Nonparametric Matching and Efficient Estimators of Homothetically Separable Functions*

Arthur Lewbel† Boston College

Oliver Linton‡ London School of Economics

April 3, 2006

Abstract

For vectors z and w and scalar v, let $r(v, z, w)$ be a function that can be nonparametrically estimated consistently and asymptotically normally, such as a distribution, density, or conditional mean regression function. We provide consistent, asymptotically normal nonparametric estimators for the functions G and H, where $r(v, z, w) = H[vG(z), w]$, and some related models. This framework encompasses homothetic and homothetically separable functions, and transformed partly additive models $r(v, z, w) = h[v + g(z), w]$ for unknown functions g and h. Such models reduce the curse of dimensionality, provide a natural generalization of linear index models, and are widely used in utility, production, and cost function applications. We also provide an estimator of G that is oracle efficient, achieving the same performance as an estimator based on local least squares knowing H.

JEL Codes: C14, C21, D24.

Keywords: Cost Function; Economies of Scale; Homogeneous Function; Homothetic Function; Index Models; Nonparametric; Oracle Efficiency; Production Function; Separability.

*This research was supported in part by the National Science Foundation through grant SES-9905010, and through a grant from the Economic and Social Science Research Council. We would like to thank David Jacho-Chavez for research assistance, and Donald Andrews, Martin Browning, Shakeeb Khan, Yuichi Kitamura, Rosa Matzkin, Whitney Newey, Peter Phillips, Joel Horowitz, Gautham Tripathi, and anonymous referees for helpful comments. All errors are our own. GAUSS and R code is available from the authors upon request.

†Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA. Phone: (617) 552-3678. http://www2.bc.edu/~lewbel/ E-mail: lewbel@bc.edu.

‡Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. http://econ.lse.ac.uk/staff/olinton/~index_own.html. E-mail address: o.linton@lse.ac.uk.

1 Introduction

Let V_i be an observed scalar and Z_i and W_i be observed vectors for $i = 1, \ldots, n$. Let $R(v, z, w)$ be some function that can be nonparametrically estimated, for example, $R(v, z, w)$ could equal $E(Y_i \mid V_i = v, Z_i = z, W_i = w)$, which is estimated with observations $\{Y_i, V_i, Z_i, W_i\}$. More generally, $R(v, z, w)$ could be a density, distribution, quantile, or hazard function, or $R(v, z, w)$ could be a utility or cost function derived from a set of estimated product or factor demands. Assume there exist unknown functions h and g and known strictly monotonic functions B_1, B_2, and B_3 such that

$$R(v, z, w) = h[B_1(B_2(v)B_3(g(z)), w] \tag{1}$$

where h is strictly monotonic on its first element. We provide consistent, asymptotically normal estimators of the functions h and g. The estimator for g has an "oracle efficiency" property as in Linton (1996), i.e., it has the same asymptotic distribution as the corresponding estimator defined when h is known. Also, replacing g and h with their estimates in equation (1) will result in an estimate of R that has a faster rate of convergence than the original nonparametric R estimate.

One leading example of equation (1) is when B_1 is the natural logarithm and B_2 and B_3 are exponentiation, which gives

$$R(v, z, w) = h[v + g(z), w]. \tag{2}$$

When R is a conditional expectation, this an example of a generalized partly linear model with unknown link function similar to Horowitz (2001) and Horowitz and Mammen (2005). Equation (2) also arises in the nonparametric regression context given the model $V = -g(Z) + e$ for some error term e. If we strengthen the usual nonparametric regression assumption $E(e \mid Z) = 0$ to an independence assumption $e \perp Z, W$, then we obtain equation (2) where $R(v, z, w)$ is the unknown conditional distribution function of Z evaluated at $Z = z$, conditional on $V = v$ and $W = w$.

Another important example of equation (1) is when B_1, B_2, and B_3 are the identity functions, which gives

$$R(v, z, w) = h[vg(z), w]. \tag{3}$$

A function $r(x, w)$ is defined to be homothetically separable in x if and only if

$$r(x, w) = h[s(x), w] \tag{4}$$

where h is strictly monotonic in s and s is linearly homogeneous. Let v be one element of x that never equals zero, and let z be the vector of all the other elements of x divided by v. Alternatively, rewrite x in polar coordinates as v, z, where v is length and z is direction. Either way, $s(x)$ is linearly homogeneous if and only if $s(x) = vg(z)$ for some unrestricted function g, so a function r

is homothetically separable in x if and only if it has the form of R in equation (3). Similarly, when w is empty, equation (3) is equivalent to the definition of a function that is homothetic in x. For example, if v is labor and z is the capital labor ratio, then equation (3) equals the definition of a homothetic production function.

In applications of homothetic separability, r may have multiple homogeneous components, that is, $r(x_0, x_1, \ldots, x_k) = h[s_1(x_1), \ldots, s_K(x_K), x_0]$ for vectors $x_0, x_1, \ldots, x_K$. In this model, each homogeneous s_k function can be estimated separately by applying the method we propose to estimate g in equation (3), taking $x = x_k$ and w equal to the union of all the elements in $x_0, x_1, \ldots, x_K$ except x_k. Then, given estimates of each g_k (and hence each s_k) function, the function h may be estimated by nonparametrically regressing r on $s_1, \ldots, s_K, x_0$. In the same way our estimator immediately extends to models like $R(v, z_0, z_1, \ldots, z_k) = h[v + \sum_k g_k(z_k), z_0]$, where each g_k is estimated by taking $z = z_k$ and w equal to the union of all the elements in $z_0, z_1, \ldots, z_K$ except z_k.

In many applications the functions h and g are of direct interest, e.g., in equation (3) the returns to scale of a homothetic production function is defined as the log derivative of h with respect to $s = vg$, and the technical rate of substitution is a function of g. Even when h and g are not of direct interest, our estimator will still be useful for speeding the rate of convergence and for testing whether functions are homothetic, homothetically separable, or more generally if they satisfy equation (1).

Homothetic and homothetically separable functions are commonly used in models of consumer preferences and firm production, e.g., $r(x, w)$ could be a utility or consumer cost function recovered from estimated consumer demand functions via revealed preference theory, or it could be a directly estimated production or producer cost function. See, e.g., Blackorby, Primont, and Russell (1978), Lewbel (1991), (1997), Matzkin (1994), Primont and Primont (1994), and Zellner and Ryu (1998).

Linear index models with $s(x) = x^\top \beta$, are a very common semiparametric specification that arises in a variety of contexts, particularly limited dependent variable models. See Powell (1994) for a survey. Replacing a linear index $x^\top \beta$ with an arbitrary linearly homogeneous function $s(x)$ is a natural generalization, particularly in contexts where economic theory gives rise to homogeneity but not necessarily linearity, such as price indices or constant returns to scale technologies.

Matzkin (1992) provides a consistent estimator for the binary threshold crossing model $y = I[s(x) + \varepsilon \geq 0]$ where $s(x)$ is linearly homogeneous and ε is independent of x. This threshold crossing model has $E(Y \mid X = x) = h[s(x)]$ where h is the distribution function of $-\varepsilon$, and so is equivalent to our framework with $r(x) = E(Y \mid X = x)$ and w empty. In an unpublished manuscript, Newey and Matzkin (1993) propose an estimator of Matzkin's (1992) model. Their estimator imposes the normalization $g(z_0) = 1$, estimates h using $E(Y \mid V = v, Z = z_0) = h(vg(z_0)) = h(v)$, and they essentially invert the corresponding h function estimate to obtain $s(x)$ from $r(x) = h[s(x)]$. Advantages of our estimators are that they can include w, they converge at a faster rate, they

include functions other than conditional means for r, they can attain oracle efficiency for s, and they do not depend upon a single arbitrarily chosen point z_0.

