View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Research Papers in Economics

cemmap
centre for microdata methods and practice

NONPARAMETRIC ESTIMATION OF HOMOTHETIC AND HOMOTHETICALLY SEPARABLE FUNCTIONS

Arthur Lewbel
Oliver Linton

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP14/03

Nonparametric Estimation of Homothetic and Homothetically Separable Functions*

Arthur Lewbel† Oliver Linton‡
Boston College London School of Economics

October, 2003

Abstract

For vectors x and w, let $r(x, w)$ be a function that can be nonparametrically estimated consistently and asymptotically normally. We provide consistent, asymptotically normal estimators for the functions g and h, where $r(x, w) = h[g(x), w]$, g is linearly homogeneous and h is monotonic in g. This framework encompasses homothetic and homothetically separable functions. 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. Extensions to related functional forms include a generalized partly linear model with unknown link function. We provide simulation evidence on the small sample performance of our estimator, and we apply our method to a Chinese production dataset.

JEL Codes: C14, C21, D24.

Keywords: Cost Function; Economies of Scale; Homogeneous Function; Homothetic Function; Index Models; Nonparametric; 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, Gary Jefferson for providing data, and Shakeeb Khan, Rosa Matzkin, and Whitney Newey for helpful comments. All errors are our own.

†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. E-mail address: lintono@lse.ac.uk.

1 Introduction

Let X_i and W_i be observed vectors for $i = 1, \ldots, n$. Let $r(x, w)$ be some function that can be nonparametrically estimated, for example, $r(x, w)$ could equal $E(Y \mid X = x, W = w)$ which is estimated with observations $\{Y_i, X_i, W_i\}$. More generally, $r(x, w)$ could be a density, distribution, quantile, or hazard function, or $r(x, w)$ could be a utility or cost function derived from a set of estimated product or factor demands. Let $\widehat{r}(x, w)$ be a consistent, asymptotically normal estimator of the function $r(x, w)$.

Assume there exist functions h and g such that

$$r(x, w) = h[g(x), w] \tag{1}$$

where g is linearly homogeneous and h is strictly monotonic on its first element[1]. This paper provides consistent, asymptotically normal estimators of the functions h and g. This in turn yields a general estimator for homothetic and homothetically separable functions. We provide limiting distributions for our estimators, and provide Monte Carlo simulations of the small sample properties of our estimator. We also provide an empirical application to estimation of homothetic production functions for chemical plants in mainland China.

We also consider some extensions, including restrictions on g other than homogeneity, and the presence of endogenous regressors. In particular, we also provide an estimator for unknown functions H and m in the model $R(u, z, w) = H[m(z) + u, w]$, which is a generalized partly linear model with unknown link function (here we observe vectors Z_i and W_i and scalar U_i for $i = 1, \ldots, n$, and a nonparametric estimate of R).

A function $r(x)$ is defined to be homothetic if and only if $r(x) = h[g(x)]$ where h is strictly monotonic and g is linearly homogeneous. When w is empty, equation (1) is homothetic. More generally, a function r is defined to be homothetically separable, and x is said to be homothetically separable in r, if equation (1) holds where g is linearly homogeneous.

Homothetic and homothetically separable functions are commonly used in models of consumer preferences and firm production. The function $r(x, w)$ 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. Some examples of homothetic functions used in economics are provided in Chiang (1984). Zellner and Ryu (1998) perform empirical comparisons of a large number of different homothetic functional forms for production. Blackorby, Primont, and

[1] Without loss of generality, this model is equivalent to having g be homogeneous of any nonzero degree κ, because any homogeneous of degree κ function can be written as $g(x)^{\kappa}$ where $g(x)$ is linearly homogeneous, and the exponent κ can be absorbed into the function h.

Russell (1978) provide an extensive study of the properties of homothetically separable functions and their applications. See also Matzkin (1994) for a general survey on imposing restrictions of economic theory on nonparametric estimators.

Linear index models, corresponding to the case where $g(x) = x^\top \beta$, are a very common semi-parametric specification that arises in a variety of contexts, particularly limited dependent variable models. See, e.g., Powell (1994) for a survey. Replacing a linear index $x^\top \beta$ with an arbitrary linearly homogeneous function $g(x)$ 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.

In applications of homothetic separability, r may have multiple homogeneous components, that is,

$$r(x_0, x_1, \ldots, x_k) = h[g_1(x_1), \ldots, g_K(x_K), x_0] \tag{2}$$

for vectors $x_0, x_1, \ldots, x_K$. In this model, each g_k function can be estimated separately by applying the method we propose to estimate g in equation (1), 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 function, the function h may be estimated by nonparametrically regressing r on $g_1, \ldots, g_K, x_0$.

In many applications the functions h and g are of direct interest, e.g., the returns to scale of a homothetic production function is defined as the log derivative of h with respect to g. Even when h and g are not of direct interest, our estimator may still be valuable for testing whether functions are homothetic or homogeneously separable, by comparing $\widehat{r}(x, w)$ to $\widehat{h}[\widehat{g}(x), w]$, and because, with our estimator, the latter model achieves a faster rate of convergence than unrestricted nonparametric estimation of r.

One obvious way to estimate equation (1) would be to parameterize the unknown functions h and g, and employ nonlinear least squares to estimate these parameters. We propose an estimation method that employs kernel or local polynomial methods to estimate h and g nonparametrically, and establish their limiting distributions. The estimator we propose uses a form of marginal integration to deal with w, while exploiting the restriction that linearly homogeneous functions g satisfy the constraint $g(x) = g(x/v)v$ for any scalar v. Once an estimate of g is obtained, the function h is estimated by nonparametrically regressing r on g and w.

Matzkin (1992) provides a consistent estimator for the binary threshold crossing model $y = I[g(x) + \varepsilon \geq 0]$ where $g(x)$ is homogeneous and ε is independent of x. This threshold crossing model has $E(y|x) = h[g(x)]$ where h is the distribution function of $-\varepsilon$, and so is equivalent to our framework with $r(x) = E(y|x)$ and w empty.[2] In an unpublished manuscript, Newey and Matzkin (1993) propose (without derivation) a limiting distribution for an estimator of Matzkin's (1992)

[2] A motivating example Matzkin provides for the threshold crossing model is where $-g(x)$ is a constant returns to

model. Our estimator is similar to theirs, though it entails fewer steps, allows for the presence of w, and our limiting distribution for h exploits dimension reduction arising from homogeneity. Matzkin (2003) considers models of the form $y = m(x, \varepsilon)$ with ε independent of x and, as one possible identifying assumption, m being homogeneous in x and an unobserved ε. In contrast, our model makes no assumptions about (and provides no estimates of) the role of unobservables other than limiting distribution theory for r, and allows for homothetic rather than just homogeneous dependence on x.

Models satisfying equations (1) or (2) without imposing homogeneity on g or g_k 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, without assuming homogeneity of g. In homothetically separable models, Pinkse's estimator will identify g up to an arbitrary monotonic transformation, whereas our estimator provides the unique (up to scale) linear homogeneous g, and exploits the homogeneity of g to obtain a faster rate of convergence than Pinkse's.

Strongly or additively separable models are models of the form $r(x) = \sum_k g_k(x_k)$. Härdle, Kim, and Tripathi (2001) provide a nonparametric estimator of additively separable models where the $g_k(x_k)$ functions are homogeneous. Other estimators of additively separable models include Friedman and Stutzle (1981), Breiman and Friedman (1985), Andrews (1991), Tjøstheim and Auestad (1994), and Linton and Nielsen (1995). Stone (1986), Hastie and Tibshirani (1990), Linton and Härdle (1996), and Horowitz and Mammen (2002) provide estimators of generalized additively separable models, defined as $r(x) = H[\sum_k g_k(x_k)]$, where H is a known function, and Horowitz (2001) extends this to the case where H is unknown. In both cases, homogeneity is not assumed or exploited. Tripathi and Kim (2001) discuss nonparametric estimators of homogeneous functions, corresponding to the special case of equation (1) where w is empty and h is a known power function.

In section 2 we discuss identification and the estimation strategy in general terms, while in section 3 the estimation algorithm is given in full detail. In section 4 we present the distribution theory for our estimators of g and of h. In section 5 we present the results of some Monte Carlo simulations and an application to Chinese production data. Section 6 discusses some extensions and conclusion. The proofs are given in the appendix.

scale cost function for a project, ε is firm's benefit or return from undertaking the project (which is unknown to the researcher), and y indicates whether the firm embarks on the project, which it does if the benefit exceeds the cost.

2 Identification

ASSUMPTION A1. Let X and W be random vectors with $\text{supp}(X, W) = \Psi_X \times \Psi_W$. There exists functions r, h and g such that $r(x, w) = h[g(x), w]$ for all $(x, w) \in \Psi_X \times \Psi_W$, where h is invertible with respect to its first element and g is linearly homogeneous. Let $x_0 \in \Psi_X$ be a constant vector such that $g(x_0) \neq 0$ and, for all $x \in \Psi_X$, there exists a scalar v_x such that $g(x) = g(v_x x_0)$. Without loss of generality, assume that $g(x_0) = 1$.

Define $\Psi_V = \{\widetilde{v} \mid \widetilde{v}x_0 \in \Psi_X,\ \widetilde{v} \neq 0\}$. For any scalar $v \neq 0$, given the function $r(x, w)$, let the function $s(r, q, w)$ be defined for any q such that $vq \in \Psi_X$ by

$$s[r(qv, w), q, w] = v \tag{3}$$

so s is the inverse of the function $r(qv, w)$ with respect to v. Invertibility of h ensures that s exists.

THEOREM 1. *Let Assumption A1 hold. Let $(V, \widetilde{W})$ be any random vector with support contained in $\Psi_V \times \Psi_W$. Then for every $x \in \Psi_X$ and $w \in \Psi_W$ the functions g and h satisfy*

$$g(x) = E\left(\frac{V}{s\left[r\left(Vx_0, \widetilde{W}\right), x, \widetilde{W}\right]}\right) \tag{4}$$

$$h[g(x), w] = E[r(X, W) \mid g(X) = g(x), W = w]. \tag{5}$$

PROOF. Let w be any element of Ψ_W. For any v, q such that $v \neq 0$ and $vq \in \Psi_X$, having $g(x)$ be linearly homogeneous implies that $g(vq) = g(q)v$, and so $r(vq, w) = h[g(q)v, w]$, which in turn implies that

$$s(r, q, w) = \frac{h^{-1}(r, w)}{g(q)}, \tag{6}$$

where h^{-1} is the inverse function of h on its first element.

Given x, the function $r = r\left(\widetilde{v}x_0, w\right)$ is well defined because $\widetilde{v} \in \Psi_V$. Also, the function $s(r, x, w)$ is well defined for $r = r\left(\widetilde{v}x_0, w\right)$ as long as there exists a $v = v(\widetilde{v}, x)$ such that $r = r\left(\widetilde{v}x_0, w\right) = r\left(vx, w\right)$, since then $s(r, x, w)$ will equal this v. The equality $r\left(\widetilde{v}x_0, w\right) = r\left(vx, w\right)$ for all w is equivalent to $g[(\widetilde{v}/v)x_0] = g(x)$, which holds taking $\widetilde{v}/v(\widetilde{v}, x) = g(x)$. It follows that the function $s\left[r\left(\widetilde{v}x_0, w\right), x, w\right]$ is well defined.

Given equation (6), we have

$$\begin{aligned} \frac{\widetilde{v}}{s\left[r\left(\widetilde{v}x_0, w\right), x, w\right]} &= \frac{g(x)\widetilde{v}}{h^{-1}\left[r\left(\widetilde{v}x_0, w\right), w\right]} \\ &= \frac{g(x)\widetilde{v}}{h^{-1}\left[h[g(x_0)\widetilde{v}, w], w\right]} = g(x) \end{aligned} \tag{7}$$

This equation holds for all $w \in \Psi_W$, $\widetilde{v} \in \Psi_V$ $\widetilde{v} \neq 0$, so it holds in expectation replacing w with $\widetilde{W}$ and $\widetilde{v}$ with V, thereby yielding equation (4), since $V \neq 0$ with probability one. Equation (5) then follows from the definition of r. ■

Theorem 1 shows that the functions g and h are identified given r. With a consistent estimator $\widehat{r}$ of the function r, a consistent estimator $\widehat{s}$ may be constructed by inverting $\widehat{r}(qv, w)$ with respect to v. A consistent estimator of $g(x)$ is then given by

$$\widehat{g}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{V_i}{\widehat{s}\left[\widehat{r}\left(V_i x_0, W_i\right), x, W_i\right]} \tag{8}$$

and a consistent estimator of h is then given by a nonparametric regression of $\widehat{r}(x, w)$ on $\widehat{g}(x), w$.[3] Monotonic regression or average contour estimation could be used for this last step. We flesh out the details of this in the next section.

Equations (4) and (8) hold just letting $V = 1$ and $\Psi_V = \{1\}$. The reason for averaging over a random V is that this reduces the effective dimensionality of the estimator, resulting in a faster rate of convergence.[4] For example, one could let V be the inverse of a continuously distributed element of X. In this case a good choice for x_0 might be the vector that has one in the same position that V occupies in X and zeros elsewhere. Another sensible choice for V would be $V = 1/||X||$. In particular, if each element of X is continuously distributed with the same support, then in this case one might take x_0 to be the vector of ones. Yet another possibility is to let $V = X^\top a$ for some nonzero constant vector a, which might be chosen to maximize efficiency of the resulting estimator. In fact, apart from convenience it is not necessary to relate the distribution of V to X at all, except in so far as the support of V has to be determined by the support of X.

Theorem 1 assumes a set Ψ_V of values that V can take on, and hence be averaged over, for any value of X, and we need to be able to estimate the function s for every V in this set. To facilitate this construction, Theorem 1 can be generalized slightly by replacing equation (3) with

$$s[r(xv/v_0(x), w), x, w] = v/v_0(x). \tag{9}$$

for any nonzero function $v_0(x)$. Equation (9) follows immediately from equation (3) by replacing q with x and v with $v/v_0(x)$. We may now choose any convenient $v_0(x)$ (we will use $v_0(x) = ||x||$), define Ψ_V in terms of v in Equation (9) instead of the original v, and estimate s accordingly (see the section on estimation for details).

[3] This estimator of $g(x)$ takes $\widetilde{W} = W$ for convenience. If the support of X, W were not rectangular, one could modify equation (8) by including the indicator $1(W_i \in \text{supp}(W|X = x))/\sum_{j=1}^{n} 1(W_j \in \text{supp}(W|X = x))$, equivalent to choosing a $\widetilde{W}$ with support that is a subset of Ψ_W. Taking $\widetilde{W}$ to be a trimmed W could also be used to avoid boundary issues in the estimation of $\widehat{r}$.

[4] For the same reason, we take $\widetilde{W} = W$ and average over it, instead of fixing $\widetilde{W}$.

Given any choice of x_0, if the assumption that $g(x_0) \neq 0$ (and hence normalizable to one) is violated then $s(r, x_0, w)$ will be infinite for all r and w. It may therefore be possible to test this assumption by, e.g., testing if $\sum_{i=1}^{n}[n\widehat{s}[\widehat{r}(X_i, W_i), x_0, W_i]^{-1}$ is significantly different from zero.

The identification in Theorem 1 does not require differentiability of h or g (though estimation may impose such smoothness) and allows some or all of the elements of X and W other than V to be discrete or otherwise not continuously distributed.

If $r(x, w) = E(Y \mid X = x, W = w)$ for some random Y, then $h[g(x), w] = E[Y \mid g(X) = g(x), W = w]$, which on estimation may yield a simpler limiting distribution than one based on equation (5).

