\, h(x) < s$. This, e.g., is related to the exclusion restriction assumption maintained in Florens, Heckman, Meghir, and Vytlacil (2002). Let $T\varepsilon = \mathbb{E}\{\varepsilon|x\} - \mathbb{E}\{\varepsilon|h(x)\}$. Our model has content if the linear moment condition $T(y - \mu_y^*) = 0$ holds for some $\mu_y^* \in L_2(x)$. For μ_y^* to be uniquely defined, by Condition (I-M) we need that $\mu_y^*(x) = \mathbb{E}\{\mu_y^*(x)|h(x)\}$ w.p.1 only for $\mu_y^*(x) = 0$ w.p.1. This reveals that μ_y^*'s of the form $\mu_y^*(x) = f(h(x))$ are not identifiable. Therefore, letting $\mathcal{M}$ denote the set of all functions in $L_2(x)$ that are not functions of $h(x)$, it follows that $(y, \mathcal{M})$ is a moment-condition model with identification function T. Next, we show how to write $(y, \mathcal{M})$ as an instrumental variables model. Let $\mathcal{T}_0$ be the null space of T; i.e., $\mathcal{T}_0$ is the set of all random variables ε such that $\mathbb{E}\{\varepsilon|x\} = \mathbb{E}\{\varepsilon|h(x)\}$ w.p.1. Because $(y, \mathcal{M})$ is a moment-condition model with identification function T, by definition there exists a unique $\mu_y^* \in \mathcal{M}$ such that $y - \mu_y^* \in \mathcal{T}_0$. It follows that $(y, \mathcal{M})$ is also an instrumental variables model with instrument space $\mathcal{T}_0^\perp$, where $\mathcal{T}_0^\perp = \{v \in L_2(x): v \perp L_2(h(x))\}$.[13]

6. CONCLUSION

In this paper we investigate some identification issues in nonparametric linear models with endogenous regressors. Our results suggest that identification in such models can fail to hold for even relatively simple designs. Therefore, if researchers are not careful, simply assuming identification and then proceeding to estimation can lead to statistical inference that may be seriously misleading. Because lack of identification here is not easily detected, we show how to deter-

mine the identifiable part of the structural function when it is underidentified by orthogonally projecting onto an appropriately defined subspace of the model space. We also examine the connection between identification of the unknown structural function and identification of its linear functionals and show that the two are closely related.

NOTES

1. A good discussion of these models, though without any endogeneity, can be found in Hastie and Tibshirani (1990).

2. Because $L_2(x)$ and $L_2(w)$ are equivalence classes of functions, equality statements in $L_2(x)$ and $L_2(w)$ hold w.p.1.

3. Because $P_{\mathcal{W}}$ is bounded, its restriction to $\mathcal{M}$ is also bounded and, hence, continuous. Now let m_1 and m_2 denote elements of $\mathcal{M}$ and $w_j = \bar{P}_{\mathcal{W}} m_j$. Suppose $w_1 = w_2$. Then $\bar{P}_{\mathcal{W}}(m_1 - m_2) = 0$. Hence, $m_1 - m_2 \perp \mathcal{W}$. It follows from Condition (I) that $m_1 = m_2$ so that $\bar{P}_{\mathcal{W}}$ is one-to-one and, by definition, the range of $\bar{P}_{\mathcal{W}}$ is $\mathcal{W}_0$. Therefore, because $\bar{P}_{\mathcal{W}}: \mathcal{M} \to \mathcal{W}_0$ is one-to-one and onto, it has an inverse $\bar{P}_{\mathcal{W}}^{-1}: \mathcal{W}_0 \to \mathcal{M}$.

4. Basis vectors that satisfy Lemma 2.2 for more general bivariate distributions can be constructed by using some of the results discussed in Buja (1990).