Models satisfying equation (4) without imposing homogeneity on s are called weakly separable. See Gorman (1959), Goldman and Uzawa (1964) and Blackorby, Primont, and Russell (1978). Pinkse (2001) provides a general nonparametric estimator of weakly separable models. Pinkse's estimator identifies $s(x)$ up to an arbitrary monotonic transformation, whereas our estimator provides the unique (up to scale) linear homogeneous $s(x) = vg(z)$ (or equivalently g up to location in equation 2)) and exploits this structure of $s(x)$ to obtain a faster rate of convergence than Pinkse.

Many estimators exist for strongly or additively separable models, which are models of the form $E(Y|x) = \sum_k s_k(x_k)$ where the functions $s_k(x_k)$ are unknown, and for generalized additively separable models, defined as $r(x) = h[\sum_k s_k(x_k)]$. Those most closely resembling our model include Härdle, Kim, and Tripathi (2001), who estimate additively separable models where the $s_k(x_k)$ functions are homogeneous, and Horowitz (2001) and Horowitz and Mammen (2005) who estimate generalized additively separable models where both h and s_k are unknown functions.

Matzkin (2003) considers models of the form $y = m(x, \varepsilon)$ with an unobserved scalar ε independent of x and, as one possible identifying assumption, m being linearly homogeneous in x and ε. In contrast, our model makes no assumptions about (and provides no estimates of) the role of unobservables other than a limiting distribution theory for an estimate of R, and allows for homothetic rather than just homogeneous dependence on x.

2 Informal Description of the Estimators

Since v is observed and the function B_2 is known, we may without loss of generality rewrite equation (1) as $R(v, z, w) = h[B_1(vB_3(g(z)), w]$ by redefining v as $B_2(v)$. Next, by defining $H(B_3, w) = h(B_1(B_3), w)$ and $G(z) = B_3(g(z))$, we may again without loss of generality rewrite equation (1) as

$$R(v, z, w) = H[vG(z), w] \tag{5}$$

We start with a consistent estimator $\widehat{R}(\cdot)$ of the function $R(\cdot)$, and provide nonparametric estimators for G and H. Estimates of the original g and h can then be readily recovered from the estimates of G and H if desired.

We could have instead started with the form $R(v, z, w) = h[B_1(B_2(v) + B_3(g(z)), w]$, simplifying as above to $R(v, z, w) = H[v + G(z), w]$, but this is slightly less general, because e.g., it only includes equation (3) as a special case when $vg(z)$ is constrained to be positive.

We first construct an initial consistent estimator of $G(z)$ by matching. For given values v, z, z', w suppose we can find a scalar u such that $R(v, z, w) = R(vu, z', w)$, a match. Then $u = U(z, z') =$

$G(z)/G(z')$. The function $U(z,z')$ can be estimated by finding a zero of the function $\widehat{R}(v,z,w) - \widehat{R}(vu,z',w)$, averaging over a range of values of v and w to improve convergence properties. The function $G(z)$ is then estimated using a sample analog of $G(z) = U(z,z')/E[U(Z,z')]$ (averaged over a range of values of z'), which holds given the free scale normalization $E[G(Z)] = 1$. One advantage of a scale normalization like this over more simply normalizing at a point like $G(z_0) = 1$ is that the resulting limiting distributions at every point z will then not depend upon the distribution of $\widehat{R}(v,z_0,w)$.

Given the function G, the function H can be defined as the conditional expectation

$$H(\gamma,w) = E\left[R(V,Z,W) \mid VG(Z) = \gamma, W = w\right]. \tag{6}$$

Therefore, given an estimate $\widehat{G}$ of G, we can estimate the function H by a regression smooth of $\widehat{R}_i$ on $V_i\widehat{G}(Z_i), W_i$.

The above steps summarize our sequential, matching-based estimator. The simultaneous estimator begins by defining the functions G and H as minimizers of $E\{R(V,Z,W) - H(VG(Z),W)\}^2]$ subject to a normalization constraint $E[G(Z)] = 1$. This is analogous to least squares estimation of parametric models, and corresponds to the common definition of a regression function $E(Y \mid X)$ as the minimizer of $E[(Y - m(X))^2]$ over all measurable functions m. We derive a representation of the first order conditions that G and H must satisfy from the Lagrangian associated with this constrained minimization. One of these first order conditions is just equation (6). For any given z, the other condition can be conveniently expressed as $G(z) = s$ where s is the solution to

$$E[\zeta(V,Z,W,s) \mid Z = z] - E[\zeta(V,Z,W,G(Z))G(Z)] = 0 \tag{7}$$

$$\zeta(V,Z,W,s) = \left[R(V,Z,W) - H\left(Vs,W\right)\right] \frac{\partial H}{\partial \gamma}\left(VG(Z),W\right) V. \tag{8}$$

Our simultaneous estimator is based on this representation of these first order conditions. We first use the sequential estimator to obtain initial consistent estimates of G, H, and $\partial H/\partial\gamma$, denoted by $\widehat{G}$, $\widehat{H}$, and $\partial\widehat{H}/\partial\gamma$. Then we define an empirical analogue of equation (7) and (8), which could be numerically solved for s to yield an estimate of $G(z)$. To simplify computation, we linearize this expression in s and solve the resulting equation explicitly for s to yield an estimator of $G(z)$. We then estimate H given our estimate of G as before, corresponding to the first order condition (6), and iterate.

3 Identification

ASSUMPTION A. *For some set $\Psi_{v,z,w}$, there exist functions R, H, and G, such that $R(v,z,w) = H(vG(z),w)$ for all $(v,z,w) \in \Psi_{v,z,w}$. Let $\Psi_{\gamma,w} = \{(\gamma,w) \mid \gamma = vG(z)$ and $(v,z,w) \in \Psi_{v,z,w}\}$. Let*

Ψ_z be the set of all z such that there exists a v, w for which $(v, z, w) \in \Psi_{v,z,w}$. Let Ψ_z^* be a nonempty set such that $\Psi_z^* \subset \Psi_z$ and for all $z' \in \Psi_z^*$, $G(z') \neq 0$. For each $(z, z') \in \Psi_z \times \Psi_z^*$, define the set $\Psi_{v,w|z,z'} = \{(v, w) \mid (v, z, w) \in \Psi_{v,z,w}$, $vG(z)/G(z'), z', w \in \Psi_{v,z,w}$, $v \neq 0$, and $H(vG(z), w)$ exists and is invertible on its first element$\}$. For all $(z, z') \in \Psi_z \times \Psi_z^*$, the set $\Psi_{v,w|z,z'}$ is nonempty. Without loss of generality, normalize the scale of G such that $\int_{z \in \Psi_z} G(z) F(dz) = 1$ where $F(dz)$ is a measure with support Ψ_z.