2.1 Matching

One way to interpret Theorem 1 is in terms of a matching estimator. For a given x, w, find $\widetilde{v}$ such that $r(\widetilde{v}x_0, w) = r(x, w)$, a match. Then $g(\widetilde{v}x_0) = g(x)$, i.e., $g(x_0)\widetilde{v} = g(x)$, so $g(x) = \widetilde{v}$. The same argument implies that $g(x) = \widetilde{v}/v$ when $v, \widetilde{v}$ satisfy $r(\widetilde{v}x_0, w) = r(vx, w)$. Our estimator essentially does this matching (replacing r with $\widehat{r}$) for a range of values of x and w, and averages over the results. Note that the set $\{v : vx \in \Psi_X\}$ varies with x, whereas we want to choose a set Ψ_V of values for V to average over that does not vary with the evaluation point x. If we replace x by $q_0(x) = x/v_0(x)$, where $v_0(x) = ||x||$, we have $r(vq_0(x), w) = r(vx/v_0(x), w) = r(vx, w)/v_0(x)$ by homogeneity. Therefore, we look for a match with $r(\widetilde{v}x_0, w) = r(vq_0(x), w)$ and then divide by $v_0(x)$. This is the matching interpretation of replacing equation (3) with equation (9). Figure 1 illustrates this construction.

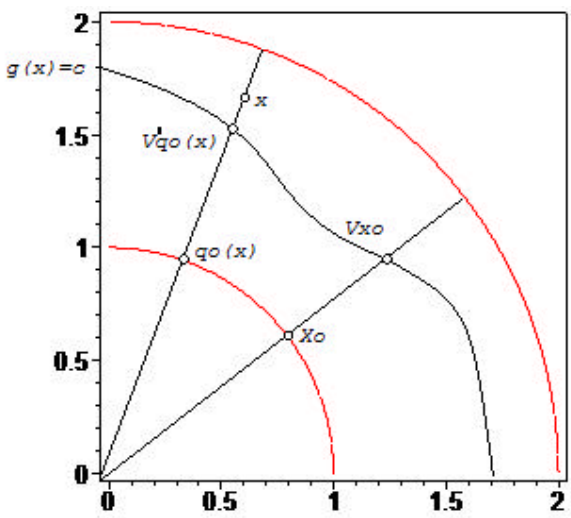


Figure 1. The points $v'q_0(x)$ and vx_0 are matched, i.e., $g(v'q_0(x)) = g(vx_0)$, which implies that $g(x) = v/v_0(x)v'$.

2.2 An Alternative Approach

Here we propose a variant of our approach based on first integrating out the contribution from w, and then matching. Define

$$\overline{r}(x) = \overline{h}[g(x)] = \int r(x,w) f_{\widetilde{W}}(w) dw \tag{10}$$

for any density $f_{\widetilde{W}}(w)$ on Ψ_W. Define $\overline{s}(r,q)$ be the inverse of the function $\overline{r}(qv)$ with respect to v. Invertibility of $\overline{h}$ ensures that $\overline{s}$ exists.

COROLLARY 1. *Let Assumption A1 hold, except that h need not be invertible with respect to its first element. Assume $\overline{r}(x)$ exists for every $x \in \Psi_X$ and that $\overline{h}$ is invertible. Let V be any random scalar with support contained in Ψ_V. Then for every $x \in \Psi_X$ the function g satisfies*

$$g(x) = E\left(\frac{V}{\overline{s}\left[\overline{r}\left(Vx_0\right), x\right]}\right) \tag{11}$$

The proof is omitted, since it follows exactly the same steps as the proof of Theorem 1. Corollary 1 suggests a slightly different estimator for $g(x)$, as follows. Let $\widetilde{r}(x)$ be some marginal integration estimator, e.g.,

$$\widetilde{r}(x) = \frac{1}{n}\sum_{i=1}^{n} \widehat{r}(x, \widetilde{W}_i),$$

where $\widetilde{W}_i$ is drawn from $f_{\widetilde{W}}(w)$. Define $\widetilde{s}(r,q)$ as the inverse of the function $\widetilde{r}(qv)$ with respect to v, and define

$$\widetilde{g}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{V_i}{\widetilde{s}\left[\widetilde{r}\left(V_i x_0\right), x\right]}. \tag{12}$$

The original estimator $\widehat{g}(x)$ inverts r and then averages over W and V. The alternative $\widetilde{g}(x)$ first averages over $\widetilde{W}$ then inverts and averages over V. This alternative requires existence of $\overline{r}(x)$, but only requires invertibility of $\overline{h}$ instead of h. In the Appendix we show that $\widetilde{g}(x)$ converges at the same rate as $\widehat{g}(x)$, but has a different limiting variance. One of the main differences is a dependence on the product of the marginal densities of V and W in place of the joint density. Neither estimator uniformly dominates the other.

2.3 Short Examples

1. Suppose that $g(x) = ||x||$, $\Psi_X \subset \mathbb{R}^2$ and contains a circle of some positive radius, $h(\gamma) = \exp(\gamma)$, and there is no w. Then $r(q \cdot v) = \exp(v \cdot ||q||)$ and $s(r,q) = \log(r)/||q||$. Let x_0 be any point on the unit circle. The set Ψ_V is the ray $\{\widetilde{v} \mid \widetilde{v}x_0 \in \Psi_X\}$. We then have $V/s\left[r\left(Vx_0\right), x\right] = V/s\left[\exp V, x\right] = ||x||$ exactly. Our Monte Carlo analysis takes this form.

2. Suppose that $g(x) = a^\top x$, where $x, a \in \mathbb{R}^d_+$, $h(\gamma) = \exp(\gamma + w)$. Then $s(r, q) = [\log(r) - w]/a^\top q$. Given any point x_0, we may freely scale a such that $a^\top x_0 = 1$. For all $x \in \Psi_X$, there must exist a scalar v_x such that $a^\top x = v_x a^\top x_0 = v_x$.

3. Suppose that $g(x) = (\sum_{j=1}^d a_j x_j^\theta)^{1/\theta}$, where $x, a \in \mathbb{R}^d_+$, $h(\gamma) = \gamma^2$, and that there is no w. Then $s(r, q) = \sqrt{r} / \left(\sum_{j=1}^d a_j x_j^\theta\right)^{1/\theta}$.

3 Estimation Details

We suppose that we observe independent and identically distributed observations $\{Z_i\}_{i=1}^n$, where $Z_i = (X_i, W_i) \in \mathbb{R}^d$ and $d = d_x + d_w$, and that we can compute an estimator $\widehat{r}(z)$ for all $z = (x, w)$ in the support of Z_i. We shall give more details about the properties that the estimator $\widehat{r}(z)$ possesses later but for now we just need that it is well-defined for all z.

There are a variety of ways of implementing our estimator. We choose a way that is convenient in practice. The main difficulty is in finding the function s for all relevant r, x, w. Let $v_0(x) = ||x||$ and $q_0(x) = x/v_0(x)$. Then, based on equation (9) and on the matching interpretation of our estimator, we define our estimator of $\widehat{s}(r, x, w)$ as any sequence that satisfies [or approximately satisfies to some order, see below]

$$\widehat{s}(r, x, w) = \frac{1}{v_0(x)} \arg\min_{v \in \Psi_V} \{v : \widehat{r}(q_0(x) \cdot v, w) = r\}. \tag{13}$$

We then define $\widehat{g}(x)$ by equation (8), where V_i are i.i.d. draws from V. This approach allows us to use the same support for V for all x.

To estimate $h(\gamma, w)$ for $\gamma \in \Psi_\Gamma = \{g(x) : x \in \Psi_X\}$, we can compute the generated regression implied by equation (5), that is compute the smooth of $\widehat{r}(X_i, W_i)$ on $\widehat{g}(X_i), W_i$. In the special case that $r(z) = E(Y \mid Z = z)$ and where there is also sample information on Y, we can alternatively estimate $h(g(x), w)$ by the smooth of Y_i on $\widehat{g}(X_i), W_i$. We will base our estimation of conditional expectations on the local polynomial kernel method. This method has been extensively analyzed and has some attractive properties like being design adaptive, and best linear minimax; see Fan and Gijbels (1996) for further discussion.[5] For any dataset $\{\widehat{Y}_i, \widehat{Z}_i\}_{i=1}^n$, the local polynomial regression of $\widehat{Y}_i$ on $\widehat{Z}_i \in \mathbb{R}^d$ of order q can be obtained from the multivariate weighted least squares criterion:

$$\sum_{i=1}^n \left[\widehat{Y}_i - \sum_{0 \le |\mathbf{j}| \le q} \theta_{\mathbf{j}} \cdot (\widehat{Z}_i - z)^{\mathbf{j}}\right]^2 K\left(\frac{\widehat{Z}_i - z}{b_*}\right), \tag{14}$$

[5] For $\widehat{h}$, local linear regression does not impose monotonicity of h on its first element. Monotonic regression could have been used instead, though note that h need not be monotonic on its first element if equation (12) is used to estimate g.

where $K(u)$ is a nonnegative weight function on $\mathbb{R}^d$ and b_* is a bandwidth parameter. Let $\widehat{m}(x) = \widehat{\theta}_0$, where $\widehat{\theta}_0$ is the minimizing intercept in (14). For any conformable vectors $x, \mathbf{j}$, $x^{\mathbf{j}} = x_1^{\mathbf{j}_1} \times \ldots \times x_d^{\mathbf{j}_d}$, while $|\mathbf{j}| = \sum_{l=1}^d \mathbf{j}_l$.

The following algorithm is used to define the estimators $\widehat{g}(X_i)$, $i = 1, \ldots, n$ and $\widehat{h}(\gamma, w)$:

- First, compute for each X_i, $i = 1, \ldots, n$ the vector $q_0(X_i)$ and scalar $v_0(X_i)$ such that $||q_0(X_i)|| = ||x_0||$ and $q_0(X_i) \cdot v_0(X_i) = X_i$. Then compute $\widehat{r}(V_j \cdot x_0, \widetilde{W}_i)$, $i, j = 1, \ldots, n$. Here, $(V_i, \widetilde{W}_i)$ are drawn from some known density $f_{V,\widetilde{W}}$.

- Then find the values

$$\widehat{V}_{(i,j)} = \arg \min_{V_k : 1 \le k \le n} \left| \widehat{r}(V_k q_0(X_i), \widetilde{W}_i) - \widehat{r}(V_j x_0, \widetilde{W}_i) \right|, \quad i, j = 1, \ldots, n,$$

  and let $\widehat{s}_{ij} = \widehat{s}[\widehat{r}(V_j x_0), X_i, \widetilde{W}_i] = \widehat{V}_{(i,j)} / v_0(X_i)$.

- Then let

$$\widehat{g}(X_i) = \frac{1}{n} \sum_{j=1}^n \frac{V_j}{\widehat{s}_{ij}}. \tag{15}$$

- For any $c = (\gamma, w)$ with $\gamma \in \{g(x) : x \in \Psi_X\}$, let $\widehat{h}(\gamma, w)$ be the intercept from a local polynomial regression of order q of $\widehat{r}(Z_i)$ on $\widehat{C}_i = (\widehat{g}(X_i), W_i)$. The computation of $\widehat{r}(Z_i)$ in this smooth may be based on a different amount of smoothing than in the previous usage in computation of $\widehat{g}$. In the special case where $r(z) = E(Y \mid Z = z)$ we can replace $\widehat{r}(Z_i)$ in the smooth by Y_i.

To apply the integrate first strategy described in section 2.2, just drop the $\widetilde{W}_i$ argument in the above calculations and use the estimator $\widetilde{r}(X_i)$ in place of $\widehat{r}(Z_i)$.

The arbitrary sign and scaling of $g(x)$ is chosen by the normalization $g(x_0) = 1$. The above estimator $\widehat{g}(x)$ may have $\widehat{g}(x_0) \neq 1$. This suggests that alternative consistent estimators such as $\widetilde{g}(x) = \widehat{g}(x) + 1 - \widehat{g}(x_0)$ or $\widetilde{g}(x) = \widehat{g}(x)/\widehat{g}(x_0)$ can be constructed that might be more accurate, at least for x in the neighborhood of x_0.

4 Distribution Theory

We present the pointwise distribution of our estimators of $g(x)$ at some $x \neq x_0$ by the two methods described, which requires an analysis of the properties of $\widehat{s}$. The technical issues are similar in some respects to those in the estimation of the mode of a density [Romano (1988)] or to those in the

estimation of maximal points of a regression function [Müller (1989)]. We then give the distribution theory for our estimator of $h(\gamma, w)$, which is a sort of generated regressors problem, see Ahn (1995). We first present our regularity conditions that are used in establishing the asymptotic normality of our estimators. We shall suppose that our estimator $\widehat{r}(.)$ of $r(.)$ satisfies an asymptotic expansion but are not more specific than this. Since the target function $r(.)$ could be a variety of things depending on the application and since a variety of estimation strategies could be contemplated for $\widehat{r}(.)$, we would like our theory to allow for this.

We shall work in polar co-ordinates, at least for X. This presupposes that the estimation strategy for $\widehat{r}(.)$ has been conducted in polar co-ordinates for X, which we would argue is quite sensible for this particular problem. The main advantage of this approach is that we obtain simple formulae for the bias and variance of our estimator. For any $x \in \Psi_X$ there exists $\rho \in \mathbb{R}, \theta \in \mathbb{R}^{d_x-1}$ with $\rho(x) = ||x||$ for some norm, where we can write $x = \phi(\rho, \theta)$ for a smooth invertible function ϕ.[6] The function $r(x, w)$ can be rewritten in terms of ρ, θ as $r(x, w) = r(\phi(\rho, \theta), w) = r_*(\phi^{-1}(x), w)$ and likewise $g(x) = g(\phi(\rho, \theta)) = g_*(\rho, \theta) = \rho g_{**}(\theta)$, where $g_{**}(\theta) = g_*(1, \theta)$. Let $\theta_0 = \theta(x_0)$ and $\rho_0 = \rho(x_0)$. Note that if $\theta(x) = \theta(x_0)$, then $g(x) = \rho(x)$, so we only consider points x with $\theta(x) \neq \theta(x_0)$.

Define for any vector $\alpha = (\alpha_1, \ldots, \alpha_d)^\top$ and function $f\colon \mathbb{R}^d \to \mathbb{R}$

$$D^\alpha f(z) = \frac{\partial^{|\alpha|} f(z)}{\partial z_1^{\alpha_1} \cdots \partial z_d^{\alpha_d}} \text{ with } |\alpha| = \sum_{j=1}^{d} \alpha_j.$$

Our assumptions are given below. We suppose that the function ϕ is arbitrarily smooth and so our main assumptions can be stated equivalently in terms of Z or Z_*; we have taken a 'mixed approach' and state some assumptions on Z and some on Z_*.

Assumption B.

B1. The random variables $(X_i, W_i, \widetilde{W}_i, V_i)$ or equivalently $(\rho_i, \theta_i, W_i, \widetilde{W}_i, V_i)$, $i = 1, \ldots, n$ are independent and identically distributed. Let $z_* = (\rho, \theta, w) \in \mathbb{R}^d$ and let $f_{Z_*}(z_*)$ $[f_Z(z)]$ be the joint density function of $Z_{*i} = (\rho_i, \theta_i, W_i)$ $[Z_i = (X_i, W_i)]$ with support $\Psi_{Z_*} = \Psi_\rho \times \Psi_\theta \times \Psi_W$ $[\Psi_Z = \Psi_X \times \Psi_W]$ a compact subset of $\mathbb{R}^d$, and let $f_{V,\widetilde{W}}(v, w)$ be the joint density function of $V_i, \widetilde{W}_i$.

B2. (a) The functions h, g, and ϕ are p-times continuously partially differentiable in all directions, which implies that $D^\alpha r_*(z_*)$ exists and is continuous on Ψ_Z for all α with $|\alpha| \leq p$; (b) The function h satisfies $\inf_{\gamma \in \Psi_\Gamma, w \in \Psi_W} |\partial h(\gamma, w)/\partial \gamma| > 0$.