5. See Petryshyn (1963) for a detailed treatment of recursive methods of this type.

6. Discontinuity of V means that slight perturbations in the response variable can lead to unbounded changes in μ_y^*, the parameter of interest associated with it. This lack of stability makes precise the sense in which some endogenous nonparametric models can be called "ill-posed." Note that sometimes a statistical problem is said to be ill-posed because of data issues; e.g., classic nonparametric regression itself can be called ill-posed because we cannot estimate the graph of an unknown function using only a finite amount of data. However, the notion of ill-posedness described here has nothing to do with sample information but is inherent to the model.

7. See, e.g., Blundell and Powell (2003) and Florens and Malavolti (2003).

8. Because $\mathbb{E}\{Q_j(x,z)|w\} = 0$, $Q_j \in \mathcal{M}_0$ for each j; i.e., $\mathcal{Q} \subseteq \mathcal{M}_0$. Next, let $m_0 \in \mathcal{M}_0$. Then $m_0(x,z) = f(x) + g(z)$ for some $f \in L_2(x)$ and $g \in L_2(z)$ such that $\mathbb{E}\{f(x) + g(z)|w\} = 0$; i.e., $\mathbb{E}\{f(x)|w\} = -\mathbb{E}\{g(z)|w\}$. Hence, writing $f(x) = \sum_{j=0}^{\infty} \alpha_j H_j(x)$ and $g(z) = \sum_{j=0}^{\infty} \beta_j H_j(z)$, it follows that $\sum_{j=0}^{\infty} \alpha_j \rho_{xw}^j H_j(w) = -\sum_{j=0}^{\infty} \beta_j \rho_{zw}^j H_j(w)$ if and only if $\sum_{j=0}^{\infty} \{\alpha_j \rho_{xw}^j + \beta_j \rho_{zw}^j\} H_j(w) = 0$. By the completeness of Hermite polynomials in $L_2(w)$, this implies that $\alpha_j \rho_{xw}^j + \beta_j \rho_{zw}^j = 0$ for each j. Therefore, $f(x) + g(z) = -\sum_{j=0}^{\infty} \beta_j Q_j(x,z)/\rho_{xw}^j \in \mathcal{Q}$; $\mathcal{M}_0 \subseteq \mathcal{Q}$.

9. There is an analogy to $\mathcal{M}_0^{\perp}$ in the specification testing literature. Suppose we want to test the null hypothesis $\mathbb{E}(y|x) = x'\theta$ against the alternative that it is false. Consider the alternative $\mathbb{E}(y|x) = x'\theta + \delta(x)$, where δ denotes a deviation from the null. It is obvious that no test will be able to reject the null if δ is a linear function of x. The only detectable perturbations are those that are orthogonal to linear functions, i.e., those satisfying $\mathbb{E}\{x\delta(x)\} = 0$.

10. Let $m_y \in \mathcal{M}_y$ be arbitrary. Then $y - \mu_y^{**} = y - P_{\mathcal{M}_0^{\perp}} m_y = y - m_y + P_{\mathcal{M}_0} m_y$. But $y - m_y \perp \mathcal{W}$ by definition of $\mathcal{M}_y$, and $P_{\mathcal{M}_0} m_y \perp \mathcal{W}$ because $P_{\mathcal{M}_0} m_y \in \mathcal{M}_0$. Therefore, $y - \mu_y^{**} \perp \mathcal{W}$. Because $\mu_y^{**} \in \mathcal{M}$, it follows that $\mu_y^{**} \in \mathcal{M}_y$.

11. Showing $\mathcal{B} \subseteq \mathcal{M}_0^{\perp}$ is easy. Next, by the projection theorem,

$$\text{proj}(g|\mathcal{M}_0)(x) = \begin{cases} g(x)h(x) + g(x+1)h(x+1) - \mathbb{E}g(x) & \text{if } -1 \leq x \leq 0, \\ g(x)h(x) + g(x-1)h(x-1) - \mathbb{E}g(x) & \text{if } 0 < x \leq 1. \end{cases}$$

Now let $g \in \mathcal{M}_0^{\perp}$. Because g is an element of $\mathcal{M}_0^{\perp}$, its projection onto $\mathcal{M}_0$ is the zero function. Hence, using the expression for $\text{proj}(g|\mathcal{M}_0)$, it follows that $g \in \mathcal{B}$; i.e., $\mathcal{M}_0^{\perp} \subseteq \mathcal{B}$. Therefore, $\mathcal{M}_0^{\perp} = \mathcal{B}$.