THEOREM 1. *Let Assumption A hold. For every $(z, z') \in \Psi_z \times \Psi_z^*$ there exists a unique $U(z, z')$ such that, for all $(v, w) \in \Psi_{v,w|z,z'}$ the equality $R(v, z, w) = R(U(z, z') v, z', w)$ holds. For every $(z, z') \in \Psi_z \times \Psi_z^*$ the function $G(z)$ satisfies*

$$G(z) = \left[\int_{z'' \in \Psi_z} U(z'', z') F(dz'') \right]^{-1} U(z, z') \tag{9}$$

and for all $(\gamma, w) \in \Psi_{\gamma,w}$, the function H satisfies

$$H(\gamma, w) = E[R(V, Z, W) \mid VG(Z) = \gamma, W = w]. \tag{10}$$

PROOF OF THEOREM 1. Having $R(v, z, w) = R(U(z, z') v, z', w)$ hold for $(v, w) \in \Psi_{v,w|z,z'}$ means that $H[vG(z), w] = H[U(z, z') vG(z'), w]$, where, at these values, the function H exists and is invertible on its first element. This equation therefore holds if and only if $vG(z) = U(z, z') vG(z')$, which requires either $U(z, z') = G(z)/G(z')$ or $v = 0$. The latter is ruled out in $\Psi_{v,w|z,z'}$, so $U(z, z')$ is uniquely given by $U(z, z') = G(z)/G(z')$, and this result holds for all $z, z' \in \Psi_z \times \Psi_z^*$. The scale normalization then gives $\int_{z'' \in \Psi_z} U(z'', z') F(dz'') = \int_{z'' \in \Psi_z} [G(z'')/G(z')] F(dz'') = 1/G(z')$ and the expression for $G(z)$ follows immediately. The function $H(\gamma, w)$ exists at $\gamma = vG(z) = U(z, z') vG(z')$, and the equation given for H follows from the definition of H. ■

The scale normalization of G in Assumption A is without loss of generality, because one may always redefine $G(z)$ as $cG(z)$ for $c \neq 0$ by redefining $H(\gamma, w)$ as $H(\gamma/c, w)$. In our application we will take F to be the distribution function of Z so the normalization is $E[G(Z)] = 1$.

By Theorem 1, $G(z)$ is identified even when Ψ_z^* and $\Psi_{v,w|z,z'}$ are singletons, but identification fails if no z' exists that yields a nonempty set $\Psi_{v,w|z,z'}$. Overidentifying information results when these sets have multiple elements.

4 Estimation

We suppose that we have a sample of data $\{V_i, Z_i, W_i, i = 1, \ldots, n\}$ and that we have an estimator $\widehat{R}(v, z, w)$ of the function $R(v, z, w)$ for all relevant values of v, z, w. This may have been computed

from data on additional variables, generically denoted Y_i, but we do not need to specify $\widehat{R}(v,z,w)$ so specifically. Let U_0, G_0, and H_0, denote the unknown true functions U, G, and H. Let $k(\cdot)$ be a univariate kernel function, and for any vector $s \in \mathbb{R}^{\dim(s)}$, let $K^s(u) = k(u_1) \times \cdots k(u_{\dim(s)})$ denote a suitable product kernel.

4.1 Matching Based Sequential Estimation of $\mathbf{G_0}$,$\mathbf{H_0}$

4.1.1 Estimation of $\mathbf{U_0}$

For estimation, we translate the matching concept into a moment condition. Define $U_0(z,z')$ as the value of U that solves $m_1(U;z,z') = 0$, where

$$m_1(U;z,z') = \int [R(v,z,w) - R(Uv,z',w)]\pi(dv,dw \mid z,z')$$

for each z, z' for some measure $\pi(dv,dw \mid z,z')$ that has support contained in $\Psi_{v,w|z,z'}$. We concentrate on the case where $\pi(dv,dw \mid z,z') = \pi(v,w \mid z,z')dvdw$ for some conditional density function $\pi(v,w \mid z,z')$ with non-trivial support, because the averaging can yield improved rates of convergence, see inter alia Linton and Nielsen (1995). In particular, for simplicity we take $\pi(dv,dw \mid z,z')$ to be $\pi(v,w) = f_{v,w}(v,w)1[(v,w) \in A]$ for some fixed set $A \subset \cap_{z,z' \in \Psi_z \times \Psi_z^*} \Psi_{v,w|z,z'}$, where $1[\cdot]$ is the indicator function. This set A does not vary with z, z'.

In practice we replace unknown quantities by estimators whence we obtain the sample moment equation $\widehat{m}_1(U;z,z') = \int [\widehat{R}(v,z,w) - \widehat{R}(Uv,z',w)]\widehat{\pi}(v,w)dvdw$, where $\widehat{R}$ is an estimate of R and $\widehat{\pi}$ is an estimate of π. The integral can be approximated by a variety of numerical methods. For example, one can use the sample observations themselves and compute the sample moment function

$$\widehat{m}_1(U;z,z') = \frac{1}{n}\sum_{i=1}^{n} [\widehat{R}(V_i,z,W_i) - \widehat{R}(UV_i,z',W_i)]1[(V_i,W_i) \in A]. \tag{11}$$

We work with this definition of $\widehat{m}_1(U;z,z')$. Define the estimator $\widehat{U}(z,z')$ for each z, z' to be any value such that

$$|\widehat{m}_1(\widehat{U}(z,z');z,z')| \leq \inf_{U \in \mathcal{U}} |\widehat{m}_1(U;z,z')| + o_p(n^{-1/2}), \tag{12}$$

where $\mathcal{U}$ is a subset of the values that $G(z)/G(z')$ could take on. This is a nonlinear optimization problem, although the parameter U is a scalar so that (11) can be computed by grid search with high accuracy.

4.1.2 Estimation of G_0

To identify $G_0(z)$ we shall make use of our normalization condition that $E[G_0(Z)] = \int G_0(z)f_Z(z)dz = 1$, where $f_Z(z)$ is the marginal density of z. It follows from Theorem 1 that $G_0(z) = \int U_0(z,z')\varpi(dz')/$

$\int U_0(z,z')\varpi(dz')f_Z(z)dz$, where $\varpi(dz')$ is any measure with support in Ψ_z. Specifically, $\varpi(dz')dz'$ could be the point mass at some point z_0 or $\varpi(dz') = \varpi(z')dz'$ with ϖ a density function on some non-trivial interval. Based on this equation we estimate $G_0(z)$ by

$$\widehat{G}(z) = \frac{\frac{1}{n}\sum_{i=1}^n \widehat{U}(z,Z_i)\varpi_f(Z_i)}{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \widehat{U}(Z_i,Z_j)\varpi_f(Z_j)}, \tag{13}$$

where $\varpi_f(z)$ is a weighting function such that $E[g^*(Z)\varpi_f(Z)] = \int g^*(z)\varpi(z)dz$ for any measurable function g^*. This estimator automatically satisfies $n^{-1}\sum_{i=1}^n \widehat{G}(Z_i) = 1$.

4.1.3 Estimation of H_0

Given the function $G_0(\cdot)$, the function $H_0(\cdot)$ is defined by equation (10). Given an estimate $\widehat{G}(\cdot)$ of $G_0(\cdot)$, we estimate $H_0(\cdot)$ by a regression smooth of $\widehat{R}_i$ on $V_i\widehat{G}(Z_i), W_i$. We use a class of kernel smoothers. Let $\gamma = vG_0(z)$, $c = (\gamma, w)$, $\widehat{\gamma} = v\widehat{G}(z)$, $\widehat{\gamma}_i = V_i\widehat{G}(Z_i)$, and let $\widehat{c} = (\widehat{\gamma}, w)$ and $\widehat{C}_i = (\widehat{\gamma}_i, W_i)$. Define the sample moment function

$$\widehat{m}_3(H;c) = \frac{1}{nb_H^{d_W+1}}\sum_{i=1}^n K^c\left(\frac{c-\widehat{C}_i}{b_H}\right)\psi\left(\widehat{R}_i - H\right), \tag{14}$$

where b_H is some bandwidth sequence and K^c is a d_W+1-dimensional product kernel. Here, $\widehat{R}_i = \widehat{R}(V_i, Z_i, W_i)$ is an estimator of R_i, while ψ is a twice continuously differentiable function with $\psi(0) = 0$ and $\psi'(0) \neq 0$. Define the estimator of $H(c)$ to be any sequence $\widehat{H}(c)$ of approximate zeros of (14) satisfying

$$|\widehat{m}_3(\widehat{H}(c);c)| \leq \inf_{H\in\mathcal{H}}|\widehat{m}_3(H;c)| + o_p(n^{-1/2}), \tag{15}$$

where $\mathcal{H}$ is some set. If $\psi(x) = x$ we obtain the standard Nadaraya-Watson kernel regression smooth of $\widehat{R}_i$ on $\widehat{C}_i$.