B3. We suppose that $\widehat{r}(z) = \widehat{r}_*(z_*)$, where $z_* = (\phi^{-1}(x), w)$, satisfies the asymptotic expansion

$$\widehat{r}_*(z_*) - r_*(z_*) = \frac{1}{nb^d} \sum_{i=1}^{n} a_n\left(z_*; \frac{z_* - Z_{*i}}{b}\right) K\left(\frac{z_* - Z_{*i}}{b}\right) u_i + b^p \beta(z_*) + R_n(z_*), \tag{16}$$

[6] For example, $x_1 = \rho \sin\theta_1$, $x_2 = \rho\cos\theta_1 \sin\theta_2, \ldots,$ $x_{d_x-1} = \rho\cos\theta_1 \cdots \cos\theta_{d_x-2} \sin\theta_{d_x-1}$, $x_{d_x} = \rho\cos\theta_1 \cdots \cos\theta_{d_x-2} \cos\theta_{d_x-1}$. We take the range $-\pi/2 < \theta_i \leq \pi/2$, $i = 1, \ldots, d_x - 2$, and $-\pi < \theta_{d_x-1} \leq \pi$.

where the components of this expansion have the following properties:

(a) The random variable u_i is i.i.d. and satisfies $E(u_i|Z_i) = 0$ and $E(|u_i|^{2+\epsilon}) < \infty$ for some $\epsilon > 0$. The function $\sigma^2(z_*) = \text{var}(u_i|Z_{*i} = z_*)$ is continuous;

(b) The random function $a_n(z_*; t)$ depends only on $Z_1, \ldots, Z_n$, is uniformly Lipschitz continuous in both its arguments, and satisfies $\sup_{z_* \in \Psi_Z, t \in \text{supp}(K)} |a_n(z_*; t) - a(z_*)| = o_p(1)$ for some function a. The non-random functions $a(.), \beta(.)$ are bounded and continuous on Ψ_Z;

(c) The kernel K takes the product form $K(u) = \prod_{j=1}^{d} k(u_j)$, where k is symmetric about zero, integrates to one, has compact support, and is continuously differentiable in all its arguments;

(d) The remainder term $R_n(\cdot)$ is a continuous function that satisfies

$$\sup_{z_* \in \Psi_{Z_*}} |R_n(z_*)| = o_p(\delta_n), \quad \text{where } \delta_n = n^{-p/(2p+d_x-1)}. \tag{17}$$

B4. For some p^* with $p > p^* > (d_w + 1)/2$, as $n \to \infty$,

$$\sup_{z \in \Psi_Z} \left| \frac{\partial \widehat{r}}{\partial x}(z) - \frac{\partial r}{\partial x}(z) \right| = o_p(n^{-1/4} b^{-(d_x-1)/4}) \tag{18}$$

$$\sup_{z \in \Psi_Z} |D^\alpha \widehat{r}(z) - D^\alpha r(z)| = o_p(1), \ |\alpha| \leq p^*. \tag{19}$$

B5. The bandwidth satisfies $b = cn^{-1/(2p+d_x-1)}$ for some c with $0 < c < \infty$.

Assumption B2(b) is similar to assumption 4 in Horowitz (2001). Assumption B3 is consistent with the estimator $\widehat{r}(z)$ being a $(p-1)^{th}$ order with local polynomial nonparametric regression estimator as in Fan and Gijbels (1996) [in which case $a(z)$ is proportional to $1/f_Z(z)$ and u_i is the regression error], or a local polynomial quantile estimator [in which case $a(z)$ is proportional to $1/f_Z(z) f_{u|Z}(0)$ and u_i is the check function of the regression error], or a nonlinear function of vectors of such estimators.[7] It is also consistent with the case where the input smoother is a marginal integration type estimator of the function $\overline{r}(x)$ defined in (10). In that case, $a(z_*) = f_W(w)/f_{Z*}(\rho, \theta, w)$, and we should elsewhere replace the argument z by x, and replace d by d_x in the dimensionality of K and in the power of b. There is one substantive difference from the usual local polynomial smoother case, which is that implicitly the smoothing window has been defined in the polar co-ordinates of X

[7] Suppose for simplicity that $\widehat{r}$ were a kernel regression estimator. In that case $a_n(z) = 1/\widehat{f}(z)$, where $\widehat{f}(z)$ is a kernel density estimator. Then taking $a(z) = 1/f(z)$ we have

$$\sup_z |a_n(z) - a(z)| \leq \frac{\sup_z \left| \widehat{f}(z) - f(z) \right|}{(\inf_z f(z)) \left(\inf_z f(z) - \sup_z |\widehat{f}(z) - f(z)| \right)} = o_p(1),$$

provided $\sup_z |\widehat{f}(z) - f(z)| = o_p(1)$. A similar argument can be made in the local polynomial case.

rather than the original co-ordinates.[8]

The rate $\delta_n = n^{-p/(2p+d_x-1)}$ in (17) is the rate at which $\widehat{g}(x)$ converges under our assumptions; in fact, this is the optimal pointwise rate of convergence for nonparametric functions of dimension $d_x - 1$ and smoothness p, Stone (1980).

Our proof technique will use some results of Andrews (1994) that uses properties of the higher derivatives, which explains why we have assumed in B4 that some higher order partials are uniformly consistent. For local polynomial regression estimators the rate of convergence of the jth-order partial derivatives is $O_p(\sqrt{\log n/nb^{d+2j}})$ [given B5], which can be achieved under some regularity conditions, see for example Masry (1996a, 1996b). The bandwidth sequence in B5 is smaller than would be optimal for the high-dimensional derivative estimation, and it is not always the case that Assumption B4 for example will be satisfied. To ensure that this does hold requires additional restrictions on d_w, p : (18) requires that $p > d_w + 3$, while (19) requires $p > (3d_w + 4)/2$. As we have discussed, to satisfy (16)-(19) we require restrictions linking d_w with p. The strongest such restriction is that $p > (3d_w+4)/2$. The greater the dimensionality of w, the more smoothness is required in order to fulfil these conditions. This technical problem shows up in semiparametric estimation, and, of more direct relevance, in the literature on estimation of additive regression functions by marginal integration, see Linton and Nielsen (1995) and Horowitz (2001). One issue here is that if the smoothness conditions are not satisfied, i.e., d_w is too large relative to p, our estimator is not 'rate optimal' in the sense of Stone (1986) [note however that our estimator is rate optimal when the restrictions are satisfied though]. Hengartner and Sperlich (2002) and Horowitz and Mammen (2002) discuss this issue and propose solutions for specific problems. We are only concerned with proposing sufficient conditions for our limit theory to hold; it may be possible to improve the theoretical results to obtain rate optimal estimators.

Define the bounded continuous functions of x

$$\begin{aligned} \beta_g(x) &= -g(x)\int[\beta(v\cdot\rho_0,\theta_0,w)-\beta(v\cdot\rho(x),\theta(x),w)]\frac{f_{V,\widetilde{W}}(v,w)}{\frac{\partial h}{\partial\gamma}(v,w)\times v}dvdw \qquad (20)\\ I(x) &= \int\sigma^2(\rho,\theta(x),w)a^2(\rho,\theta(x),w)\frac{f^2_{V,\widetilde{W}}(\rho g(x)/\rho(x),w)f_{Z_*}(\rho,\theta(x),w)d\rho dw}{\left[\frac{\partial h}{\partial\gamma}(\rho g(x)/\rho(x),w)\times\rho\right]^2}. \end{aligned}$$

THEOREM 2. *Suppose that Assumptions A1 and B1-B5 hold. Then,*

$$\sqrt{nb^{d_x-1}}\left(\widehat{g}(x)-g(x)-b_n^p\beta_g(x)\right)\Longrightarrow N\left(0,\quad(x)\right),$$

[8] This just amounts to another type of multivariate smoothing window: other common approaches include the rectangular window, and the elliptical window. The shape of the smoothing window will typically have a material affect on the bias constant but not on the variance.

where $(x) = {}_1(x) + {}_2(x)$ *and*

$$ {}_1(x) = ||k||_2^{2(d_x-1)} g^2(x) I(x_0) \quad ; \qquad {}_2(x) = ||k||_2^{2(d_x-1)} g^2(x) I(x). $$

To aid inference we next show how to compute analytic standard errors. Let $\widehat{\ }(x) = \widehat{\ }_1(x) + \widehat{\ }_2(x)$, where

$$ \begin{aligned} \widehat{\ }_1(x) &= \widehat{g}(x)^2 \frac{1}{nb^{d_x-1}} \sum_{i=1}^{n} \widehat{u}_i^2 \prod_{j=1}^{d_x-1} k^2\left(\frac{\theta_{j0}-\theta_{ji}}{b}\right) \widehat{a}^2(\rho_i, \theta_0, W_i) \frac{f^2_{V,\widetilde{W}}(\rho_i/\rho_0, W_i)}{\left[\frac{\partial \widehat{h}}{\partial \gamma}(\rho_i/\rho_0, w) \times \rho_i\right]^2} \\ \widehat{\ }_2(x) &= \widehat{g}(x)^2 \frac{1}{nb^{d_x-1}} \sum_{i=1}^{n} \widehat{u}_i^2 \prod_{j=1}^{d_x-1} k^2\left(\frac{\theta_{j0}-\theta_{ji}}{b}\right) \widehat{a}^2(\rho_i, \theta(x), W_i) \frac{f^2_{V,\widetilde{W}}(\rho_i \widehat{g}(x)/\rho(x), W_i)}{\left[\frac{\partial \widehat{h}}{\partial \gamma}(\rho_i \widehat{g}(x)/\rho(x), w) \times \rho_i\right]^2}, \end{aligned} $$

where hats denote (nonparametric) estimators of the corresponding population quantities. Under suitable regularity conditions $\widehat{\ }(x) \to^p \ (x)$. The estimates of $\partial \widehat{h}/\partial \gamma$ come from the slope estimates from (14).

We now discuss the asymptotic distribution in some special cases. In fact, we just look at the first terms in (x) as the same comments apply to the second ones. In the special case that $\widehat{r}(z)$ is a local linear estimator of a regression function r,

$$ I(x_0) = \int \frac{\sigma^2(\rho, \theta_0, w)}{f_{Z_*}(\rho, \theta_0, w)} \frac{f^2_{V,\widetilde{W}}(\rho/\rho_0, w) d\rho dw}{\left[\frac{\partial h}{\partial \gamma}(\rho/\rho_0, w) \times \rho\right]^2}. \tag{21} $$

In the special case that the input smoother is a local polynomial based marginal integration type estimator of the function $\overline{r}(x)$ defined in (10),

$$ I(x_0) = \int \frac{\sigma^2(\rho, \theta_0)}{f_{Z_*}(\rho, \theta_0, w)} \frac{f^2_V(\rho/\rho_0) f^2_{\widetilde{W}}(w) d\rho dw}{\left[\frac{\partial \overline{h}}{\partial \gamma}(\rho/\rho_0) \times \rho\right]^2}. \tag{22} $$

There are two differences between (21) and (22). The first has to do with the fact that (21) depends on the joint density of $V, \widetilde{W}$, while (22) depends only on the product of the marginals. The second difference is that in the integrate first approach the denominator has $[\partial \overline{h}(\rho/\rho_0)/\partial \gamma]^2$ instead of $[\partial h(\rho/\rho_0, w)/\partial \gamma]^2$. We next make a comparison of (21) and (22) in the special case that $V, \widetilde{W}$ are mutually independent. Define $\tau(\rho, w) = g^2(x)\sigma^2(\rho, \theta_0) f^2_V(\rho/\rho_0) f^2_{\widetilde{W}}(w)/\rho^2 f_{Z_*}(\rho, \theta_0, w)$, then we compare

$$ \int \frac{\tau(\rho, w)}{\left[\frac{\partial h}{\partial \gamma}(\rho/\rho_0, w)\right]^2} d\rho dw \text{ with } \int \frac{\tau(\rho, w)}{\left[\frac{\partial \overline{h}}{\partial \gamma}(\rho/\rho_0)\right]^2} d\rho dw. \tag{23} $$

It can be seen that the comparison in (23) could go either way, i.e., the variances could have either ranking [a good analogy is with the comparison between $E(1/Y^2)$ and $1/E^2(Y)$]. The biases of the two estimators are also different in general.

We now turn to the distribution theory of $\widehat{h}(\gamma, w)$. This is a classic generated regressors problem [like in Ahn (1995)] except that the estimator of g is a bit more complicated than in his case and the dependent variable is also generated here. The rate of convergence is determined by the effective dimensionality $d^\dagger = \max\{d_x - 1, d_w + 1\}$ [along with the worst case smoothness]. In the case where $d_w + 1 > d_x - 1$, the dominant term comes from the smooth of $\widehat{r}(X_i, W_i)$ on the known covariates $g(X_i), W_i$. In the case where $d_w + 1 > d_x - 1$, the dominant term comes from the estimation of $\widehat{g}(x)$, while when $d_w + 1 = d_x - 1$ the two terms contribute equally. We make the following additional assumption:

Assumption C. The bandwidth in (14) is chosen to be $b_* = c^* n^{-1/(2p+d^\dagger)}$ for some c^* with $0 < c^* < \infty$. We suppose that $\widehat{r}(Z_i)$ in (14) is computed with a bandwidth b_r such that $b_r = b_*/\log n$.

Let $f_{g,W}$ be the density of $(g(X), W)$, which we assume to exist and to inherit the smoothness properties of g, f_Z, and suppose that (γ, w) is an interior point for which $f_{g,W}(\gamma, w) > 0$. Define

$$\Xi(\gamma, w) = \frac{E\left[f_Z^2(Z)a^2(Z)\sigma^2(Z)|g(X) = g(x), W = w\right]}{f_{g,W}(\gamma, w)}.$$

When $a = 1/f_Z$, $\Xi(\gamma, w) = E[\sigma^2(Z)|g(X) = g(x), W = w]/f_{g,W}(\gamma, w)$, which can be recognized as the covariate dependent part of the asymptotic variance that would result were g known and a standard local polynomial smoother used. In the appendix, in (54), we define the kernel constants $\varphi(K)$ and $\psi(K)$.

Theorem 3. *Suppose that Assumptions A1, B1–B5, and C hold. Then, there exists a bounded continuous function $\beta_h(\cdot)$ such that*

$$\sqrt{nb^{*d^\dagger}}\left(\widehat{h}(\widehat{g}(x), w) - h(\widehat{g}(x), w) - b_*^p \beta_h(g(x), w)\right) \Longrightarrow N\left(0, \Sigma(g(x), w)\right),$$

where

$$\Sigma(g(x), w) = \varphi(K)\Xi(g(x), w)1(d_w + 1 \geq d_x - 1) + \psi(K)\left[\frac{\partial h}{\partial \gamma}(g(x), w)\right]^2 g^2(x) I(x_0) 1(d_w + 1 \leq d_x - 1).$$

Standard errors can be computed from this formula as above.

5 Numerical Results

5.1 Monte Carlo

In this section we describe a small monte carlo experiment. The design is $g(x) = ||x||/\sqrt{2}$, $h(g) = \exp(g)$, and $y = h[g(x)] + \varepsilon$ so there is no w. The distribution of X is uniform over the area

$\Psi_X = \{x : ||x||^2 \leq 2 \text{ and } x_j > 0.2\}$. We take $V = ||X||$ and $x_0 = (1, 1)$, which makes $g(x_0) = 1$ and $\Psi_V = (0.08, \sqrt{2})$. The nonparametric functions used in each step of the estimation are constructed using ordinary kernel regressions with a Gaussian kernel. We report results for three different sample sizes and three different error variances, for a total of nine designs. Each design is estimated using three different bandwidths h_1, h_2, and h_3, where h_2 is given by Silverman's rule ($1.06n^{-1/5}$ times the square root of the average of the regressor variances), $h_1 = 0.5 * h_2$, and $h_3 = 1.5 * h_2$. These kernel and bandwidth choices are not likely to be optimal for our setting, but are chosen because they are commonly used in applications and are easy to calculate.