12. Recall that $\mathcal{M}_0 = span\{Q_j(x,z)\}$. Now let $A_k \in \mathcal{A}$. Because $\mathbb{E}\{A_k(x,z)Q_j(x,z)\} = 0$ for all j, $A_k \in \mathcal{M}_0^\perp$. Therefore, $\mathcal{A} \subseteq \mathcal{M}_0^\perp$. Next, let $m_1 \in \mathcal{M}_0^\perp$. Then $m_1(x,z) = f(x) + g(z)$ for some $f \in L_2(x)$ and $g \in L_2(z)$. But because m_1 is orthogonal to $\mathcal{M}_0$ by definition, $\mathbb{E}\{f(x) + g(z)\}Q_k(x,z) = 0$ for all k. Hence, writing $f(x) = \sum_{j=0}^{\infty} \alpha_j H_j(x)$ and $g(z) = \sum_{j=0}^{\infty} \beta_j H_j(z)$, we have $\sum_{j=0}^{\infty} \alpha_j \mathbb{E}\{H_j(x)Q_k(x,z)\} + \sum_{j=0}^{\infty} \beta_j \mathbb{E}\{H_j(z)Q_k(x,z)\} = 0$ for $k = 0,1,\ldots$. Thus $(\rho_{xz}^j \rho_{xw}^j - \rho_{zw}^j)\alpha_j + \beta_j(\rho_{xw}^j - \rho_{xz}^j \rho_{zw}^j) = 0$ for $j \geq 1$, and $f(x) + g(z) = \alpha_0 + \beta_0 + \sum_{j=1}^{\infty} \beta_j A_j(x,z)/(\rho_{zw}^j - \rho_{xz}^j \rho_{xw}^j)$. It follows that $m_1 \in \mathcal{A}$; i.e., $\mathcal{M}_0^\perp \subseteq \mathcal{A}$.

13. Define $\mathcal{W} = \{v \in L_2(x) : v \perp L_2(h(x))\}$ and let v and ε denote arbitrary elements of $\mathcal{W}$ and $\mathcal{T}_0$, respectively. Then by the properties of ε and v, $\mathbb{E}\{v(x)\varepsilon\} = \mathbb{E}\{v(x)\mathbb{E}[\varepsilon|x]\} = \mathbb{E}\{v(x)\mathbb{E}[\varepsilon|h(x)]\} = 0$; i.e., $\mathcal{W} \perp \mathcal{T}_0$, implying that $\mathcal{W} \subseteq \mathcal{T}_0^\perp$. Next, let $u \in \mathcal{T}_0^\perp$. Because the random variable $a(x) = \mathbb{E}\{u|h(x)\}$ satisfies $\mathbb{E}\{a(x)|x\} = \mathbb{E}\{a(x)|h(x)\}$, we obtain that $\mathbb{E}\{u|h(x)\}$ is an element of $\mathcal{T}_0$. Hence, $u \perp \mathbb{E}\{u|h(x)\}$, which implies that $\mathbb{E}\{u|h(x)\} = 0$ w.p.1. Therefore, $u \perp L_2(h(x))$. Now write $u = u_1 + u_2$, where $u_1 \in L_2(x)$ and $u_2 \in L_2^\perp(x)$. Because $\mathbb{E}\{u_2|x\} = 0$ w.p.1 and $\mathbb{E}\{u_2|h(x)\} = \mathbb{E}\{\mathbb{E}[u_2|x]|h(x)\} = 0$ w.p.1, it follows that $\mathbb{E}\{u_2|x\} = \mathbb{E}\{u_2|h(x)\}$ w.p.1; i.e., $u_2 \in \mathcal{T}_0$. But because $u \in \mathcal{T}_0^\perp$, $0 = \mathbb{E}\{uu_2\} = \mathbb{E}\{u_2^2\}$ implies that $u_2 = 0$ w.p.1. Hence, $u = u_1 \in L_2(x)$. Thus $u \in L_2(x)$ and $u \perp L_2(h(x))$; i.e., $u \in \mathcal{W}$. Because u was chosen arbitrarily in $\mathcal{T}_0^\perp$, we conclude that $\mathcal{T}_0^\perp \subseteq \mathcal{W}$.