We suppose that $\widehat{C}_i$ in (14) is computed as in (13) with the bandwidth b_G, but that $\widehat{R}_i$ is computed with a different, 'small', bandwidth b_0. The extra bandwidth b_0 does not play an important role in comparison with b_H and b_G and we shall assume that it is smaller in magnitude. If Y_i is observed and $R_i = E[Y_i|V_i, Z_i, W_i]$, then we can replace $\widehat{R}_i$ in (14) by Y_i, which amounts to taking $b_0 = 0$.

We also define an estimator of $\partial H(c)/\partial\gamma$ by differentiating $\widehat{H}(\widehat{c})$ with respect to γ and denote this by $\partial\widehat{H}(\widehat{c})/\partial\gamma$. Alternatively, one can use a local polynomial method and take the coefficient on $(\widehat{\gamma}_i - \widehat{\gamma})$ as the estimate of $\partial H(\widehat{c})/\partial\gamma$.

4.2 Simultaneous Estimation of H_0 and G_0

Our strategy for improving the efficiency of the estimators we defined above is based on using a more general definition of the functions $H_0(\cdot)$ and $G_0(\cdot)$. They can be defined as minimizers of the

functional

$$E\{R(V,Z,W) - H(VG(Z),W)\}^2] = \int [R(v,z,w) - H(vG(z),w)]^2 f_X(v,z,w)dvdzdw \tag{16}$$

subject to the restriction that $E[G(Z)] = 1$, where $f_X(v,z,w)$ is the joint density of the random variables $X = (V,Z,W)$. For the remainder of the paper, we now let X contain W, unlike the introduction, since this will ease later notation. When Y_i is observed and R is the regression of Y on V,Z,W, one could replace the criterion (16) by the more standard $E[\{Y - H(VG(Z),W)\}^2]$. The criterion and the subsequent first order conditions are the same by iterated expectation. This simultaneous definition of the functional parameters $H_0(\cdot)$ and $G_0(\cdot)$ as minimizers of a population objective function is natural and is used in other contexts. See Mammen, Linton, and Nielsen (1999) for a discussion in the context of additive nonparametric regression.

To find a characterization of the solutions to (16) we follow Weinstock (1952, Chapter 4) in our treatment. Define the objective functional

$$\mathcal{L}(H,G,\lambda) = \int [R(v,z,w) - H(vG(z),w)]^2 f_X(v,z,w)dvdzdw + \lambda \left[\int G(z)f_Z(z)dz - 1\right]$$

for each H, G, λ. Letting $G(\cdot) = G_0(\cdot) + \epsilon\tau(\cdot)$ and $H(\cdot) = H_0(\cdot) + \delta\eta(\cdot)$ we find the following first order conditions:

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}(H_0 + \delta\eta, G_0 + \epsilon\tau, \lambda)}{\partial \delta}\Big|_{\downarrow \epsilon=0,\delta=0} \\ &= -\int [R(v,z,w) - H_0(vG_0(z),w)]\,\eta(vG_0(z),w)\, f_X(v,z,w)dvdzdw \end{aligned} \tag{17}$$

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}(H_0 + \delta\eta, G_0 + \epsilon\tau)}{\partial \epsilon}\Big|_{\downarrow \epsilon=0,\delta=0} \\ &= -\int [R(v,z,w) - H_0(vG_0(z),w)]\,\frac{\partial H_0}{\partial \gamma}(vG_0(z),w)\, v\tau(z) f_X(v,z,w)dvdzdw + \lambda \int \tau(z) f_Z(z)dz, \end{aligned} \tag{18}$$

for all measurable and smooth test functions η, τ for which these expectations are well-defined, which is a necessary condition for a local minimum, see Sagan (1969), Theorem 1.7 for example. By invoking the Euler-Lagrange theorem and using the law of iterated expectation we obtain the necessary condition

$$\mathcal{L}_H(H_0,G_0)(s,u) = -E[\{R(V,Z,W) - H_0(VG_0(Z),W)\} \mid VG_0(Z) = s, W = u]\, f_{vG_0(z),w}(s,u) = 0 \tag{19}$$

corresponding to (17), where $f_{vG_0(z),w}(s,u)$ is the density function of the random variable $(VG_0(Z),W)$ [heuristically, this can be seen by setting the directions to be the Dirac deltas $\tau(z) = 1(z = t)$ and $\eta(vG_0(z),w) = 1(vG_0(z) = s, w = u)$]. For equation (18), we obtain the necessary condition that

$$E\left[\{R(V,Z,W) - H_0(VG_0(Z),W)\}\frac{\partial H_0}{\partial \gamma}(VG_0(Z),W)\, V \mid Z = t\right] f_Z(t) = \lambda f_Z(t) \tag{20}$$

for all t. The equations (19) and (20) and the constraint $\int G_0(z) f_Z(z) dz = 1$ determine the system. Multiplying (20) by $G_0(t)$ and integrating over t and using the law of iterated expectations we obtain

$$\lambda = E\left[[R(V,Z,W) - H_0\left(VG_0(Z),W\right)] \frac{\partial H_0}{\partial \gamma}\left(VG_0(Z),W\right) VG_0(Z) \right].$$

Then substituting into (20) and dividing through by $f_Z(t)$ we obtain the equation (for all t):

$$\begin{aligned} &E\left[[R(V,Z,W) - H_0\left(VG_0(Z),W\right)] \frac{\partial H_0}{\partial \gamma}\left(VG_0(Z),W\right) V \mid Z = t \right] \\ &-E\left[[R(V,Z,W) - H_0\left(VG_0(Z),W\right)] \frac{\partial H_0}{\partial \gamma}\left(VG_0(Z),W\right) VG_0(Z) \right] = 0. \end{aligned} \tag{21}$$

Equation (19) is linear in H_0 given G_0, and we obtain $H_0(s,u) = E\left[R(V,Z,W) \mid VG_0(Z) = s, W = u\right]$. Equation (21) is non-linear in G_0 even given H_0; also the second term makes (21) an integral equation in $G_0(\cdot)$. One could try to solve empirical versions of (19) and (21). See for comparison Linton and Mammen (2005) who work only with linearized integral equations around an initial consistent estimator. Instead we shall pursue a strategy that makes use of the preliminary estimators obtained previously and does not require the solution of an integral equation.