For each estimated function g and h we calculate four criteria summarizing goodness of fit. These are integrated mean squared error IMSE, integrated mean absolute error IMAE, pointwise mean squared error PMSE, and pointwise mean absolute error PMAE. Results are based on a hundred simulations of each design and bandwidth. These are reported in Tables 1 and 2.

Table 1 shows that, for estimation of g the largest bandwidth produces superior estimates in all designs, with error criteria reduced by roughly 1/2 to 1/4 relative to estimates based on the smallest bandwidth. When estimating g using the large bandwidth, most criteria in most designs are approximately halved when the sample size is increased from 100 to 500 observations. Estimates of h, shown in Table 2, are generally less accurate than the estimates of g. The best choice of bandwidth for h varies across designs, but the differences in fit across different bandwidths is less pronounced for h than for g. The improvement in the fit of h when increasing the sample size from 100 to 500 varies across designs, with decreases in the error measures ranging from about 40% to 80%.

5.2 Application to Nonparametric Production Function Estimation

Let y be the log output of a firm and x be a vector of inputs. Starting from Shephard (1953), many parametric production function models of the form $y = r(x) + \varepsilon$ have been estimated that impose homotheticity. A recent example is Zellner and Ryu (1998), who provide empirical comparisons of a large number of different homothetic production functional forms. Regarding nonparametric models, Hanoch and Rothschild (1972) test whether a homothetic production function exists that could, without statistical errors, generate a given data set, and Primont and Primont (1994) provide a way to construct these homothetic production functions. However, while the production functions in these last two papers are nonparametric, by assumption they have no statistical errors and hence no associated distribution theory.

Given a homothetic production function $E(y|x) = r(x) = h[g(x)]$ with linearly homogeneous g, a property of production that is empirically important is returns to scale, defined as

$$S(g) = \frac{\partial h(g)}{\partial \ln g}$$

Other important properties are measures of substitutability of inputs, such as the technical rate of substitution and the elasticity of substitution. When x consists of just two elements, for example, capital K and labor L, then a simple measure of substitutability is

$$\alpha(K/L) = \frac{\partial \ln g\,(K/L, 1)}{\partial \ln(K/L)}$$

Note in interpreting this measure that $g\,(K/L, 1) = g\,(K, L)\,/L$. The substitutability measure $\alpha(K/L)$ equals a constant α when $g(x) = K^{\alpha}L^{1-\alpha}$, that is, when the production function $r(x)$ is a monotonic transformation of a Cobb Douglas, which is a common specification for homothetic production.

The data set consists of observations of chemical manufacturing firms in mainland China in two time periods, 1995 and 2001. For each firm, we observe the net value of real fixed assets K, the number of employees L, and Y defined as the log of value-added real output. Output and capital are measured in thousands of Yuan converted to the base year 2000 using a general price deflator for the Chinese chemical industry. For details regarding the collection and construction of this data, see Jefferson et. al. (2002). To eliminate extreme outliers, which may be due to gross errors in data reported by some firms, we sort the data by K/L and remove the bottom and top 2.5% of observations. This leaves a total sample size of 1638 firms in 2001 and 1560 firms in 1995.

We consider both nonparametric and parametric estimates of the production function $r(K, L)$. The parametric model we employ is a homothetic Translog production function, in which log output $Y = h[g\,(K, L)] + \epsilon$ with

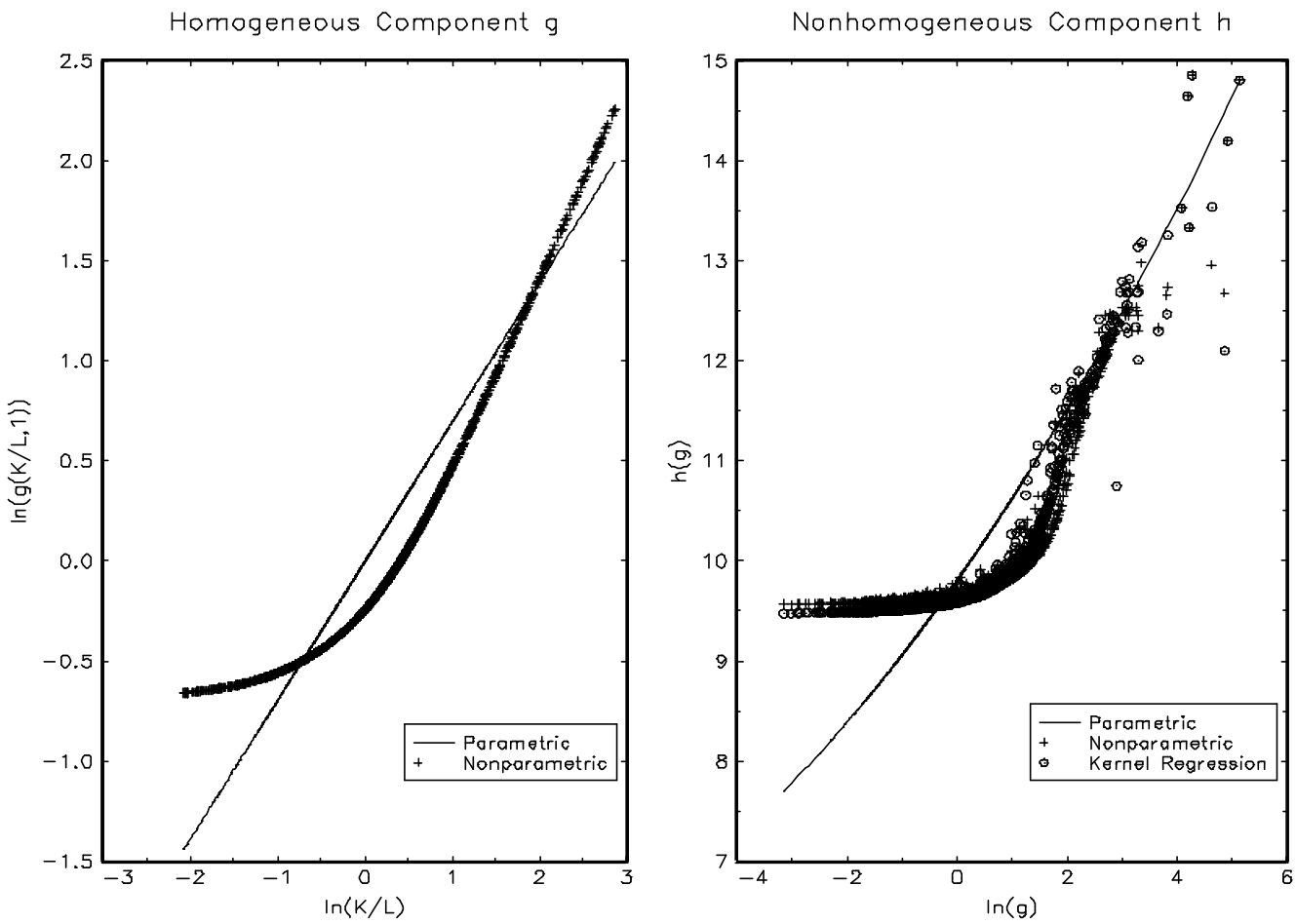
$$\begin{aligned} g\,(K, L) &= \left(\frac{K}{L}\right)^{\alpha} L \\ h\,(g) &= \beta_0 + \beta_1 \ln(g) + \beta_2 \ln(g)^2 \end{aligned}$$

Fitting this model by nonlinear least squares in each of the years of data yields the parameter estimates reported in Table 3 (standard errors are in parentheses).

TABLE 3: Parametric Translog Estimates

	α	β_0	β_1	β_2
2001 Translog	0.696	9.815	0.783	0.036
	(0.043)	(0.031)	(0.028)	(0.012)
1995 Translog	0.478	9.585	0.961	0.045
	(0.046)	(0.024)	(0.041)	(0.017)

Figures 2 and 3 show homothetic Translog and homothetic nonparametric estimates $\widehat{g}(K/L,1)$ and $\widehat{h}(g)$ in 2001. Figure 3 also shows fits from a simple nonhomothetic kernel regression of Y on K, L, that is, the initial unconstrained estimator of the function r. For simplicity, at each nonparametric estimation step we used ordinary kernel regressions with a normal kernel and bandwidth given by Silverman's rule. Without loss of generality, both the parametric and nonparametric models use the same scale normalization $g(1,1) = 1$.



Figures 2 and 3

The nonparametric fits of r and those of h shown in Figure 3 are quite similar, indicating that the imposition of homotheticity is reasonable for this data set. The nonparametric estimates of the functions g and h are roughly similar to the parametric Translog model estimates, but show quite a bit more curvature, departing most markedly from the parametric model for g at low capital to labor ratios and from the model for $h(g)$ at low values of g.

These differences are greatly magnified when one calculates the returns to scale $S(g)$ and the substitution measure $\alpha(K/L)$. For the Translog model, $S(g) = \beta_1 + 2\beta_2 \ln(g)$ and $\alpha(K/L)$ equals the constant α. For the nonparametric model we use the approximation $\widehat{S}(\widehat{g}_i) \approx [\widehat{h}(\widehat{g}_{i+1}) - \widehat{h}(\widehat{g}_{i-1})]/(\widehat{g}_{i+1} -$

$\widehat{g}_{i-1}$) after sorting the data by $\widehat{g}_i$ for each firm i, and similarly for $\widehat{\alpha}(K/L)$. Figures 4 and 5 show the results of these calculations for 2001, and Table 4 provides summary statistics for both years.

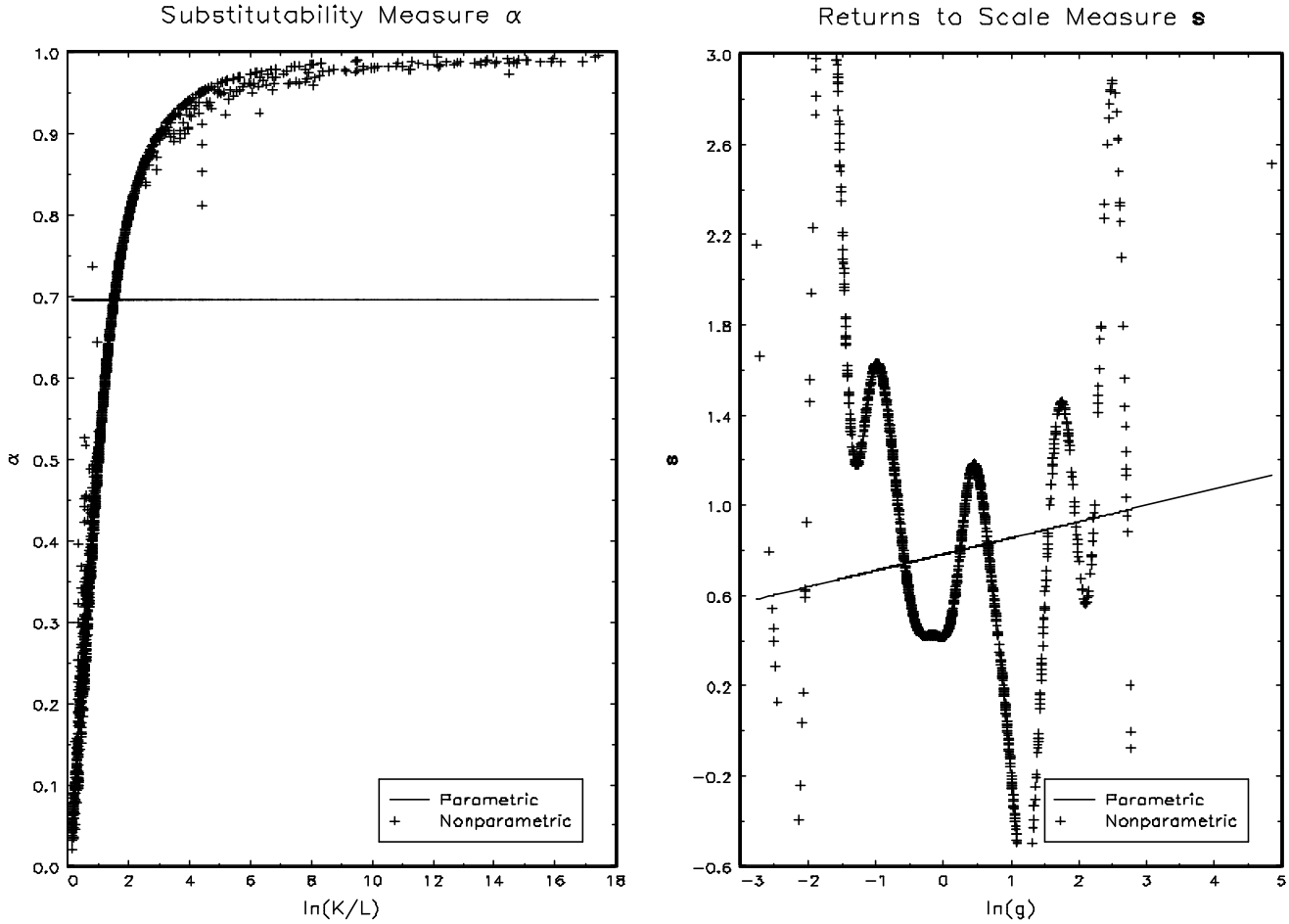
TABLE 4: Substitutability and Returns to Scale Estimates

	α parametric	α nonparametric	S parametric	S nonparametric
mean 2001	0.696	0.537	0.788	0.821
standard deviation 2001	0.000	0.281	0.082	1.286
mean 1995	0.478	0.562	0.968	1.101
standard deviation 1995	0.000	0.225	0.072	1.528

Unlike the popular homothetic Translog model, which assumes α constant, the nonparametric estimates have α sharply increasing at low capital labor ratios and leveling off only at high levels. This result indicates likely inadequacies of the parametric model. The assumption of a constant α may be more reasonable for advanced economies like the United States, which tend to have higher capital labor ratios.

The models also differ in returns to scale $S(g)$. Both models imply similar returns to scale on average, but the parametric model has $S(g)$ mildly increasing, based on a small but statistically significant positive estimate of $\widehat{\beta}_2$. In contrast, the nonparametric estimates (which are likely to be undersmoothed), are roughly U shaped, with a majority of the data in the decreasing portion of the U. Given the substantial variability of the nonparametric $\widehat{S}$, it is difficult to draw conclusions about the dependence of S on g.

The estimates based on 1995 data are broadly similar to 2001. The major difference between the two years, which can be seen in Table 4, is that average returns to scale appear to have declined over time, from approximately constant returns with average S near one in 1995, to decreasing returns with S near 0.8 in 2001. This finding could be an artifact of substantial ownership reform during this period. Many larger firms in the Chinese chemical industry may still be state-owned in 2001, while many smaller enterprises were privatized after 1995 and so could have substantially restructured, thereby enhancing their productivity. Combining these into a single cross section might then create the appearance of decreasing returns on average. This could explain the overall difference in mean S between the two years, but would explain the observed patterns in $S(g)$ within each year, though as noted above these departures of $S(g)$ from a constant are at best weakly estimated. Changes over time may more generally be due to changes in technology, demand, and other aspects of China's increasing economic liberalization and growth over this time period.



Figures 4 and 5

6 Extensions and Conclusions

We have provided a general nonparametric estimator for homothetic and homothetically separable functions, and demonstrated it in a monte carlo simulation and in an empirical production function application. We conclude by describing some extensions of our methodology.

6.1 A Generalized Partly Linear Model With Unknown Link Function

Instead of homotheticity, consider nonparametric estimation of unknown functions H and m in a model of the form

$$R(u, z, w) = H[m(z) + u, w] \tag{24}$$

given a nonparametric estimate of the function R. A sample of u, z, w is observed, where z and w are vectors and u is a scalar. Here $m(z) + u$ is the partly linear form, the function m is not constrained

to be homogeneous, and H is an unknown link function.