REFERENCES

Ai, C. & X. Chen (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71, 1795–1843.

Blundell, R. & J.L. Powell (2003) Endogeneity in nonparametric and semiparametric regression models. In M. Dewatripont, L. Hansen, & S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications*, vol. 2, pp. 312–357. Cambridge University Press.

Buja, A. (1990) Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics* 18, 1032–1069.

Carrasco, M., J.-P. Florens, & E. Renault (2002) Linear Inverse Problems in Structural Econometrics. Manuscript, University of Rochester.

Darolles, S., J.-P. Florens, & E. Renault (2002) Nonparametric Instrumental Regression. Manuscript, University of Toulouse.

Das, M. (2001) Instrumental Variables Estimation of Nonparametric Models with Discrete Endogenous Regressors. Manuscript, Columbia University.

Engle, R., C. Granger, J. Rice, & A. Weiss (1986) Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* 81, 310–320.

Florens, J.-P. (2003) Inverse problems and structural econometrics: The example of instrumental variables. In M. Dewatripont, L. Hansen, & S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications*, vol. 2, pp. 284–311. Cambridge University Press.

Florens, J.-P., J. Heckman, C. Meghir, & E. Vytlacil (2002) Instrumental Variables, Local Instrumental Variables, and Control Functions. CEMMAP Working paper CWP19/02.

Florens, J.-P. & L. Malavolti (2003) Instrumental Regression with Discrete Variables. Manuscript, University of Toulouse.

Florens, J.-P., M. Mouchart, & J. Rolin (1990) *Elements of Bayesian Statistics*. Marcel Dekker.

Granger, C. & P. Newbold (1976) Forecasting transformed series. *Journal of the Royal Statistical Society, Series B* 38, 189–203.

Hall, P. & J.L. Horowitz (2003) Nonparametric Methods for Inference in the Presence of Instrumental Variables. CEMMAP Working paper CWP02/03.

Hastie, T. & R. Tibshirani (1990) *Generalized Additive Models*. Chapman and Hall.

Kress, R. (1999) *Linear Integral Equations*, 2nd ed. Springer-Verlag.

Kreyszig, E. (1978) *Introductory Functional Analysis with Applications*. Wiley.

Li, K.-C. (1984) Regression models with infinitely many parameters: Consistency of bounded linear functionals. *Annals of Statistics* 12, 601–611.

Linton, O., E. Mammen, J. Nielsen, & C. Tanggaard (2001) Yield curve estimation by kernel smoothing methods. *Journal of Econometrics* 105, 185–223.

Loubes, J. & A. Vanhems (2003) Saturation Spaces for Regularization Methods in Inverse Problems. Manuscript, Université Paris-Sud.

Milne, W. (1929) On the degree of convergence of the Gram-Charlier series. *Transactions of the American Mathematical Society* 31, 422–443.

Newey, W.K. & D. McFadden (1994) Large sample estimation and hypothesis testing. In R. Engle & D. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp. 2111–2245. Elsevier Science B.V.

Newey, W.K. & J.L. Powell (2003) Instrumental variables estimation of nonparametric models. *Econometrica* 71, 1557–1569.