Letting $\zeta_i(\Gamma) = [R(V_i,Z_i,W_i) - H\left(V_i\Gamma,W_i\right)](\partial H/\partial\gamma)(v_i G_0(Z_i),W_i)V_i$, the conditions (21) can be represented as $E[\zeta_i(G_0(Z_i))|Z_i = t] - E[\zeta_i(G_0(Z_i))G_0(Z_i)] = 0$. Denote consistent estimators of G_0, H_0, and $\partial H_0/\partial\gamma$ by $\widehat{G}$, $\widehat{H}$, and $\partial\widehat{H}/\partial\gamma$ respectively. Let t_{ni} denote a trimming sequence that is needed to ensure that $\widehat{H}$ and $\partial\widehat{H}/\partial\gamma$ are computed at interior points. Define the sample moment function

$$\widehat{m}_4(\Gamma;z) = \frac{1}{nb_1^{d_Z}} \sum_{i=1}^n K^z\left(\frac{z - Z_i}{b_1}\right) \left[\widehat{\zeta}_i(\Gamma) - \frac{1}{n}\sum_{i=1}^n \widehat{\zeta}_i(\widehat{G}(Z_i))\widehat{G}(Z_i) \right] \tag{22}$$

$$\widehat{\zeta}_i(\Gamma) = \left[\widehat{R}(V_i,Z_i,W_i) - \widehat{H}\left(V_i\Gamma,W_i\right)\right] \frac{\partial\widehat{H}}{\partial\gamma}\left(V_i\widehat{G}(Z_i),W_i\right) V_i t_{ni},$$

where b_1 is a bandwidth and K^z is a d_Z-dimensional kernel. Define the estimator $\widetilde{G}(z)$ for each z to be any value such that

$$|\widehat{m}_4(\widetilde{G}(z);z)| \leq \inf_{\Gamma\in\mathcal{G}} |\widehat{m}_4(\Gamma;z)| + o_p(n^{-1/2}), \tag{23}$$

where the set $\mathcal{G}$ can be chosen to be a small or even shrinking neighborhood of $\widehat{G}(z)$. The moment condition $\widehat{m}_4(\Gamma;z)$ is like the numerator of a regression smooth of $\widehat{\zeta}_i(\Gamma) - n^{-1}\sum_{i=1}^n \widehat{\zeta}_i(\widehat{G}(Z_i))\widehat{G}(Z_i)$ on Z_i, in this case a kernel regression smooth; it can be considered as an approximation to (21) multiplied by $f_Z(z)$. This approximation to (21) is chosen for convenience. In particular, we replace Γ by the preliminary estimator everywhere except inside the basic residual $\widehat{R}(V_i,Z_i,W_i) - \widehat{H}\left(V_i\Gamma,W_i\right)$, as in Hastie and Tibshirani (1990). Note that if Y_i were observed and $R(V_i,Z_i,W_i) = E[Y_i|V_i,Z_i,W_i]$, then

one can replace $\widehat{R}(V_i, Z_i, W_i)$ in $\widehat{\zeta}_i(\Gamma)$ by Y_i. The bandwidth b_1 does not play a big role in the sequel and we shall assume as above that it is smaller in magnitude than the other smoothing parameters.

Given $\widetilde{G}$, we then compute $\widetilde{H}$ as the nonparametric regression of $\widehat{R}_i$ on $V_i\widetilde{G}(Z_i)$. We still call the estimators $\widetilde{G}$ and $\widetilde{H}$ simultaneous because they are based on the simultaneous definition of G_0, H_0 given above, in particular, the estimator $\widetilde{G}(z)$ makes use of information about $H_0(\cdot)$.

4.3 Discussion

Our estimators require selection of sets Ψ_z^* and A (or more generally $\Psi_{v,w|z,z'}$), that is, the sets to use for matching and averaging. For efficiency of the sequential estimator it is desirable to average over large sets, but this is of less importance if they are just being used to generate starting values for the simultaneous estimator.

One procedure for selecting these sets is to search over the data to find observations j and k such that, for each observation i, there exists a nonzero match U_i where $\widehat{R}(V_j, Z_i, W_j) = \widehat{R}(V_jU_i, Z_k, W_j)$ and both V_j, Z_i, W_j and V_jU_i, Z_k, W_j are in neighborhoods of observed data. One could then take Ψ_z^* to be a neighborhood of the union of all such Z_k and A to be a neighborhood of the union of all such V_j, W_j. Alternatively, one could just search for a single observation k such that, for each i, $\widehat{R}(V_i, Z_i, W_i) \simeq \widehat{R}(V_iU_i, Z_k, W_i)$ and V_iU_i, Z_k, W_i is in a neighborhood of observed data, then take Ψ_z^* to be a neighborhood of z_k and let $\Psi_{v,w|Z_i,z'}$ be a neighborhood of V_i, W_i. Consistency of the initial $\widehat{G}(Z_i)$ estimator doesn't require these sets to have positive measure, e.g., they could just be the singletons $\Psi_z^* = \{Z_k\}$ and $\Psi_{v,w|Z_i,z'} = \{V_i, W_i\}$.

If matching on w is a problem then one could first replace $\widehat{R}(v, z, w)$ with $n^{-1}\sum_{i=1}^n \widehat{R}(v, z, W_i)$ and, in the theorems, replace $R(v, z, w)$ with $E[R(V, Z, W) \mid V, Z]$ obtaining the estimator $\widehat{U}$ with the result, as if there were no w. Then, once the initial consistent $\widehat{G}$ is obtained from this $\widehat{U}$, go back to using the original $\widehat{R}$ for estimating $\widehat{H}$ and for the simultaneous estimator.

5 Distribution Theory

In the working paper version of this article, Lewbel and Linton (2005), we provide the pointwise asymptotic distribution of our estimators $\widehat{U}(z, z')$, $\widehat{G}(z)$, $\widehat{H}(v\widehat{G}(z), w)$, $\widetilde{G}(z)$ and $\widetilde{H}(v\widetilde{G}(z), w)$. Our strategy is to write the estimators as solving a sample first order condition and then to employ a general theory we develop for this type of procedure. We now summarize without proof the limiting distributions of the sequential estimators.

We assume that our estimator $\widehat{R}(\cdot)$ of $R(\cdot)$ satisfies an asymptotic expansion but are not more specific about how the estimator is defined. This generality is useful because the target function $R(\cdot)$

could be a variety of things depending on the application and a variety of estimation strategies could be employed for $\widehat{R}(\cdot)$. Define for any vector $\alpha = (\alpha_1, \ldots, \alpha_d)^\top$ and function $f\colon \mathbb{R}^d \to \mathbb{R}$

$$D^\alpha f(x) = \frac{\partial^{|\alpha|} f(x)}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \text{ with } |\alpha| = \sum_{j=1}^{d} \alpha_j.$$

Assumption B.

B1. *The random variables* (V_i, Z_i, W_i), $i = 1, \ldots, n$ *are independent and identically distributed. Let* $x = (v, z, w) \in \mathbb{R}^d$ *and let* $f_X(x)$ *be the joint density function of* $X_i = (V_i, Z_i, W_i)$ *with support* $\Psi_X = \Psi_v \times \Psi_z \times \Psi_W$ *a compact subset of* $\mathbb{R}^d$. *We assume without loss of generality that* $V_i \geq 0$.

B2. (a) *The functions* H_0 *and* G_0 *are* p*-times continuously partially differentiable in all arguments, which implies that* $D^\alpha R(x)$ *exists and is continuous on* Ψ_X *for all* α *with* $|\alpha| \leq p$; *(b) The function* H *satisfies* $\inf_{\gamma \in \Psi_\gamma, w \in \Psi_W} |\partial H(\gamma, w)/\partial \gamma| > 0$, *where* $\Psi_\gamma = \{\gamma = vG_0(z) : (v, z) \in \Psi_v \times \Psi_z\}$; *(c) The function* G_0 *satisfies* $\inf_{z \in \Psi_z} |G_0(z)| > 0$.