Examples of models of this type include reservation price and willingness to pay models[9], and models of generalized separability. Recent nonparametric or semiparametric estimators of this and closely related models include Linton and Härdle (1996), Chen and Randall (1997), Creel and Loomis (1997), An (2000), Horowitz (2001), McFadden (1999), Lewbel, Linton, and McFadden (2001), and Horowitz and Mammen (2002).

Theorem 1 uses homogeneity in $r(qv, w) = h[g(q)v, w]$ and essentially estimates g by solving this equation for v. We may similarly solve equation (24) for the scalar u to estimate the function m, as follows.

ASSUMPTION B1. Let Z and W be random vectors and U a random scalar with $\text{supp}(U, Z, W) = \Psi_U \times \Psi_Z \times \Psi_W$. There exists functions R, H and m such that $R(u, z, w) = H[m(z) + u, w]$ for all $(u, z, w) \in \Psi_U \times \Psi_Z \times \Psi_W$, where H is invertible with respect to its first element. Let $z_0 \in \Psi_Z$ be a constant vector and impose the free normalization $m(z_0) = 0$.

Given the function $R(u, z, w)$, define the function S by $S[R(u, z, w), z, w] = u$. By equation (24), the function S exists as long as H is invertible on its first element.

COROLLARY 2. *Let Assumption B1 hold. Let $(\widetilde{U}, \widetilde{W})$ be any random vector with support contained in $\Psi_U \times \Psi_W$. Then for every $(u, z, w) \in \Psi_U \times \Psi_Z \times \Psi_W$ the functions m and H satisfy*

$$m(z) = E[\widetilde{U} - S[R(\widetilde{U}, z_0, \widetilde{W}), z, \widetilde{W}]] \tag{25}$$

$$H[m(z) + u, w] = E[R(U, Z, W) \mid U = u, m(Z) = m(z), W = w]. \tag{26}$$

PROOF. Mirroring the proof of Theorem 1, we have $S(R, z, w) = H^{-1}(R, w) - m(z)$, so for any $(u, z, w) \in \Psi_U \times \Psi_Z \times \Psi_W$,

$$\begin{aligned} u - S[R(u, z_0, w), z, w] &= u - [H^{-1}(R(u, z_0, w), w) - m(z)] \\ &= u - [H^{-1}(H[m(z_0) + u, w], w) - m(z)] = m(z) \end{aligned}$$

[9] In the reservation price or willingness to pay model $y = I[-m(z) + \varepsilon \leq u]$, where u is the price or cost of a good to an individual, $-m(z) + \varepsilon$ is the individual's reservation price or willingness to pay for the good (which depends on observables z and an unobserved taste parameter ε), y is the indicator of whether the individual buys (or is willing to buy) the good at price u, $H(\varepsilon, w)$ is the conditional distribution function of ε given $W = w$. It is then assumed that $\varepsilon \mid U, Z, W \sim \varepsilon \mid W$ and $R(u, z, w) = E(Y \mid U = u, Z = z, W = w)$.

This equation holds for all $w \in \Psi_W$, $u \in \Psi_U$, so it holds in expectation replacing w with $\widetilde{W}$ and u with $\widetilde{U}$. ■

With a consistent estimator $\widehat{R}$ of the function R, a consistent estimator $\widehat{S}$ is constructed by inverting $\widehat{R}(u, z, w)$ with respect to u. Analogous to $\widehat{g}$, the corresponding consistent estimator of $m(z)$ is then given by

$$\widehat{m}(z) = \frac{1}{n} \sum_{i=1}^{n} U_i - S[R(U_i, z_0, W_i), z, \widetilde{W}_i] \tag{27}$$

and a consistent estimator of H is a nonparametric regression of $\widehat{r}(u, z, w)$ on $\widehat{m}(z) + u, w$. One could instead estimate H by regressing $\widehat{r}(u, z_0, w)$ on u, w, but that would only provide estimates of H over $\Psi_U \times \Psi_W$ instead of over $\text{supp}[M(Z) + U] \times \Psi_W$.

6.2 Endogenous Regressors

Consider estimation of $g(x)$ in the model $y = H[g(x), w, \varepsilon]$ where ε is unobserved. If $\varepsilon \perp X, W$ then $r(x, w) = E(Y \mid X = x, W = w) = h[g(x), w]$, and our estimator can be applied. However, suppose instead that some of the covariates X, W are endogenous, and so are correlated with ε. Then estimation of $g(x)$ is still possible, under the following conditions. Assume that we observe a vector of exogenous covariates Z. This Z can include exogenous elements of X and W, if any. Define $m_x(z) = E(X \mid Z = z)$, $U_x = X - m_x(Z)$, $m_w(z) = E(W \mid Z = z)$, $U_w = W - m_w(Z)$, and let $U = U_x, U_w$. Then by construction $\varepsilon \mid X, W, Z \backsim \varepsilon \mid U, Z$. Define $r(x, w, u) = E(Y \mid X = x, W = w, U = u)$ and $h[g(x), w, u] = E[H[g(x), w, \varepsilon] \mid U = u]$. Assume that $\varepsilon \mid U, Z \backsim \varepsilon \mid U$. It then follows that $r(x, w, u) = h[g(x), w, u]$. This is the form required for application of our original estimator, redefining w as w, u. If u were observed, then our estimator could be immediately applied without change to this equation. Since u is not observed, it must be estimated.

We therefore have the following estimation procedure. First, estimate $\widehat{m}_x$ and $\widehat{m}_w$ by nonparametric regressions of X and W on Z. Then let $\widehat{U}_{xi} = \widehat{m}_x(Z_i)$, for $i = 1, \ldots, n$, and similarly for $\widehat{U}_{wi}$, which together give $\widehat{U}_i$. Compute $\widehat{r}$ as a nonparametric regression of Y on $X, W, \widehat{U}$. Then apply the homotheticity estimator of the previous section, replacing W everywhere with $W, \widehat{U}$. Consistency of the resulting estimator follows from uniform consistency of the estimators in each step.

The key assumption that $\varepsilon \mid U, Z \backsim \varepsilon \mid U$ is the form of endogeneity analyzed in the control function models of Blundell and Powell (2000), (2001). It also yields a nonparametric triangular system similar to Newey, Powell, and Vella (1999) and Imbens and Newey (2001).

For every element of X, W that is also in Z, the corresponding element of U and $\widehat{U}$ will be identically zero, and hence can be ignored.

The above procedure yields estimates of the functions g and h. Recovery of the function H will in general require some additional structure. Once $g(x)$ is known, it can be treated as an observable endogenous regressor, and estimation of H (or of identifiable functionals of H that are of applied interest) then reduces to estimation of a nonparametric triangular system. Examples of estimators for such systems include Blundell and Powell (2000), (2001), Imbens and Newey (2001) and Chesher (2001). See also Matzkin (2003) for a homogeneity based method of identifying models with nonadditive errors.

6.3 Partly Linear g

Since linear functions are homogeneous, a natural way to further reduce dimensionality is with a partly linear specification for g, that is, $g(x) = x_1^\top \beta + g_2(x_2)$ for subvectors x_1, x_2, with g_2 homogeneous. Given an initial consistent estimator $\widehat{g}(x)$ using our estimator, β and g_2 could then be obtained by applying a partly linear regression estimator such as Robinson (1988), treating $\widehat{g}(x)$ as the dependent variable in a partly linear regression model. Alternatively, if g_2 is not constrained to be homogeneous, then the estimator of equation (27) could be applied first, taking V to be an element of x_1, and then treating $\widehat{m}(z)$ as the dependent variable in a partly linear regression to obtain $g_2(x_2)$ and the remaining portion of $x_1^\top \beta$.

A Appendix

Define the image sets

$$\Psi_{\widehat{r}}(w) = \{r : \widehat{r}(v \cdot x_0, w) = r,\ v \in \Psi_V\} \text{ and } \Psi_r(w) = \{r : r(v \cdot x_0, w) = r,\ v \in \Psi_V\}$$

for any $w \in \Psi_{\widetilde{W}}$. These sets trace out the values that the functions $r(.)$ and $\widehat{r}(.)$ can take - they are both intervals due to the continuity of $r, \widehat{r}$. By assumption, $\Psi_r(w)$ is non-empty for any $w \in \Psi_{\widetilde{W}}$. By the uniform consistency of $\widehat{r}(.)$, the set $\Psi_{\widehat{r}}(w)$ is also non-empty for any $w \in \Psi_{\widetilde{W}}$ with probability tending to one. In fact, for any $\epsilon > 0$

$$\Pr\left[\sup_{w \in \Psi_{\widetilde{W}}} \rho_H(\Psi_{\widehat{r}}(w), \Psi_r(w)) > \epsilon\right] \leq \Pr\left[\sup_{x \in \Psi_X, w \in \Psi_{\widetilde{W}}} |\widehat{r}(x,w) - r(x,w)| > \epsilon\right] \to 0, \tag{28}$$

where the Hausdorff distance ρ_H between two compact subsets A, B of $\mathbb{R}$ is $\rho_H(A,B) = \sup_{y \in B} \inf_{x \in A} |x - y|$. It follows that $\sup_{w \in \Psi_{\widetilde{W}}} \rho_H(\Psi_{\widehat{r}}(w), \Psi_r(w)) \longrightarrow^p 0$.

Define the event

$$\mathcal{A}_n = \left\{\widehat{r}(V_i x_0, \widetilde{W}_i) \in \Psi_r(\widetilde{W}_i) \text{ for all } i\right\},$$

which is contained in the event $\{\Psi_{\widehat{r}}(w) \subseteq \Psi_r(w)$ for all $w\}$. Note that $\mathcal{A}_n^c = \{\rho_H(\Psi_{\widehat{r}}(w), \Psi_r(w)) > \epsilon$ for some $\epsilon > 0$ and some $w\}$. A consequence of (28) is that $\Pr[\mathcal{A}_n^c] \to 0$. In the sequel, we have to compute $\Pr[\mathcal{B}_n]$ for various events $\mathcal{B}_n$. Since

$$\Pr[\mathcal{B}_n] \leq \Pr[\mathcal{B}_n \cap \mathcal{A}_n] + \Pr[\mathcal{A}_n^c] \leq \Pr[\mathcal{B}_n \cap \mathcal{A}_n] + o(1),$$

we can restrict attention to the event $\mathcal{B}_n \cap \mathcal{A}_n$. This argument ensures that we can replace, for example, the set $\Psi_{\widehat{r}}(w)$ by the set $\Psi_r(w)$ in the sequel.

A.1 Proof of Theorem 2

The proof of this theorem relies on two lemmas that are stated and proved below. We first establish the properties of $\widehat{g}(x)$. Using the fact that $1/a - 1/b = -(a-b)/ab$, we have

$$\begin{aligned}
\widehat{g}(x) - g(x) &= \frac{1}{n}\sum_{i=1}^{n} \frac{V_i}{\widehat{s}[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]} - g(x) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{V_i}{s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]} - g(x) \\
&\quad -\frac{1}{n}\sum_{i=1}^{n} V_i \left[\frac{\widehat{s}[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] - s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]}{\widehat{s}[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]}\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n} V_i \left[\frac{\widehat{s}[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] - s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]}{s^2[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]}\right] \\
&\quad +\frac{1}{n}\sum_{i=1}^{n} V_i \left[\frac{\left(\widehat{s}[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] - s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]\right)^2}{\widehat{s}^2[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]}\right],
\end{aligned}$$

because

$$\frac{1}{n}\sum_{i=1}^{n} \frac{V_i}{s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]} - g(x) \equiv 0.$$

Therefore, we just need to consider the terms

$$T_n = -\frac{1}{n}\sum_{i=1}^{n} V_i \left[\frac{\widehat{s}[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] - s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]}{s^2[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]}\right] \tag{29}$$

$$R_n = \frac{1}{n}\sum_{i=1}^{n} V_i \left[\frac{\left(\widehat{s}[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] - s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]\right)^2}{\widehat{s}^2[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]}\right]. \tag{30}$$

Remainder term. By the Cauchy-Schwarz inequality we have

$$|R_n| \leq \left(\frac{1}{n}\sum_{i=1}^{n} V_i^2\right)^{1/2} \frac{\left(\max_{1\leq i\leq n}\left|\widehat{s}[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] - s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]\right|\right)^2}{\min_{1\leq i\leq n} \widehat{s}^2[\widehat{r}(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i]}. \tag{31}$$

By the triangle inequality

$$\max_{1\leq i\leq n}\left|\widehat{s}[\widehat{r}(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]-s[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]\right|$$

$$\begin{aligned}\leq\ & \max_{1\leq i\leq n}\left|\widehat{s}[\widehat{r}(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]-s[\widehat{r}(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]\right| \\ & +\max_{1\leq i\leq n}\left|s[\widehat{r}(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]-s[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]\right|.\end{aligned}$$

We have

$$\begin{aligned}\max_{1\leq i\leq n}\left|\widehat{s}[\widehat{r}(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]-s[\widehat{r}(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]\right| &\leq \sup_{\widetilde{W}\in\Psi_{\widetilde{W}}}\sup_{r\in\Psi_r(\widetilde{W})}\left|\widehat{s}[r,x,\widetilde{W}]-s[r,x,\widetilde{W}]\right| \\ &\leq o_p(\delta_n^{1/2})\end{aligned}$$

by Lemma 1. Because s is a differentiable function with bounded continuous first derivative, we have $\max_{1\leq i\leq n}|s[\widehat{r}(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]-s[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]|=o_p(\delta_n^{1/2})$ also using the uniform consistency assumptions about $\widehat{r}$. The denominator terms in (31) are bounded away from zero with probability tending to one by the reverse triangle inequality and the uniform convergence, hence $R_n=o_p(\delta_n)$.

Leading Term. By a Taylor series expansion

$$\widehat{s}[\widehat{r}(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]-s[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]$$

$$\begin{aligned}=\ & \widehat{s}[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]-s[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i] \\ & +\frac{\partial s}{\partial r}[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]\times[\widehat{r}(V_i x_0,\widetilde{W}_i)-r(V_i x_0,\widetilde{W}_i)] \\ & +\left(\frac{\partial\widehat{s}}{\partial r}[\overline{r}_i,x,\widetilde{W}_i]-\frac{\partial s}{\partial r}[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]\right)\times[\widehat{r}(V_i x_0,\widetilde{W}_i)-r(V_i x_0,\widetilde{W}_i)],\end{aligned}$$

where $\overline{r}_i$ are intermediate points between $\widehat{r}(V_i x_0,\widetilde{W}_i)$ and $r(V_i x_0,\widetilde{W}_i)$. Substituting this expression we have

$$T_n = -\frac{1}{n}\sum_{i=1}^n V_i\left[\frac{\widehat{s}[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]-s[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]}{s^2[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]}\right] \tag{32}$$

$$-\frac{1}{n}\sum_{i=1}^n V_i\frac{\partial s}{\partial r}[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]\left[\frac{[\widehat{r}(V_i x_0,\widetilde{W}_i)-r(V_i x_0,\widetilde{W}_i)]}{s^2[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]}\right] \tag{33}$$

$$-\frac{1}{n}\sum_{i=1}^n V_i\left(\frac{\partial\widehat{s}}{\partial r}[\overline{r}_i,x,\widetilde{W}_i]-\frac{\partial s}{\partial r}[r(V_i x_0,\widetilde{W}_i),x,\widetilde{W}_i]\right)\times[\widehat{r}(V_i x_0,\widetilde{W}_i)-r(V_i x_0,\widetilde{W}_i)]. \tag{34}$$

The term (34) is bounded in probability by a constant times

$$\sup_{z\in\Psi_Z}\sup_{r\in\Psi_r(w)}\left|\frac{\partial\widehat{s}}{\partial r}(r,x,w)-\frac{\partial s}{\partial r}(r,x,w)\right|\times\sup_{z\in\Psi_Z}|\widehat{r}(x,w)-r(x,w)|=o_p(\delta_n)$$

using the same inequality as in (31) and applying B4 and Lemma 1.