Newey, W.K., J.L. Powell, & F. Vella (2003) Nonparametric estimation of triangular simultaneous equations models. *Econometrica* 67, 565–603.

Petryshyn, W. (1963) On a general iterative method for the approximate solution of linear operator equations. *Mathematics of Computation* 17, 1–10.

Pinkse, J. (2000) Nonparametric two-step regression estimation when regressors and error are independent. *Canadian Journal of Statistics* 28, 289–300.

Robinson, P.M. (1988) Root-*N*-consistent semiparametric regression. *Econometrica* 56, 931–954.

Roehrig, C.S. (1988) Conditions for identification in nonparametric and parametric models. *Econometrica* 56, 433–447.

Wahba, G. (1990) *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, no. 59.

Wooldridge, J.M. (2002) *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

APPENDIX: Proofs

Proof of (2.5). By the orthogonality of Hermite polynomials,

$$\int_{\mathbb{R}} \{\hat{\mu}_n(x) - \mu_y^*(x)\}^2 \phi(x)\, dx = \sum_{j=0}^{k_n} \frac{(\hat{\gamma}_j - \gamma_j)^2}{j!\,\rho^{2j}} + \sum_{j=k_n+1}^{\infty} \frac{\gamma_j^2}{j!\,\rho^{2j}}.$$

Hence,

$$\text{MISE} = \mathbb{E} \int_{\mathbb{R}} \{\hat{\mu}_n(x) - \mu_y^*(x)\}^2 \phi(x)\, dx \le \rho^{-2k_n} \sum_{j=0}^{k_n} \frac{\mathbb{E}(\hat{\gamma}_j - \gamma_j)^2}{j!} + \sum_{j=k_n+1}^{\infty} \frac{\gamma_j^2}{j!\,\rho^{2j}}.$$

But because $\text{var}(yH_j(w)) = \mathbb{E}\{\text{var}(y|w)H_j^2(w)\} + \text{var}\{\mathbb{E}(y|w)H_j(w)\}$ and the maps $w \mapsto \mathbb{E}(y|w)$ and $w \mapsto \text{var}(y|w)$ are bounded by assumption,

$$\text{var}(yH_j(w)) \le \mathbb{E}\{\text{var}(y|w)H_j^2(w)\} + \mathbb{E}\{\mathbb{E}^2(y|w)H_j^2(w)\} \le c\mathbb{E}H_j^2(w) = cj!,$$

where c is a generic constant. Thus, $\mathbb{E}(\hat{\gamma}_j - \gamma_j)^2 = \text{var}(yH_j(w))/n \le cj!/n$. It follows that the MISE is majorized by $c\rho^{-2k_n}k_n/n + \sum_{j=k_n+1}^{\infty} \gamma_j^2/j!\,\rho^{2j}$. But $\sum_{j=k_n+1}^{\infty} \gamma_j^2/j!\,\rho^{2j} \le ck_n^{-\alpha}$ for some $\alpha > 0$ under some smoothness conditions on μ_y^*; see, e.g., Milne (1929, Cor. I). The desired result follows. ∎

Proof of Lemma 2.1. Let S denote the statement "For each function $f(x)$ such that $\mathbb{E}f(x) = 0$ and $\text{var}\,f(x) > 0$, there exists a function $g(w)$ such that $f(x)$ and $g(w)$ are correlated." First, suppose that S is not true. Hence, there exists a nonzero function $f(x)$

satisfying $\mathbb{E}f(x) = 0$ such that $\mathbb{E}\{f(x)g(w)\} = 0$ for all $g(w)$. But this implies that $\mathbb{E}\{f(x)|w\} = 0$ w.p.1; i.e., $pdf(x|w)$ is not complete. Next, suppose that the conditional distribution of $x|w$ is not complete. Hence, there exists a function $f(x)$ such that $\mathbb{E}\{f(x)|w\} = 0$ w.p.1 but $f(x) \neq 0$; i.e., $\operatorname{var}f(x) > 0$. Clearly, this implies that $f(x)$ is uncorrelated with all functions of w. The desired result follows. ∎