B3. *The estimator* $\widehat{R}(x)$ *satisfies the uniform asymptotic expansion as* $n \to \infty$

$$\widehat{R}(x) - R(x) = \frac{1}{nb_G^d} \sum_{i=1}^{n} \left[\sum_{j=1}^{J} a_{nj}(x) K_j \left(\frac{x - X_i}{b_G} \right) \right] u_i + b_G^p \beta_R(x) + \mathcal{R}_n(x), \tag{24}$$

where the components of (24) have the following properties: (a) The random variables (u_i, X_i) *are i.i.d. with* $E(u_i|X_i) = 0$ *a.s. and* $\sup_{x \in \Psi_X} E(|u_i|^\kappa | X_i = x) < \infty$ *for some* $\kappa > 2$. *The function* $\sigma^2(x) = var(u_i|X_i = x)$ *is continuous a.s.; (b) The deterministic functions* $a_{nj}(x)$, $j = 1, \ldots, J$, *satisfy for all vectors* α *with* $|\alpha| \leq 1$: $\lim_{n \to \infty} \sup_{x \in \Psi_X} |D^\alpha a_{nj}(x) - D^\alpha a_j(x)| = 0$, *where* $D^\alpha a_j(x)$ *are bounded and continuous on* Ψ_X. *The non-random functions* $a_j(x) = a_j f_X^{-1}(x)$, *where* a_j *are constants depending only on* $K_1, \ldots, K_J$; *(c) The functions* K_j *take the product form* $K_j(u) = k_{j1}(u_1) \times \cdots \times k_{jd}(u_d) = K_j^v(u_v) K_j^z(u_z) K_j^w(u_w)$, *grouping terms in an obvious way. Here,* k_{jl} *have compact support and are twice continuously differentiable; (d) (i) The bandwidth satisfies* $b_G = \lambda_G n^{-1/(2p + d_Z)}$ *for some* λ_G *with* $0 < \lambda_G < \infty$; *(ii)* $p > d_W + 5/2$; *(e) The non-random function* $\beta_R(\cdot)$ *is continuous on* Ψ_X; *(f) The remainder term satisfies*

$$\sup_{x \in \Psi_X} |\mathcal{R}_n(x)| = o_p(b_G^p) + O_p\left(\frac{\log n}{nb_G^d} \right). \tag{25}$$

B4. *The weight functions* π, ϖ *are continuous on their supports. The set* A *has non-zero Lebesgue measure.*

The working paper version of this article discusses this and other assumptions in detail. For now we simply note that these assumptions allow $\widehat{R}(x)$ to be a nonparametric kernel regression or kernel density estimator, or a local polynomial nonparametric regression estimator as in Fan and

Gijbels (1996) and Masry (1996), or a local nonlinear least squares estimator as in Gozalo and Linton (2000), or a conditional cumulative distribution function estimator of, say, $\Pr(Y \leq y|X = x)$ at some point y. There is no requirement that $\widehat{R}(x)$ is smooth or even continuous in x, only that it is well approximated by a function (the left hand side of (24)) that is smooth.

Assumption B is sufficient to ensure that $\sup_{x\in\Psi_x} |\widehat{R}(x) - R(x)| = O_p((\log n/nb_G^d)^{1/2}) + O_p(b_G^p)$, and Assumptions A and B suffice for deriving a limiting normal distribution for the matching estimators $\widehat{U}(z,z')$ and $\widehat{G}(z)$. The estimators $\widehat{U}(z,z')$ and $\widehat{G}(z)$ converge to $U_0(z,z')$ and $G_0(z)$ at a rate $n^{p/(2p+d_Z)}$ under our assumptions and this is the optimal pointwise rate of convergence for nonparametric functions of dimension d_Z and smoothness p, Stone (1980), and so would be the optimal rate for G_0 when H_0 is known.

ASSUMPTION C. 1. *Let $f_{\gamma,z,W}$ be the density of (γ_i, Z_i, W_i) with marginal density f_C the density of $C_i = (\gamma_i, W_i)$ with support Ψ_C, where $C_i = (\gamma_i, W_i)$, and $\gamma_i = V_i G_0(Z_i)$. Suppose that $c = (vG_0(z), w)$ is an interior point of Ψ_C for which $f_C(c) > 0$. 2. The kernel k has compact support, is twice continuously differentiable, and is of order p, that is, $\int t^j k(t)dt = 0$ for all $j \leq p$. 3. The bandwidth $b_H = \lambda_H n^{-1/(2p+d_W+1)}$ for some λ_H with $0 < \lambda_H < \infty$. 4. $b_0/\min\{b_G, b_H\} \to 0$ and $n^{p+\min\{d_W+1,d_Z\}} b_0^d/(\log n)^2 \to \infty$.*

For some constants κ_2, κ_3 depending on $K_1, \ldots, K_J$, and λ_G, λ_H (these are defined in the working paper version), define:

$$\Omega_2(z) = \frac{\kappa_2 \int \frac{\pi^2(v,w)\sigma^2(v,z,w)}{f_X(v,z,w)} dvdw}{\left(\int \frac{\partial H_0}{\partial \gamma}(vG_0(z), w) v\pi(v,w) dvdw\right)^2} \tag{26}$$

$$\Omega_3(c) = \left[\frac{\partial H_0}{\partial \gamma}(c)v\right]^2 \Omega_2(z) 1(d_W + 1 \geq d_Z) + \frac{\kappa_3 E\left[\sigma^2(Z)|C = c\right]}{\psi'(0)^2 f_C(c)} 1(d_W + 1 \leq d_Z).$$

THEOREM 2. *Suppose that assumptions A and B hold and that z is an interior point of Ψ_Z. Then, there exists a bounded continuous function $\beta_2(z)$ such that as $n \to \infty$ with $\delta_{nG} = n^{p/(2p+d_Z)}$,*

$$\delta_{nG}\left[\widehat{G}(z) - G_0(z) - b^p \beta_2(z)\right] \Longrightarrow N(0, \Omega_2(z)).$$

Suppose that also assumption C holds. Then, there exists bounded continuous functions $\beta_{3G}(c), \beta_{3H}(c)$ such that as $n \to \infty$ with $\delta_{nH} = \min\{n^{p/(2p+d_Z)}, n^{p/(2p+d_W+1)}\}$,

$$\delta_{nH}\left[\widehat{H}(\widehat{c}) - H_0(c) - b_G^p \beta_{3G}(c) - b_H^p \beta_{3H}(c)\right] \Longrightarrow N(0, \Omega_3(c)).$$

To save space we have omitted explicit expressions for the bias terms in Theorem 2. These are in general quite complicated, although they simplify when the first stage estimates $\widehat{G}$ and $\widehat{H}$ are

undersmoothed. Consistent standard errors can be obtained by an obvious plug-in method, which is defined explicitly in the working paper.

The limiting distribution for $\widehat{g}(z) = B_3^{-1}[\widehat{G}(z)]$ is immediately obtained from Theorem 2 using the delta method, and for $\widehat{H}[B_1(vG(z)), w] = h(vG(z), w)$ by redefining γ as $B_1[vG(z)]$, or equivalently by the same derivation as in Theorem 2, defining $\widehat{H}$ as the regression smooth of $\widehat{R}$ on $B_1[v\widehat{G}(z)], w$.