To analyze (32) we invoke the expansion obtained in Lemma 1 below. Specifically, from (58) below we have

$$\begin{aligned}
&\widehat{s}[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] - s[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] \\
&= -\frac{\widehat{r}(x \cdot s(r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i), \widetilde{W}_i) - r(x \cdot s(r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i), \widetilde{W}_i)}{\frac{\partial h}{\partial \gamma}(s(r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i) \cdot g(x), \widetilde{W}_i) \times g(x)} + o_p(\delta_n). \\
&= -\frac{\widehat{r}(x \cdot V_i/g(x), \widetilde{W}_i) - r(x \cdot V_i/g(x), \widetilde{W}_i)}{\frac{\partial h}{\partial \gamma}(V_i, \widetilde{W}_i) \times g(x)} + o_p(\delta_n).
\end{aligned}$$

This expansion is uniform over $i = 1, \ldots, n$. The last line follows from the fact that by substitution, we have $s(r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i) = V_i/g(x)$ given the normalization on x_0. Regarding (33), from the definition of $s(r, q, w)$ we have

$$\frac{\partial s}{\partial r}(r, q, w) = \frac{1}{\frac{\partial h}{\partial \gamma}(h^{-1}(r, w), w)) \times g(q)}$$

so that

$$\frac{\partial s}{\partial r}[r(V_i x_0, \widetilde{W}_i), x, \widetilde{W}_i] = \frac{1}{\frac{\partial h}{\partial \gamma}(V_i, \widetilde{W}_i) \times g(x)}.$$

Substituting back into T_n, we get

$$\begin{aligned}
\widehat{g}(x) - g(x) &= \frac{1}{n}\sum_{i=1}^{n} \omega(V_i, \widetilde{W}_i) \cdot [\widehat{r}(V_i \cdot x_0, \widetilde{W}_i) - r(V_i x_0, \widetilde{W}_i)] \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} \omega(V_i, \widetilde{W}_i) \cdot [\widehat{r}(V_i \cdot x/g(x), \widetilde{W}_i) - r(V_i \cdot x/g(x), \widetilde{W}_i)] + o_p(\delta_n),
\end{aligned} \tag{35}$$

where

$$\omega(v, w) = -\frac{g(x)}{\frac{\partial h}{\partial \gamma}(v, w) \times v}. \tag{36}$$

The expression (35) is an average of the estimation error of a $d_x + d_w$ dimensional nonparametric regression over the support of the random variables $V, \widetilde{W}$. Because of this averaging, $\widehat{g}(x)$ behaves like a $d_x - 1$ dimensional smoother, see Linton and Nielsen (1995) for comparison. There is some trickyness here due to the fact that the V variable defines an integration path in $\mathbb{R}^{d_x}$ - but because we are using the polar co-ordinates this becomes standard. We now replace the sums in (35) by integrals, which follows by two applications of Lemma 2 with $u = x_0/g(x_0)$ and $u = x/g(x)$, respectively. Thus

$$\begin{aligned}
\widehat{g}(x) - g(x) &= \int [\widehat{r}(v \cdot x_0, w) - r(v \cdot x_0, w)]\omega(v, w) f_{V,\widetilde{W}}(v, w) dv dw \\
&\quad - \int [\widehat{r}(v \cdot x/g(x), w) - r(v \cdot x/g(x), w)]\omega(v, w) f_{V,\widetilde{W}}(v, w) dv dw + o_p(\delta_n) \\
&\equiv T_{n1} + T_{n2} + o_p(\delta_n).
\end{aligned}$$

Then substitute in the expansion (16) for $\widehat{r}(x,w)-r(x,w)$ and using the assumed properties

$$T_{n1}=\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d_x-1}k_b\left(\theta_{j0}-\theta_{ji}\right)u_i\widehat{\xi}_n(Z_i)+O_p(b^p)+o_p(\delta_n),\quad \text{where} \tag{37}$$

$$\begin{aligned}\widehat{\xi}_n(Z_i) &= \int k_b\left(v\cdot\rho_0-\rho_i\right)\prod_{l=1}^{d_w}k_b\left(w_l-W_{li}\right)a_n\left(v\rho_0,\theta_0,w;\frac{v\rho_0-\rho_i}{b},\frac{\theta_0-\theta_i}{b},\frac{w-W_i}{b}\right)\\ &\quad\times\omega(v,w)f_{V,\widetilde{W}}(v,w)dvdw.\end{aligned}$$

The integral in $\widehat{\xi}_n(Z_i)$ is over $\Psi_V\times\Psi_{\widetilde{W}}$, and $k_b(.)=k(./b)/b$. The $O_p(b^p)$ term in (37) is the integrated bias of $\widehat{r}(v\cdot x_0,w)-r(v\cdot x_0,w)$, and we defer analysis of this, see below. We next approximate $\widehat{\xi}_n(Z_i)$ by simpler random variables. First make a change of variables $v\mapsto u=(v\cdot\rho_0-\rho_i)/b$. It follows that $v=(\rho_i+ub)/\rho_0$ so that $dv=b\ du/\rho_0$. The range of integration over which u is taken expands to $\pm\infty$ as $b\to 0$ since the points $v\cdot\rho_0$ for all $v\in\Psi_V$ are interior to Ψ_X. Also, changing variables $w\mapsto t=(w-W_i)/b$, we get

$$\begin{aligned}\widehat{\xi}_n(Z_i) &= \frac{1}{\rho_0}\int k(u)\prod_{j=1}^{d_w}k\left(t_j\right)a_n\left(\rho_i+ub,\theta_0,W_i+tb;u,\frac{\theta_0-\theta_i}{b},t\right)\\ &\quad\times\omega((\rho_i+ub)/\rho_0,W_i+tb)f_{V,\widetilde{W}}((\rho_i+ub)/\rho_0,W_i+tb)dudt.\end{aligned} \tag{38}$$

We shall replace $\widehat{\xi}_n(Z_i)$ by the leading term

$$\widetilde{\xi}_n(Z_i)=\frac{1}{\rho_0}a(\rho_i,\theta_0,W_i)\omega(\rho_i/\rho_0,W_i)f_{V,\widetilde{W}}(\rho_i/\rho_0,W_i)=-g(x)\frac{a(\rho_i,\theta_0,W_i)f_{V,\widetilde{W}}(\rho_i/\rho_0,W_i)}{\frac{\partial h}{\partial\gamma}(\rho_i/\rho_0,W_i)\times\rho_i},$$

using the following argument. We have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d_x-1}k_b\left(\theta_{j0}-\theta_{ji}\right)u_i\widehat{\xi}_n(Z_i)\\ =\ &\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d_x-1}k_b\left(\theta_{j0}-\theta_{ji}\right)u_i\widetilde{\xi}_n(Z_i)+\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d_x-1}k_b\left(\theta_{j0}-\theta_{ji}\right)u_i[\widehat{\xi}_n(Z_i)-\widetilde{\xi}_n(Z_i)]\\ \equiv\ &T_{n1}^*+R_{n1}^*.\end{aligned}$$

Since a_n [and hence $\widehat{\xi}_n$] depends only on $Z_1,\ldots,Z_n$, we have

$$E[R_{n1}^*|Z_1,\ldots,Z_n]=0. \tag{39}$$

By the mutual independence of u_i, u_j given $Z_1, \ldots, Z_n$, the conditional variance of R^*_{n1} is just a sum of conditional expectations. Therefore, we have

$$\begin{aligned}
\operatorname{var}[R^*_{n1}|Z_1,\ldots,Z_n] &= \frac{1}{n^2}\sum_{i=1}^{n}\left[k_b\left(\theta_{j0}-\theta_{ji}\right)\right]^2 E(u_i^2|Z_i)[\widehat{\xi}_n(Z_i)-\widetilde{\xi}_n(Z_i)]^2 \\
&\leq \left[\max_{1\leq i\leq n}\left|\widehat{\xi}_n(Z_i)-\widetilde{\xi}_n(Z_i)\right|\right]^2 \times \sup_{z\in\Psi_Z}\sigma^2(z)\frac{1}{n^2}\sum_{i=1}^{n}\prod_{j=1}^{d_x-1}\left[k_b\left(\theta_{j0}-\theta_{ji}\right)\right]^2 \\
&= o_p(1)\times\frac{1}{n^2b^{2d_x-2}}\sum_{i=1}^{n}\prod_{j=1}^{d_x-1}k^2\left(\frac{\theta_{j0}-\theta_{ji}}{b}\right) = o_p(n^{-1}b^{-(d_x-1)}), \qquad (40)
\end{aligned}$$

since $n^{-2}\sum_{i=1}^{n}\prod_{j=1}^{d_x-1}\left[k_b\left(\theta_{j0}-\theta_{ji}\right)\right]^2 = O_p(n^{-1}b^{-(d_x-1)})$ by standard theory for kernel density estimators, while $\max_{1\leq i\leq n}|\widehat{\xi}_n(Z_i)-\widetilde{\xi}_n(Z_i)| = o_p(1)$ using the triangle inequality and the following bounds:

$$\begin{aligned}
\max_{1\leq i\leq n}\sup_{u,t\in\operatorname{supp}(k)}|\omega((\rho_i+ub)/\rho_0, W_i+tb)-\omega(\rho_i/\rho_0, W_i)| &= o_p(1) \\
\max_{1\leq i\leq n}\sup_{u,t\in\operatorname{supp}(k)}\left|f_{V,\widetilde{W}}((\rho_i+ub)/\rho_0, W_i+tb)-f_{V,\widetilde{W}}(\rho_i/\rho_0, W_i)\right| &= o_p(1) \\
\sup_{n\geq n_0}\max_{1\leq i\leq n}\sup_{u,t,v\in\operatorname{supp}(k)}|a_n\left(\rho_i+ub,\theta_0,W_i+tb;u,v,t\right)-a(\rho_i,\theta_0,W_i)| &= o_p(1),
\end{aligned}$$

which follow from the smoothness and boundedness of the functions ω, a, and $f_{V,\widetilde{W}}$, and assumption B3 about a_n. Together, (37), (39) and (40) imply that $T_{n1} = T^*_{n1} + O_p(b^p) + o_p(\delta_n)$.

The random variable T^*_{n1} is a sample average of independent random variables depending on (ρ_i, θ_i, W_i), with

$$\begin{aligned}
E(T^*_{n1}) &= 0 \\
\operatorname{var}(T^*_{n1}) &= \frac{1}{nb^{2(d_x-1)}}\frac{1}{\rho^2(x_0)}\int\sigma^2(\rho,\theta,w)\prod_{j=1}^{d_x-1}k\left(\frac{\theta_{j0}-\theta_j}{b}\right)^2\times \\
&\quad a^2(\rho,\theta_0,w)\omega^2(\rho/\rho_0,w)f^2_{V,\widetilde{W}}(\rho/\rho_0,w)f_{Z_*}(\rho,\theta,w)d\rho d\theta dw.
\end{aligned}$$

We change variables once again $\theta_j \longmapsto (\theta_{j0}-\theta_j)/b = t_j$, $j = 1,\ldots,d_x-1$, i.e., $\theta_j = \theta_{j0} - t_j b$, and

obtain

$$\begin{aligned}
\mathrm{var}(T_{n1}^*) &= \frac{1}{nb^{(d_x-1)}}\frac{1}{\rho^2(x_0)}\int \sigma^2(\rho,\theta_0-tb,W)\prod_{j=1}^{d_x-1}k^2(t_j)\times \\
&\quad a^2(\rho,\theta_0,w)\omega^2(\rho/\rho_0,w)f^2_{V,\widetilde{W}}(\rho/\rho_0,w)f_{Z_*}(\rho,\theta_0-tb,w)d\rho dt dw \\
&\simeq \frac{g^2(x)||k||_2^{2(d_x-1)}}{nb^{(d_x-1)}}\int \sigma^2(\rho,\theta_0,w)a^2(\rho,\theta_0,w)\frac{f^2_{V,\widetilde{W}}(\rho/\rho_0,w)f_{Z_*}(\rho,\theta_0,w)}{\left[\frac{\partial h}{\partial\gamma}(\rho/\rho_0,w)\times\rho\right]^2}d\rho dw \qquad (41)
\end{aligned}$$

using (36). The approximation in (41) is valid by dominated convergence.

Similar arguments apply to T_{n2}. We have

$$\begin{aligned}
T_{n2} &= \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d_x-1}k_b\left(\theta_j(x/g(x))-\theta_{ji}\right)u_i\int k_b\left(v\cdot\rho(x/g(x))-\rho_i\right)\times\prod_{l=1}^{d_w}k_b\left(w_l-W_{li}\right) \\
&\quad a_n(v\rho(x/g(x)),\theta(x/g(x)),w;\frac{v\rho(x/g(x))-\rho_i}{b},\frac{\theta(x)-\theta_i}{b}\frac{w-W_i}{b})\omega(v,w)f_{V,\widetilde{W}}(v,w)dvdw \\
&\quad +O_p(b^p)+o_p(\delta_n),
\end{aligned}$$

by substituting in the expansion (16) for $\widehat{r}(v\cdot x/g(x),w)-r(v\cdot x/g(x),w)$ and using the assumed properties. We have $T_{n2}=T_{n2}^*+O_p(b^p)+o_p(\delta_n)$, where

$$\begin{aligned}
T_{n2}^* &= \frac{1}{n}\frac{g(x)}{\rho(x)}\sum_{i=1}^{n}u_i\prod_{j=1}^{d_x-1}k_b\left(\theta_j(x)-\theta_{ji}\right)a(\rho_i,\theta(x),W_i)\omega(\rho_i g(x)/\rho(x),W_i)f_{V,W}(\rho_i g(x)/\rho(x),W_i) \\
&= \frac{g(x)}{n}\sum_{i=1}^{n}u_i\prod_{j=1}^{d_x-1}k_b\left(\theta_j(x)-\theta_{ji}\right)\frac{a(\rho_i,\theta(x),W_i)}{\frac{\partial h}{\partial\gamma}(\rho_i g(x)/\rho(x),W_i)\rho_i}f_{V,W}(\rho_i g(x)/\rho(x),W_i),
\end{aligned}$$

which has the variance as stated in Theorem 2 after changing variables and applying dominated convergence. Note that $\theta_j(x/g(x))=\theta_j(x)$ and $\rho(x/g(x))=\rho(x)/g(x)$.

The two terms T_{n1}^*,T_{n2}^* are asymptotically independent because for any x,x_0, with $\theta(x)\neq\theta(x_0)$, the windows $k_b\left(\theta_j(x)-\theta_{ji}\right)$ and $k_b\left(\theta_j(x_0)-\theta_{ji}\right)$ have small overlap. By the Lindeberg CLT, T_{n1}^*,T_{n2}^* are asymptotically normal, see Gozalo and Linton (2000). Finally, the $O_p(b^p)$ bias term is just obtained by integrating $\beta(v\cdot x_0,w)$ with respect to $\omega(v,w)f_{V,\widetilde{W}}(v,w)dvdw$ and $\beta(v\cdot x/g(x),w)$ with respect to $\omega(v,w)f_{V,\widetilde{W}}(v,w)dvdw$. ■

A.2 Proof of Theorem 3

We outline the proof strategy. The dependent variable $\widehat{r}(Z_i)$ in the local polynomial regression can be decomposed as the sum of three terms, and so therefore can the local polynomial estimator $\widehat{h}$. We deal with each of these three terms separately. In each case, it is necessary to expand out $\widehat{g}(X_i)$, which is the covariate in the local polynomial regression. The higher order terms in this expansion can be shown to be small using the properties of $\widehat{g}(x) - g(x)$ established in Theorem 2. We then have to manipulate the leading terms to obtain the dominant effects using integration by parts and moment calculation.