Proof of Lemma 2.2. We first consider the series expansion for Vy. Let $y \in \mathcal{Y}$. Hence, there exists a unique (because Condition (I) is assumed to hold) $\mu_y^* \in \mathcal{M}$ such that $y - \mu_y^* \perp \mathcal{W}$. This implies that $P_{\mathcal{W}}y - P_{\mathcal{W}}\mu_y^* = 0$. Because $\mu_y^* \in \mathcal{M}$, we can write $\mu_y^* = \sum_{j=0}^{\infty} \alpha_j m_j$ for some constants $\alpha_0, \alpha_1, \dots$. Hence, $P_{\mathcal{W}}\mu_y^* = \sum_{j=0}^{\infty} \alpha_j P_{\mathcal{W}}m_j$, and using the fact that $\langle m_i, P_{\mathcal{W}}m_j \rangle = 0$ for $i \neq j$, it follows that $\langle m_i, P_{\mathcal{W}}\mu_y^* \rangle = \alpha_i \langle m_i, P_{\mathcal{W}}m_i \rangle$. Hence, $\alpha_i = \langle m_i, P_{\mathcal{W}}y \rangle / \langle m_i, P_{\mathcal{W}}m_i \rangle$ because $P_{\mathcal{W}}y = P_{\mathcal{W}}\mu_y^*$. Therefore, $Vy = \mu_y^* = \sum_{i=0}^{\infty} \{\langle m_i, P_{\mathcal{W}}y \rangle / \langle m_i, P_{\mathcal{W}}m_i \rangle\}m_i$. Next, we consider the series for $\bar{P}_{\mathcal{W}}^{-1}$. Let $\tilde{w}$ be any element of $\mathcal{W}_0$ and let $\tilde{m} = \bar{P}_{\mathcal{W}}^{-1}\tilde{w}$. Note that $P_{\mathcal{W}}\bar{P}_{\mathcal{W}}^{-1}\tilde{w} = \tilde{w}$ by definition of $\bar{P}_{\mathcal{W}}^{-1}$. Because $\tilde{m} \in \mathcal{M}$ and $\mathcal{M}$ is contained in $\mathcal{Y}$, by the previous result

$$V\tilde{m} = \sum_{j=0}^{\infty} \frac{\langle m_j, P_{\mathcal{W}}\tilde{m} \rangle}{\langle m_j, P_{\mathcal{W}}m_j \rangle} m_j = \sum_{j=0}^{\infty} \frac{\langle m_j, P_{\mathcal{W}}\bar{P}_{\mathcal{W}}^{-1}\tilde{w} \rangle}{\langle m_j, P_{\mathcal{W}}m_j \rangle} m_j = \sum_{j=0}^{\infty} \frac{\langle m_j, \tilde{w} \rangle}{\langle m_j, P_{\mathcal{W}}m_j \rangle} m_j.$$

But by (2.2), $V\tilde{m} = \bar{P}_{\mathcal{W}}^{-1}P_{\mathcal{W}}\tilde{m} = \bar{P}_{\mathcal{W}}^{-1}P_{\mathcal{W}}\bar{P}_{\mathcal{W}}^{-1}\tilde{w} = \bar{P}_{\mathcal{W}}^{-1}\tilde{w}$. The desired result follows. ∎