5.1 Efficiency

We give the distribution theory for the efficient estimators $\widetilde{G}(z)$ and $\widetilde{H}(\widetilde{c})$ in the working paper. Additional conditions are required on the trimming, and on the kernel K^z and bandwidth b_1 in (22). One obtains pointwise asymptotic normality for the centred estimators at the same rates δ_{nG} and δ_{nH} as in Theorem 2 with asymptotic variances:

$$\begin{aligned}
\Omega_4(z) &= \frac{\kappa_2 \int \sigma^2(v,z,w) \left[\frac{\partial H_0}{\partial \gamma}(vG_0(z),w)\right]^2 v^2 f_X(v,z,w)dvdw}{\left(\int \left[\frac{\partial H_0}{\partial \gamma}(vG_0(z),w)\right]^2 v^2 f_X(v,z,w)dvdw\right)^2} \\
\Omega_5(c) &= \left[\frac{\partial H_0}{\partial \gamma}(c)v\right]^2 \Omega_4(z) 1(d_W + 1 \geq d_Z) + \frac{\kappa_3 E\left[\sigma^2(Z)|C=c\right]}{\psi'(0)^2 f_C(c)} 1(d_W + 1 \leq d_Z).
\end{aligned}$$

By the Cauchy-Schwarz inequality, the simultaneous estimators $\widetilde{G}(z)$ and $\widetilde{H}(\widetilde{c})$ are at least as efficient under homoskedasticity, i.e., $\sigma^2(v,z,w) = \sigma^2$, as the sequential estimators $\widehat{G}(z)$ and $\widehat{H}(\widehat{c})$. Furthermore, they can be oracle efficient, Linton (1996), as we next discuss. Suppose that $G_0(\cdot)$ was defined as the minimizer (for known $H_0(\cdot)$) of $E[\{R(V,Z,W) - H_0(VG(Z),W)\}^2]$ with respect to $G(\cdot)$ subject to the constraint that $E[G(Z)] = 1$. This leads to the sample moment condition for each z

$$\widetilde{m}_4(\Gamma; z) = \frac{1}{n}\sum_{i=1}^n K^z\left(\frac{z - Z_i}{b}\right)\{Y_i - H_0(V_i\Gamma, W_i)\}\frac{\partial H_0}{\partial \gamma}(V_i G_0(Z_i), W_i) V_i,$$

where K^z and b are a generic kernel and bandwidth. Let $\overline{G}(z)$ be the estimator that is any approximate zero of $\widetilde{m}_4(\Gamma; z)$. This estimator is a natural benchmark against which to measure the performance of our estimators. The distribution theory for $\overline{G}(z)$ follows from arguments of Gozalo and Linton (2000), and to first order, the distribution of our feasible estimator $\widetilde{G}(z)$ is equivalent to the distribution of $\overline{G}(z)$. When first stage estimates are undersmoothed, the bias terms in the limit distributions of $\widetilde{G}(z)$ and $\overline{G}(z)$ are also the same.

We now turn to the estimation of H_0 or rather the full regression function $R(v,z,w) = H_0(vG_0(z),w)$. When $d_W + 1 > d_Z$, the rate of convergence of $\widehat{H}(\widehat{c})$ and $\widetilde{H}(\widetilde{c})$ is optimal because it is the same as the rate of convergence of the infeasible regression estimator based on knowing $G_0(\cdot)$. Furthermore,

$\widetilde{H}(\widetilde{c})$ has the oracle property that it achieves the same asymptotic variance as the infeasible regression estimator. When $d_W + 1 \leq d_Z$, the benchmark rate of convergence is provided by an estimator that makes use of known H_0, for example $H_0(\widehat{c})$, say, and our estimator indeed has the optimal rate in this case.

In the presence of heteroskedasticity, i.e., $\sigma^2(v, z, w) \neq \sigma^2$ for some v, z, w, one could alter the criterion in (16) from least squares to weighted least squares; the resulting estimator will involve an additional weighting factor due to the heteroskedasticity. Although weighting for heteroskedasticity can improve efficiency, it may result in more cumbersome procedures. Even in the homoskedastic case, the sequential $\widehat{G}$ and $\widehat{H}$ are simpler to compute and entail fewer assumptions (e.g., they don't require trimming) and so could be considered more robust in this regard.

6 Extensions and Conclusions

We proposed estimators of equation (1), simplifying without loss of generality to $R = H[vG(z), w]$ for unknown H and G. Important special cases include homothetically separable models which fit immediately in this form, and $r(v, z, w) = h[v + g(z), w]$. Instead of transforming the latter case into H and G, one could directly estimate h and g using our methodology by matching additively instead of multiplicatively, finding u such that $r(v, z, w) = r(v + u, z', w)$, making $u = u(z, z') = g(z) - g(z')$.

Further results appear in the working paper version of this article, including detailed proofs, consistent standard errors, a Monte Carlo simulation that shows small sample results in general accord with our asymptotic theory, and an empirical application to estimation of a value added homothetic production function for industrial firms in mainland China.

7 Appendix

We sketch the arguments used to establish Theorem 2. More detail and formal proofs are in the working paper version. The first step is to establish the consistency of $\widehat{U}(z, z')$ uniformly over z'. It suffices to show that

$$\sup_{z,z' \in \Psi_z} \sup_{U \in \widehat{\mathcal{U}}} |\widehat{m}_1(U; z, z') - m_1(U; z, z')| \xrightarrow{P} 0, \tag{27}$$

where $\widehat{\mathcal{U}}$ is the set for which $\widehat{m}_1(U; z, z')$ is well-defined. We show that this set $\widehat{\mathcal{U}}$ converges to $\mathcal{U}$, which is non-empty by assumption A. The result (27) follows from the expansion in assumption B using the triangle inequality and a ULLN of Andrews (1987). The identification argument in Theorem 1 implies that $\widehat{U}(z, z')$ is consistent, and indeed uniformly consistent. We next obtain a uniform asymptotic expansion for $\widehat{U}(z, z')$ that implies pointwise asymptotic normality, but also establishes

uniform rates on the remainder term. This is done using a general theory we develop based on a modification of the framework of Pakes and Pollard (1989). First, one extends the result in (27) to allow for a rate $\delta_n = b_G^{-1}$. Under our conditions, the population moment function $m_1(U; z, z')$ is several times continuously differentiable in U with non-zero first derivative. These conditions ensure that the uniform rate for $\widehat{m}_1(U; z, z') - m_1(U; z, z')$ translates into the same uniform rate for $\widehat{U}(z, z') - U_0(z, z')$. Second, we establish a stochastic equicontinuity condition that holds for U close to $U_0(z, z')$. For every sequence of positive numbers $\{\epsilon_n\}$ that converges to zero

$$\sup_{z,z' \in \Psi_Z} \sup_{\delta_n |U - U_0(z,z')| \leq \epsilon_n} \delta_{nG} \left| \widehat{m}_1(U; z, z') - m_1(U; z, z') - \widehat{m}_1(U_0(z, z'); z, z') \right| \xrightarrow{P} 0. \tag{28}$$