We first develop some notation that will be useful in writing out the local polynomial estimator. Let $d^* = d_x - 1$, $d^{**} = d_w + 1$. Let $c = (g(x), w)$, $\widehat{c} = (\widehat{g}(x), w)$, $C_i = (g(X_i), W_i)$, and $\widehat{C}_i = (\widehat{g}(X_i), W_i)$ all vectors in $\mathbb{R}^{d^{**}}$. Following the notation of Masry (1996a,b), let $N_\ell = \ell + d^{**} - 1!/\ell!(d^{**} - 1)!$ be the number of distinct d^{**}-tuples $\mathbf{j}$ with $|\mathbf{j}| = \sum j_k = \ell$. Arrange these N_ℓ d^{**}-tuples as a sequence in a lexicographical order and let ϕ_ℓ^{-1} denote this one-to-one map. Define $\delta_n^\dagger = n^{-1/2} b_*^{-d^\dagger/2} = n^{-p/(2p+d^\dagger)}$.

We write

$$\widehat{h}(\widehat{c}) = e_0^\top \widehat{M}_n^{-1} \widehat{D}_n \tag{42}$$

where $e_0 = (1, 0, \ldots, 0)^\top$, while $\widehat{M}_n$ and $\widehat{D}_n$ are symmetric $N \times N$ ($N = \sum_{\ell=0}^q N_\ell \times 1$, $q = p - 1$) matrix and $N \times 1$ dimensional column vector respectively defined as

$$\widehat{M}_n = \begin{bmatrix} \widehat{M}_{n,0,0} & \widehat{M}_{n,0,1} & \ldots & \widehat{M}_{n,0,q} \\ \vdots & \widehat{M}_{n,1,1} & \ldots & \widehat{M}_{n,1,q} \\ \vdots & & \ddots & \vdots \\ \widehat{M}_{n,q,0} & \ldots & \ldots & \widehat{M}_{n,q,q} \end{bmatrix}, \quad \widehat{D}_n = \begin{bmatrix} \widehat{D}_{n,0} \\ \widehat{D}_{n,1} \\ \vdots \\ \widehat{D}_{n,q} \end{bmatrix},$$

where $\widehat{M}_{n,|m|,|k|}$is a $N_{|m|} \times N_{|k|}$ dimensional submatrix with the (l, l') element given by

$$\left[\widehat{M}_{n,|m|,|k|}\right]_{l,l'} = \frac{1}{nb_*^{d^{**}}} \sum_{i=1}^n \left(\frac{\widehat{c} - \widehat{C}_i}{b_*}\right)^{\phi_{|m|}(l) + \phi_{|k|}(l')} K\left(\frac{\widehat{c} - \widehat{C}_i}{b_*}\right),$$

and $\widehat{D}_{n,|m|}$ is a $N_{|m|}$ dimensional subvector whose l-th element is given by

$$\left[\widehat{D}_{n,|m|}\right]_l = \frac{1}{nb_*^{d^{**}}} \sum_{i=1}^n \left(\frac{\widehat{c} - \widehat{C}_i}{b_*}\right)^{\phi_{|m|}(l)} K\left(\frac{\widehat{c} - \widehat{C}_i}{b_*}\right) \widehat{r}(Z_i).$$

Write

$$\widehat{r}(Z_i) = \widehat{r}(Z_i) - r(Z_i) + h(g(X_i), W_i) - h(\widehat{g}(X_i), W_i) + h(\widehat{g}(X_i), W_i),$$

where $r(Z_i) = h(g(X_i), W_i)$, and make a Taylor series expansion of $h(\widehat{g}(X_i), W_i)$ around $h(\widehat{g}(x), w)$

$$h(\widehat{C}_i) = \frac{1}{\mathbf{m}!} \sum_{0 \le |\mathbf{m}| \le q} (D^{\mathbf{m}} h)(\widehat{c})(\widehat{C}_i - \widehat{c})^{\mathbf{m}} + \widehat{\Delta}_i,$$

where $\widehat{\Delta}_i$ is the remainder term from a qth order Taylor expansion. Therefore, using the decomposition in Masry (1996, p576) we have

$$\widehat{h}(\widehat{c}) - h(\widehat{c}) = e_0^\top \widehat{M}_n^{-1}[\widehat{U}_{n1} + \widehat{U}_{n2} + \widehat{U}_{n3}], \tag{43}$$

where $\widehat{U}_{ns}$, $s = 1, 2, 3$ are $N \times 1$ vectors $\widehat{U}_{ns} = [\widehat{U}_{ns,0}, \widehat{U}_{ns,1}^\top, \ldots, \widehat{U}_{ns,q}^\top]$, where $\widehat{U}_{ns,|m|}$ are $N_{|m|}$ dimensional subvectors whose l-th elements are given by:

$$\begin{aligned}
\left[\widehat{U}_{n1,|m|}\right]_l &= \frac{1}{nb_*^{d**}} \sum_{i=1}^n \left(\frac{\widehat{c} - \widehat{C}_i}{b_*}\right)^{\phi_{|m|}(l)} K\left(\frac{\widehat{c} - \widehat{C}_i}{b_*}\right) [\widehat{r}(Z_i) - r(Z_i)] \\
\left[\widehat{U}_{n2,|m|}\right]_l &= \frac{1}{nb_*^{d**}} \sum_{i=1}^n \left(\frac{\widehat{c} - \widehat{C}_i}{b_*}\right)^{\phi_{|m|}(l)} K\left(\frac{\widehat{c} - \widehat{C}_i}{b_*}\right) [h(g(X_i), W_i) - h(\widehat{g}(X_i), W_i)] \\
\left[\widehat{U}_{n3,|m|}\right]_l &= \frac{1}{nb_*^{d**}} \sum_{i=1}^n \left(\frac{\widehat{c} - \widehat{C}_i}{b_*}\right)^{\phi_{|m|}(l)} K\left(\frac{\widehat{c} - \widehat{C}_i}{b_*}\right) \widehat{\Delta}_i.
\end{aligned}$$

We have suppressed notationally the dependence of the $\widehat{U}$ quantities on $\widehat{c}$.

We next analyze the properties of $\widehat{U}_{n1}, \widehat{U}_{n2}$, and $\widehat{U}_{n3}$. For notational simplicity we only consider in detail the first element of these vectors $\widehat{U}_{n1,0}, \widehat{U}_{n2,0}$, and $\widehat{U}_{n3,0}$, since the arguments are the same for the other elements.

Properties of $\widehat{U}_{n1,0}$. Let

$$\widetilde{U}_{n1,0} = \frac{1}{n} \sum_{i=1}^n k_{b_*}(g(x) - g(X_i)) \prod_{l=1}^{d_w} k_{b_*}(w_l - W_{li}) [\widehat{r}(Z_i) - r(Z_i)].$$

By the mean value theorem

$$\begin{aligned}
&\widehat{U}_{n1,0} - \widetilde{U}_{n1,0} \qquad (44) \\
&= \frac{1}{nb_*^2} \sum_{i=1}^n k'\left(\frac{\overline{g}(x) - \overline{g}_i}{b_*}\right) \prod_{l=1}^{d_w} k_{b_*}(w_l - W_{li}) [\widehat{g}(x) - g(x) - \widehat{g}(X_i) + g(X_i)] [\widehat{r}(Z_i) - r(Z_i)],
\end{aligned}$$

where $\overline{g}_i, \overline{g}(x)$ are intermediate points. Then

$$\begin{aligned}
|\widehat{U}_{n1,0} - \widetilde{U}_{n1,0}| &\leq 2\frac{\sup_t |k'(t)|}{b_*^2} \left(\sup_x |\widehat{g}(x) - g(x)|\right) \left(\sup_x |\widehat{r}(z) - r(z)|\right) \frac{1}{n} \sum_{i=1}^n \prod_{l=1}^{d_w} |k_{b_*}(w_l - W_{li})| \\
&= O_p\left(n^{-1/2} b^{-(d_x-1)/2} n^{-1/2} b_r^{-d/2} b_*^{-2} \log n\right) = o_p(n^{-1/2} b_*^{-d^\dagger/2}), \qquad (45)
\end{aligned}$$

because $nb^{d_x-1} b_*^{d+4-d^\dagger} (b_r/b_*)^d \to \infty$ provided $2p > d_w + 3$. We have used the facts that: $\sup_x |\widehat{g}(x) - g(x)| = O_p(n^{-1/2} b^{-(d_x-1)/2} \sqrt{\log n})$, which follows from standard arguments applied to our expansion

for $\widehat{g}(x)-g(x)$ obtained in Theorem 2 [see Masry (1996ab)], and $\sup_x |\widehat{r}(z)-r(z)| = O_p(n^{-1/2}b_r^{-d/2}\sqrt{\log n})$, which follows from B3.

Following the proof of Lemma 2, we show that

$$\widetilde{U}_{n1,0} = \int k_{b_*}\left(g(x)-g(X)\right)\prod_{l=1}^{d_w} k_{b_*}\left(w_l - W_l\right)\left[\widehat{r}(Z)-r(Z)\right] f_Z(Z)dZ + o_p(\delta_n^*). \tag{46}$$

Then substituting from the expression (16) for $\widehat{r}-r$ and interchanging summation with integration we have

$$\widetilde{U}_{n1,0} = \frac{1}{n}\sum_{i=1}^{n} u_i \tau_{n1,0,i} + O_p(b_r^p) + o_p(\delta_n^*) \equiv \widetilde{U}_{n1,0}^* + o_p(b_*^p) + o_p(\delta_n^*), \quad \text{where}$$

$$\tau_{n1,0,i} = \frac{1}{b_r^d}\int a_n\left(Z_*; \frac{Z_* - Z_{*i}}{b}\right) k_{b_*}\left(g(x)-g(X)\right)\prod_{l=1}^{d_w} k_{b_*}\left(w_l - W_l\right) K\left(\frac{Z_* - Z_{*i}}{b_r}\right) f_{Z_*}(Z_*)dZ_*.$$

Define

$$\widetilde{U}_{n1,0}^{**} = \frac{1}{n}\sum_{i=1}^{n} \tau_{n1,0,i}^* u_i, \quad \text{where} \tag{47}$$

$$\tau_{n1,0,i}^* = f_{\rho,\theta,W}(\rho_i,\theta_i,W_i)a(Z_{*i})\frac{1}{b_*^{d_w+1}}k\left(\frac{g(x)-g(X_i)}{b_*}\right)\prod_{l=1}^{d_w} k\left(\frac{w_l - W_{li}}{b_*}\right).$$

Note that $\widetilde{U}_{n1,0}^{**}$ is a sum of independent mean zero random variables with variance of order $n^{-1}b_*^{-(d_w+1)}$; it is like the stochastic part of a d_w+1 dimensional weighted kernel estimator. Furthermore, it is asymptotically normal by the Lindeberg CLT.

We now show that we can approximate $\widetilde{U}_{n1,0}^*$ by $\widetilde{U}_{n1,0}^{**}$. Write $g(x) = \rho(x)g_{**}(\theta(x))$ and $g(X) = \rho g_{**}(\theta)$, and change variables: $\rho \mapsto t = (\rho-\rho_i)/b_r$, $\theta \mapsto s = (\theta-\theta_i)/b_r$, $W \mapsto u = (W-W_i)/b_r$, we have

$$\frac{\rho(x)g_{**}(\theta(x)) - \rho g_{**}(\theta)}{b_*} \mapsto \frac{\rho(x)g_{**}(\theta(x)) - (\rho_i + tb_r)g_{**}(\theta_i + sb_r)}{b_*}.$$

By the Mean Value Theorem,

$$(\rho_i + tb_r)g_{**}(\theta_i + sb_r) - \rho_i g_{**}(\theta_i)$$

$$= \quad tb_r g_{**}(\theta_i + sb_r) + \rho_i s b_r g_{**}'(\theta_i + \overline{s}b_r) \equiv b_r\lambda_n(\rho_i,\theta_i,s,t), \tag{48}$$

where $\overline{s}$ lies between 0 and s, and under our assumptions the function $\lambda_n(\cdot,\cdot,\cdot,\cdot)$ is bounded and uniformly continuous in all its arguments. It follows that

$$\begin{aligned}\tau_{n1,0,i} \quad = \quad & \frac{1}{b_*}\int k\left(\frac{g(x)-g(X_i)}{b_*} + \frac{b_r}{b_*}\lambda_n(\rho_i,\theta_i,s,t)\right) a_n\left(\rho_i + tb_r, \theta_i + sb_r, W_i + ub; \frac{b_r}{b_*}(t,s,u)\right) k(t) \\ & \times \frac{1}{b_*^{d_w+1}}\prod_{l=1}^{d_w} k(s_l)\, k\left(\frac{w_l - W_{li}}{b_*} - \frac{b_r}{b_*}u_l\right) k(u_l)\, f_{\rho,\theta,W}(\rho_i + tb_r, \theta_i + sb_r, W_i + ub_r)dtdsdu.\end{aligned}$$

Then consider $\tau_{n1,0,i} - \tau^*_{n1,0,i}$. This includes a number of terms like, for example,

$$\begin{aligned} r_{ni} &= a(Z_{*i}) f_{\rho,\theta,W}(\rho_i, \theta_i, W_i) \frac{1}{b_*} k'\left(\frac{g(x) - g(X_i)}{b_*}\right) \frac{1}{b_*^{d_w+1}} \prod_{l=1}^{d_w} k\left(\frac{w_l - W_{li}}{b_*}\right) \\ &\quad \times \frac{b_r}{b_*} \int \lambda_n(\rho_i, \theta_i, s, t) k(t) \prod_{l=1}^{d_w} k(s_l)\, dtds. \end{aligned}$$

Note that the integral $\int \lambda_n(\rho_i, \theta_i, s, t) k(t) \prod_{l=1}^{d_w} k(s_l)\, dtds$ is finite by the properties of $\lambda_n(\cdot,\cdot,\cdot,\cdot)$. Since k has bounded support, k' is also zero outside that support and so $n^{-1}\sum_{i=1}^n u_i r_{ni} = O_p(b_r/b_*) \times O_p(\widetilde{U}^{**}_{n1,0}) = o_p(\delta^*_n)$. Similar comments apply to the other remainder terms, i.e., $n^{-1}\sum_{i=1}^n u_i(\tau_{n1,0,i} - \tau^*_{n1,0,i}) = o_p(\delta^*_n)$.