Proof of Lemma 2.3. Note that $m_{n+1} - m_n = aP_{\mathcal{M}}\bar{P}_{\mathcal{W}}(m_0 - m_n)$. If $\lim_{n\uparrow\infty} m_n$ exists, then $m_{n+1} - m_n \to 0$ as $n \uparrow \infty$ so that $P_{\mathcal{M}}\bar{P}_{\mathcal{W}}(m_n - m_0) \to 0$ as $n \uparrow \infty$. Because $P_{\mathcal{M}}\bar{P}_{\mathcal{W}}(m_n - m^*) \to 0$ as $n \uparrow \infty$, it follows that $P_{\mathcal{M}}\bar{P}_{\mathcal{W}}(m_0 - m^*) = 0$. Let $m = m_0 - m^*$; then $P_{\mathcal{W}}m \perp \mathcal{M}$. Hence, $\langle P_{\mathcal{W}}m, m \rangle = \langle P_{\mathcal{W}}m, P_{\mathcal{W}}m \rangle = 0$ so that $P_{\mathcal{W}}m = 0$. It follows from Condition (I) that $m = 0$, i.e., that $m^* = m_0$, proving the result. ∎

Proof of Lemma 2.4. We show that (i) $\Rightarrow$ (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i) and (i) $\Rightarrow$ (iv) $\Rightarrow$ (v) $\Rightarrow$ (i).

First, suppose V is continuous. Let $m_1, m_2, \dots$ be a sequence in $\mathcal{M}$ such that $\lim_{n\uparrow\infty} P_{\mathcal{W}}m_n = 0$ and let $y_n = P_{\mathcal{W}}m_n$. Note that for each n, $y_n - m_n = P_{\mathcal{W}}m_n - m_n \perp \mathcal{W}$ so that $y_n \in \mathcal{Y}$. Hence, $m_n = Vy_n$. Because $y_n \to 0$ as $n \uparrow \infty$, it follows that $m_n \to 0$. Therefore, (iii) holds.

Next, assume that (iii) holds. Let $y_1, y_2, \dots$ denote a sequence in $\mathcal{Y}$ such that $\lim_{n\uparrow\infty} y_n = y$ for some $y \in \mathcal{U}$. We need to show that $y \in \mathcal{Y}$. Observe that because $P_{\mathcal{W}}V(y_n - y) = P_{\mathcal{W}}(y_n - y)$ by (2.2) and $P_{\mathcal{W}}$ is continuous, it follows that $P_{\mathcal{W}}(y_n - y) \to 0$ as $n \uparrow \infty$. But as $Vy_n - Vy$ is in $\mathcal{M}$, by (iii) we have $Vy_n - Vy \to 0$ as $n \uparrow \infty$. This shows that $y_n - Vy_n - (y - Vy) \to 0$ as $n \uparrow \infty$. Because $y_n \in \mathcal{Y}$, we know that $y_n - Vy_n \perp \mathcal{W}$ for each n. Thus, by the continuity of the inner product, $\langle y - Vy, w \rangle = 0$ for all $w \in \mathcal{W}$. Hence, $y \in \mathcal{Y}$ and $\mathcal{Y}$ is closed.

Next, suppose that $\mathcal{Y}$ is closed. Then because V is a closed linear operator with domain $\mathcal{Y}$, continuity of V follows by the closed graph theorem; see, e.g., Kreyszig (1978).

Now, let (i) hold. This implies that $\mathcal{Y}$ is closed. Let $w_1, w_2, \dots$ denote a sequence in $\mathcal{W}_0$ with limit $w \in \mathcal{W}$ and, for each $n = 1, 2, \dots$, let $m_n \in \mathcal{M}$ satisfy $w_n = P_{\mathcal{W}}m_n$. Because $w_n - m_n = P_{\mathcal{W}}m_n - m_n \perp \mathcal{W}$, it follows that $w_n \in \mathcal{Y}$ and $m_n = Vw_n$. Thus $w \in \mathcal{Y}$ because $\mathcal{Y}$ is closed. Next, $\lim_{n\uparrow\infty} Vw_n = Vw$ by continuity of V. Hence, $\lim_{n\uparrow\infty} P_{\mathcal{W}}(m_n - Vw) = \lim_{n\uparrow\infty} w_n - P_{\mathcal{W}}Vw = 0$, so that $w_n \to P_{\mathcal{W}}Vw$ as $n \uparrow \infty$. Because $Vw \in \mathcal{M}$, we know that $P_{\mathcal{W}}Vw \in \mathcal{W}_0$. Hence, $\mathcal{W}_0$ is closed.