This is established by taking derivatives of the leading term of (24) rather than $\widehat{R}$ itself. Third, one obtains an expansion for $\widehat{m}_1(U_0(z, z'); z, z')$,

$$\begin{aligned} & \widehat{m}_1(U_0(z, z'); z, z') - m_1(U_0(z, z'); z, z') \\ = \; & \kappa \frac{1}{n} \sum_{i=1}^{n} u_i \frac{\pi(V_i, W_i)}{f_X(V_i, z, W_i)} \sum_{j=1}^{J} \frac{1}{b^{d_Z}} K_j^z \left(\frac{z - Z_i}{b} \right) + \\ & -\kappa \frac{1}{n} \sum_{i=1}^{n} u_i \frac{\pi(V_i / U_0(z, z'), W_i)}{U_0(z, z') f_X(V_i, z', W_i)} \sum_{j=1}^{J} \frac{1}{b^{d_Z}} K_j^z \left(\frac{z' - Z_i}{b} \right) + \\ & + b_{G'}^p \beta_U(z, z') + o_p(\delta_{nG}^{-1}), \end{aligned} \tag{29}$$

where $\beta_U(z, z')$ is a bounded continuous function and κ is a kernel related constant. The remainder terms are uniformly small by some arguments based on U-statistics and an exponential inequality. The non-zero first derivative on m_1 ensures that the expansion for $\widehat{m}_1(U_0(z, z'); z, z') - m_1(U_0(z, z'); z, z')$ translates into an expansion for $\widehat{U}(z, z')$ to $U_0(z, z')$ after dividing through by the 'Hessian', $\partial m_1(U_0(z, z'); z, z')/\partial U$.

To obtain the asymptotics for $\widehat{G}(z)$ we use the expansion for $\widehat{U}(z, z')$ to $U_0(z, z')$ and the standard approach to dealing with the integration type of estimators, Linton and Nielsen (1995), which here is based on U-statistic arguments. In this case, the second stochastic term in (29) 'integrates out', i.e., is of smaller order after integration over z'. The bias of $\widehat{G}(z)$ is the integrated bias of $\widehat{U}(z, z')$.

To obtain the asymptotic distribution of $\widehat{H}(\widehat{c})$, we use the uniform asymptotic expansion for $\widehat{G}(\cdot)$. Write $\widehat{H}(\widehat{c}) - H_0(c) = \widehat{H}(\widehat{c}) - H(\widehat{c}) + H(\widehat{c}) - H_0(c)$. We apply the delta method to obtain $H(\widehat{c}) - H_0(c) \simeq [\partial H_0(\gamma, w)/\partial \gamma](\widehat{\gamma} - \gamma) = [\partial H_0(\gamma, w)/\partial \gamma] v [\widehat{G}(z) - G(z)]$, and the asymptotic distribution of this term follows from the expansion of $\widehat{G}(z) - G(z)$. The tedious part is to obtain the distribution of $\widehat{H}(\widehat{c}) - H_0(\widehat{c})$, since this involves the generated regressors. This analysis involves a Taylor expansion of the kernel out to second order and then application of U-statistic techniques.

The asymptotics for $\widetilde{G}(z)$ and $\widetilde{H}(\widetilde{c})$ involve similar arguments. The difficult part here is that to obtain the full efficiency it is necessary to obtain expansions for $\widehat{G}(\cdot), \widehat{H}(\cdot)$, and $\partial \widehat{H}(\cdot)/\partial \gamma$ over a

set that expands to cover the whole support. Related recent work of Horowitz and Mammen (2005) proposes using series estimates of these preliminary quantities.

8 References

Andrews, D. W. K., (1987), "Consistency in Nonlinear Econometric Models: A Generic Uniform Law of Large Numbers," *Econometrica*, 55, 1465-1471.

Blackorby, C., D. Primont and R. R. Russell, (1978), *Duality, Separability, and Functional Structure: Theory and Economic Applications*. New York: North Holland.

Fan, J., and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications* Chapman and Hall.

Goldman, S. M. and H. Uzawa, (1964), "A Note On Separability and Demand Analysis," *Econometrica,* 32, 387-398.

Gorman, W. M., (1959), "Separable Utility and Aggregation," *Econometrica,* 27, 469-481.

Gozalo, P., and O. Linton (2000): "Local nonlinear least squares estimation: Using parametric information nonparametrically," *The Journal of Econometrics* 99, 63-106.

Härdle, W., W. Kim and G. Tripathi, (2001): "Nonparametric Estimation of Additive Models With Homogeneous Components," *Economics Essays: A Festschrift for Werner Hildenbrand*, eds. G. Debreu, W. Neuefeind, and W. Trockel, 159-179, Berlin: Springer.

Hanoch, G. and M. Rothschild (1972), "Testing the Assumptions of Production Theory: A Nonparametric Approach," *Journal of Political Economy,* 80, 256-275.

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

Horowitz, J., (2001), "Nonparametric Estimation of a Generalized Additive Model With An Unknown Link Function," *Econometrica,* 69, 499-513.

Horowitz, J., and Mammen, E. (2005), "Rate-Optimal estimation for a general class of nonparametric regression models with unknown link functions," Manuscript, University of Mannheim.

Lewbel, A. (1991), "The Rank of Demand Systems: Theory and Nonparametric Estimation," *Econometrica*, 59, 711-730.

Lewbel, A. (1997), "Consumer Demand Systems and Household Equivalence Scales," Handbook of Applied Econometrics, Volume II: Microeconomics, M. H. Pesaran and P. Schmidt, eds., Oxford: Blackwell Publishers Ltd.

Lewbel, A., and O. Linton (2005). "Nonparametric Matching and Efficient Estimators of Homothetically Separable Functions" Working Paper available at
http://personal.lse.ac.uk/lintono/working_papers.htm

LINTON, O. B. (1996), "Efficient estimation of additive nonparametric regression models," *Biometrika* 84, 469-474.

LINTON, O. AND E. MAMMEN (2005): "Estimating semiparametric ARCH(∞) models by kernel smoothing," *Econometrica* 73, 771-836.

LINTON, O. AND J. P. NIELSEN (1995), "A kernel method of estimating structured nonparametric regression based on marginal integration," *Biometrika*, 82, 93-100.

MAMMEN, E., O. LINTON, AND NIELSEN, J. P. (1999): "The existence and asymptotic properties of a backfitting projection algorithm under weak conditions," *Annals of Statistics* 27, 1443-1490.

MASRY, E. (1996), "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," *J. Time Ser. Anal.* 17, 571-599.

MATZKIN, R. L. (1992), "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica*, 60, 239-70

MATZKIN, R. L. (1994), "Restrictions of Economic Theory in Nonparametric Methods," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, 2523-2558, Amsterdam: Elsevier.

MATZKIN, R. L. (2003), "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339-1375.

NEWEY, W. K., AND R. L. MATZKIN (1993), "Kernel Estimation of Nonparametric Limited Dependent Variable Models," unpublished manuscript.

PAKES, A., AND D. POLLARD (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica* 57, 1027-1057.

PINKSE, J., (2001), "Nonparametric Regression Estimation Using Weak Separability," unpublished manuscript.

POWELL, J. L., (1994), "Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, 2443-2521, Amsterdam: Elsevier.

PRIMONT, D. AND D. PRIMONT, (1994), "Homothetic Non-parametric Production Models," *Economics Letters,* 45, 191-195.

SAGAN, H. (1969). *Introduction to the Calculus of Variations.* Dover Inc., New York.

STONE, C.J. (1980), "Optimal rates of convergence for nonparametric estimators," *Annals of Statistics,* 8, 1348-1360.

WEINSTOCK, R. (1952): *Calculus of Variations with applications to physics and engineering.* Dover Publications inc, New York.

ZELLNER, A. AND H. RYU (1998), "Alternative Functional Forms For Production, Cost and Returns to Scale," *Journal of Applied Econometrics*, 13, 101-127.