Next, assume that $\mathcal{W}_0$ is closed. Because $\mathcal{W}_0 \subseteq \mathcal{Y}$, the result follows by taking $\mathcal{Y}_0 = \mathcal{W}_0$.

Finally, suppose that (v) holds. Then the restriction of V to $\mathcal{Y}_0$ is closed, and, because $\mathcal{Y}_0$ is closed, that restriction is continuous by the closed graph theorem. Let $m_1, m_2, \ldots$ denote a sequence in $\mathcal{M}$ such that $P_{\mathcal{W}} m_n \to 0$ as $n \uparrow \infty$ and let $w_n = P_{\mathcal{W}} m_n$ so that $w_1, w_2, \ldots$ is a sequence in $\mathcal{W}_0$ such that $w_n \to 0$ as $n \uparrow \infty$. Note that $w_n - m_n = P_{\mathcal{W}} m_n - m_n \perp \mathcal{W}$ for each n, implying that $P_{\mathcal{W}}(w_n - m_n) = 0$; i.e., $P_{\mathcal{W}} w_n = \bar{P}_{\mathcal{W}} m_n$ because $\bar{P}_{\mathcal{W}}$ is the restriction of $P_{\mathcal{W}}$ to $\mathcal{M}$. Hence, $m_n = V w_n$ by (2.2). Because V is continuous on $\mathcal{Y}_0$ and $\mathcal{W}_0 \subseteq \mathcal{Y}_0$, it follows that $\lim_{n \uparrow \infty} m_n = V \lim_{n \uparrow \infty} w_n = 0$. Thus (iii) holds. But we have already shown that (iii) implies (i). The desired result follows. ∎

Proof of Theorem 4.1. $\rho(\mu_y^*)$ is identified if and only if all $\mu_y^* \in \mathcal{M}$ for which $y - \mu_y^* \perp \mathcal{W}$ yield the same value of $\rho(\mu_y^*)$. Suppose that Condition (I-F) holds and that m_1, m_2 are elements of $\mathcal{M}$ satisfying $y - m_j \perp \mathcal{W}$ for $j = 1, 2$. Then $m_1 - m_2 \perp \mathcal{W}$, so, by Condition (I-F), $\rho(m_1 - m_2) = \rho(m_1) - \rho(m_2) = 0$. Hence, $\rho(\mu_y^*)$ is identified. Next, suppose that all m for which $y - m \perp \mathcal{W}$ yield the same value of $\rho(m)$. Suppose $y - m_1 \perp \mathcal{W}$ and $m \in \mathcal{M}$ satisfies $m \perp \mathcal{W}$. Then $y - (m_1 + m) \perp \mathcal{W}$. Because $\rho(\mu_y^*)$ is identified, $\rho(m_1 + m) = \rho(m_1) + \rho(m) = \rho(m_1)$ so that $\rho(m) = 0$. Hence, Condition (I-F) holds. ∎

Proof of Theorem 4.2. Suppose Condition (I) holds and $m \in \mathcal{M}$ satisfies $m \perp \mathcal{W}$. Then, by Condition (I), $m = 0$ so that $\rho(m) = 0$ for any linear functional ρ. Next, suppose that Condition (I-F) holds for any continuous linear functional ρ and let $m \in \mathcal{M}$ satisfy $m \perp \mathcal{W}$. Consider the bounded linear functional $\rho(m_1) = \langle m, m_1 \rangle$, where $m_1 \in \mathcal{M}$. Because Condition (I-F) holds, it follows that $\langle m, m \rangle = 0$. But this implies that $m = 0$. Hence, Condition (I) holds. ∎