By the same sequence of arguments we can show that the leading term of $[\widehat{U}_{n1,|j|}]_l$ is

$$\left[\widetilde{U}^{**}_{n1,|m|}\right]_l = \frac{1}{n}\sum_{i=1}^n \tau^*_{n1,|m|,l,i} u_i, \quad \text{where}$$

$$\tau^*_{n1,|m|,l,i} = f_{\rho,\theta,W}(\rho_i, \theta_i, W_i) a(Z_{*i}) \frac{1}{b_*^{d_w+1}} k\left(\frac{g(x) - g(X_i)}{b_*}\right) \prod_{l=1}^{d_w} k\left(\frac{w_l - W_{li}}{b_*}\right) \left(\frac{c - C_i}{b_*}\right)^{\phi_{|m|}(l)}.$$

In conclusion

$$\widehat{U}_{n1} = \widetilde{U}^{**}_{n1} + o_p(\delta^*_n).$$

Properties of $\widehat{U}_{n2,0}$. Define

$$\widetilde{U}_{n2,0} = \frac{-1}{n}\sum_{i=1}^n k_{b_*}(g(x) - g(X_i)) \prod_{l=1}^{d_w} k_{b_*}(w_l - W_{li}) \frac{\partial h}{\partial \gamma}(g(X_i), W_i)[\widehat{g}(X_i) - g(X_i)].$$

Then by Taylor expanding $h(\widehat{g}(X_i), W_i)$ about $h(g(X_i), W_i)$ and $k_{b_*}(\widehat{g}(x) - \widehat{g}(X_i))$ about $k_{b_*}(g(x) - g(X_i))$ we obtain

$$\begin{aligned} \widehat{U}_{n2,0} - \widetilde{U}_{n2,0} &= \frac{-1}{nb^{*2}} \sum_{i=1}^n k'(\overline{g}(x) - \overline{g}_i) \prod_{l=1}^{d_w} k_{b_*}(w_l - W_{li}) \frac{\partial h}{\partial \gamma}(g(X_i), W_i)[\widehat{g}(X_i) - g(X_i)] \\ &\quad \times [\widehat{g}(x) - g(x) - \widehat{g}(X_i) + g(X_i)] \\ &\quad + \frac{-1}{2n} \sum_{i=1}^n k_{b_*}(\widehat{g}(x) - \widehat{g}(X_i)) \prod_{l=1}^{d_w} k_{b_*}(w_l - W_{li}) \frac{\partial^2 h}{\partial \gamma^2}(\widetilde{g}_i, W_i) \\ &\quad \times [\widehat{g}(x) - g(x) - \widehat{g}(X_i) + g(X_i)]^2, \end{aligned}$$

where $\overline{g}(x), \overline{g}_i, \widetilde{g}_i$ are intermediate values. The remainder term $\widehat{U}_{n2,0} - \widetilde{U}_{n2,0}$ is small by the same arguments as those used in (45). Namely, with probability tending to one, for some $C < \infty$ we have $|\widehat{U}_{n2,0} - \widetilde{U}_{n2,0}| \leq C\left(\sup_x |\widehat{g}(x) - g(x)|\right)^2 / b_*^2 = O_p(n^{-1} b_*^{-2} b^{-(d_x-1)} \log n) = o_p(\delta^*_n)$.

Following the proof of Lemma 2 we show that

$$\widetilde{U}_{n2,0} = -\int k_{b_*}(g(x)-g(X))\prod_{l=1}^{d_w} k_{b_*}(w_l - W_l)\frac{\partial h}{\partial \gamma}(g(X),W)[\widehat{g}(X)-g(X)]f_Z(X,W)dXdW + o_p(\delta_n). \tag{49}$$

We substitute into $\widetilde{U}_{n2,0}$ the expansion we already obtained for $\widehat{g}$ in Theorem 2, and interchanging summations and integrals we obtain

$$\widetilde{U}_{n2,0} = A_{n1} + A_{n2} + O_p(b^p) + o_p(\delta_n), \quad \text{where:}$$

$$\begin{aligned} A_{n1} &= -\frac{1}{n}\sum_{i=1}^{n} u_i \prod_{l=1}^{d_x-1} k_b(\theta_{l0}-\theta_{li})\frac{a(\rho_i,\theta_0,W_i)}{\frac{\partial h}{\partial \gamma}(\rho_i/\rho_0,W_i)\rho_i} f_{V,\widetilde{W}}(\rho_i/\rho_0,W_i)\times \\ &\quad \int k_{b_*}(g(x)-g(X))\prod_{l=1}^{d_w} k_{b_*}(w_l-W_l)\frac{\partial h}{\partial \gamma}(g(X),W)g(X)f_Z(X,W)dXdW \\ A_{n2} &= \frac{1}{n}\sum_{i=1}^{n} u_i \int \prod_{l=1}^{d_x-1} k_b(\theta_l(X)-\theta_{li})\frac{a(\rho_i,\theta(X),W_i)}{\frac{\partial h}{\partial \gamma}(\rho_i g(X)/\rho(X),W_i)\rho_i} f_{V,\widetilde{W}}(\rho_i g(X)/\rho(X),W_i)\times \\ &\quad k_{b_*}(g(x)-g(X))\prod_{l=1}^{d_w} k_{b_*}(w_l-W_l)\frac{\partial h}{\partial \gamma}(g(X),W)g(X)f_Z(X,W)dXdW. \end{aligned}$$

Note that

$$\begin{aligned} &\int k_{b_*}(g(x)-g(X))\prod_{l=1}^{d_w} k_{b_*}(w_l-W_l)\frac{\partial h}{\partial \gamma}(g(X),W)g(X)f_Z(X,W)dXdW \\ &= \int k_{b_*}(g(x)-g)\prod_{l=1}^{d_w} k_{b_*}(w_l-W_l)\frac{\partial h}{\partial \gamma}(g,W)gf_{g,W}(g,W)dgdW \\ &= \int k(t)\prod_{l=1}^{d_w} k(s_l)\frac{\partial h}{\partial \gamma}(g(x)+tb,w+sb_*)[g(x)+tb_*]f_{g,W}(g(x)+tb,w+sb_*)dtds \\ &= g(x)f_{g,W}(g(x),w)\left.\frac{\partial h(\gamma,w)}{\partial \gamma}\right|_{\gamma=g(x)} + o(1) \end{aligned}$$

by a change of variables [$g \mapsto t = (g(x)-g)/b_*$ and $W \mapsto s = (w-W)/b_*$] and dominated convergence, using $\int k(t)\prod_{l=1}^{d_w} k(s_l)\,dtds = 1$. It follows that

$$\begin{aligned} A_{n1} &= -g(x)f_{g,W}(g(x),w)\left.\frac{\partial h(\gamma,w)}{\partial \gamma}\right|_{\gamma=g(x)} \\ &\quad \times\frac{1}{n}\sum_{i=1}^{n} u_i \prod_{l=1}^{d_x-1} k_b(\theta_{l0}-\theta_{li})\frac{a(\rho_i,\theta_0,W_i)f_{V,\widetilde{W}}(\rho_i/\rho_0,W_i)}{\frac{\partial h}{\partial \gamma}(\rho_i/\rho_0,W_i)\rho_i}[1+o_p(1)] \\ &= O_p(n^{-1/2}b^{-(d_x-1)/2}). \end{aligned} \tag{50}$$

Regarding A_{n2}, we can write $A_{n2} = n^{-1}\sum_{i=1}^{n} u_i t_{ni}$, where:

$$\begin{aligned} t_{ni} &= \int k_{b_*}\left(g(x)-g(X)\right)\prod_{l=1}^{d_x-1} k_b\left(\theta_l(X)-\theta_{li}\right)\prod_{l=1}^{d_w} k_{b_*}\left(w_l-W_l\right)\times \\ &\quad \frac{a(\rho_i,\theta(X),W_i)}{\frac{\partial h}{\partial\gamma}(\rho_i g(X)/\rho(X),W_i)\rho_i} f_{V,\widetilde{W}}(\rho_i g(X)/\rho(X),W_i)\frac{\partial h}{\partial\gamma}(g(X),W)g(X)f_Z(X,W)dXdW \\ &= \int k_{b_*}\left(\rho(x)g_{**}(\theta(x))-\rho g_{**}(\theta)\right)\prod_{l=1}^{d_x-1} k_b\left(\theta_l-\theta_{li}\right)\prod_{l=1}^{d_w} k_{b_*}\left(w_l-W_l\right)\times \\ &\quad \frac{a(\rho_i,\theta,W_i)}{\frac{\partial h}{\partial\gamma}(\rho_i g_{**}(\theta),W_i)\rho_i} f_{V,\widetilde{W}}(\rho_i\theta,W_i)\frac{\partial h}{\partial\gamma}(\rho g_{**}(\theta),W)\rho g_{**}(\theta)f_{Z*}(\rho,\theta,W)d\rho d\theta dW, \end{aligned}$$

after changing to polar co-ordinates where $g(x)=\rho(x)g_{**}(\theta(x))$ and $g(X)=\rho g_{**}(\theta)$. We now change variables $\theta\mapsto t=(\theta-\theta_i)/b$, $W\mapsto u=(w-W)/b_*$, $\rho\mapsto v=(\rho-\rho_i)/b_*$, and get

$$\begin{aligned} \frac{\rho(x)g_{**}(\theta(x))-\rho g_{**}(\theta)}{b_*} &\mapsto \frac{\rho(x)g_{**}(\theta(x))-(\rho_i+vb_*)g_{**}(\theta_i+tb)}{b_*} \\ &= \frac{\rho(x)g_{**}(\theta(x))-\rho_i g_{**}(\theta_i)}{b_*}-\lambda_n^*(\rho_i,\theta_i,v,t), \end{aligned}$$

where $\lambda_n^*(\rho_i,\theta_i,v,t)=vg_{**}(\theta_i+tb)+\rho_i t g'_{**}(\theta_i+\bar{t}b)(b/b_*)$ with $\bar{t}$ being intermediate values. When $\limsup b/b_*<\infty$, $\lambda_n^*(\rho_i,\theta_i,v,t)$ is bounded and uniformly continuous. We have

$$\begin{aligned} t_{ni} &= \int k\left(\frac{g(x)-g(X_i)}{b_*}-\lambda_n^*(\rho_i,\theta_i,v,t)\right)\prod_{l=1}^{d_w} k\left(u_l\right)\prod_{l=1}^{d_x-1} k\left(t_l\right)\times \\ &\quad \times\frac{a(\rho_i,\theta_i+tb,W_i)}{\frac{\partial h}{\partial\gamma}(\rho_i g_{**}(\theta_i+tb),W_i)\rho_i} f_{V,\widetilde{W}}(\rho_i g_{**}(\theta_i+tb),W_i) \\ &\quad \times\frac{\partial h}{\partial\gamma}((\rho_i+b_*v)g_{**}(\theta_i+tb),w+b_*u)(\rho_i+b_*v)g_{**}(\theta_i+tb) \\ &\quad \times f_{Z*}(\rho_i+b_*v,\theta_i+tb,w+b_*u)dvdudt. \end{aligned}$$

It follows that t_{ni} is bounded in probability and depends only on Z_i. Therefore, $A_{n2}=O_p(n^{-1/2})$.

In conclusion $\widehat{U}_{n2,0}=\widetilde{U}_{n2,0}^{**}+o_p(n^{-1/2}b^{-(d_x-1)/2})$, where

$$\widetilde{U}_{n2,0}^{**} = -g(x)f_{g,W}(g(x),w)\frac{\partial h(g(x),w)}{\partial\gamma}\frac{1}{n}\sum_{i=1}^{n} u_i\prod_{l=1}^{d_x-1} k_b\left(\theta_{l0}-\theta_{li}\right)\frac{a(\rho_i,\theta_0,W_i)f_{V,\widetilde{W}}(\rho_i/\rho_0,W_i)}{\frac{\partial h}{\partial\gamma}(\rho_i/\rho_0,W_i)\rho_i} \quad (51)$$

Furthermore, it is easy to see that the leading term of $\left[\widehat{U}_{n2,|m|}\right]_l$ is

$$
\begin{aligned}
&= -\frac{1}{n}\sum_{i=1}^{n} u_i \prod_{l=1}^{d_x-1} k_b\left(\theta_{l0}-\theta_{li}\right) \frac{a(\rho_i,\theta_0,W_i)}{\frac{\partial h}{\partial \gamma}(\rho_i/\rho_0,W_i)\rho_i} f_{V,\widetilde{W}}(\rho_i/\rho_0,W_i) \int \frac{\partial h}{\partial \gamma}(g(X),W)g(X) \times \\
&\quad \left(\frac{g(x)-g(X)}{b_*},\frac{w-W}{b_*}\right)^{\phi_{|m|}(l)} k_{b_*}\left(g(x)-g(X)\right) \prod_{l=1}^{d_w} k_{b_*}\left(w_l-W_l\right) f_Z(X,W)dXdW \\
&\simeq -\frac{1}{n}\sum_{i=1}^{n} u_i \prod_{l=1}^{d_x-1} k_b\left(\theta_{l0}-\theta_{li}\right) \frac{a(\rho_i,\theta_0,W_i)}{\frac{\partial h}{\partial \gamma}(\rho_i/\rho_0,W_i)\rho_i} f_{V,\widetilde{W}}(\rho_i/\rho_0,W_i) \times \int \prod_{l=1}^{d_w+1} k\left(v_l\right) v^{\phi_{|m|}(l)} dv
\end{aligned}
$$

and in fact

$$
\left[\widehat{U}_{n2,|m|}\right]_l = \int \prod_{l=1}^{d_w+1} k\left(v_l\right) v^{\phi_{|m|}(l)} dv \times \widetilde{U}^{**}_{n2,0} + o_p(n^{-1/2}b^{-(d_x-1)/2}).
$$

In conclusion

$$
\widehat{U}_{n2} = \widetilde{U}^{**}_{n2} + o_p(\delta^*_n),
$$

where $\widetilde{U}^{**}_{n2}$ is the vector with components $\int \prod_{l=1}^{d_w+1} k\left(v_l\right) v^{\phi_{|m|}(l)} dv \times \widetilde{U}^{**}_{n2,0}$.

Properties of $\widehat{U}_{n3,0}$. We have

$$
\widehat{U}_{n3,0} = O_p(b^p_*) \tag{52}
$$

by a lengthy argument similar to those already given using the fact that $\widehat{\Delta}_i = O_p(b^p_*)$.

We have established the properties of the three numerator terms in (43); it remains to determine the properties of the denominator and combine the results. Define the $N \times 1$ vector $m = (m_0, m_1^\top, \ldots, m_q^\top)^\top$ and $N \times N$ dimensional matrices

$$
M = \begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,q} \\ M_{1,0} & M_{1,1} & \cdots & M_{1,q} \\ \vdots & & & \vdots \\ M_{q,0} & M_{q,1} & \cdots & M_{q,q} \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \Gamma_{0,0} & \Gamma_{0,1} & \cdots & \Gamma_{0,q} \\ \Gamma_{1,0} & \Gamma_{1,1} & \cdots & \Gamma_{1,q} \\ \vdots & & & \vdots \\ \Gamma_{q,0} & \Gamma_{q,1} & \cdots & \Gamma_{q,q} \end{bmatrix}, \tag{53}
$$

where m_i are $N_i \times 1$ vectors and $M_{i,j}, \Gamma_{i,j}$ are $N_i \times N_j$ dimensional matrices whose ℓ and (ℓ, m) element are respectively $\mu_{\phi_i(\ell)}$, $\mu_{\phi_i(\ell)+\phi_j(m)}$ and $\nu_{\phi_i(\ell)+\phi_j(m)}$, with for $0 \le |\mathbf{j}| \le 2q$, $\mu_{\mathbf{j}}(k) = \int u^{\mathbf{j}} \prod_{l=1}^{d_w+1} k(u_l)du$ and $\nu_{\mathbf{j}}(k) = \int u^{\mathbf{j}} \prod_{l=1}^{d_w+1} k(u_l)^2 du$. Masry (1996, 3.9). Define also

$$
\varphi(K) = \left[M^{-1}\Gamma M^{-1}\right]_{0,0} \text{ and } \psi(K) = ||k||_2^{2(d_x-1)} \left[M^{-1}mm^\top M^{-1}\right]_{0,0}, \tag{54}
$$

where $[A]_{0,0}$ signifies the upper-left element of matrix A.

Defining the matrix $\widetilde{M}_n$ with components

$$
\left[\widetilde{M}_{n,|m|,|k|}\right]_{l,l'} = \frac{1}{nb_*^{d^{**}}} \sum_{i=1}^{n} \left(\frac{c-C_i}{b_*}\right)^{\phi_{|m|}(l)+\phi_{|k|}(l')} K\left(\frac{c-C_i}{b_*}\right),
$$

we show that $\widehat{M}_n = \widetilde{M}_n + O_p(\delta_n) + O_p(\delta_n^*)$. By Masry (1996, Corollary 1)

$$\widetilde{M}_n = E\left(\widetilde{M}_n\right) + O_p\left(\sqrt{\frac{\log n}{nb_*^{d_w+1}}}\right), \tag{55}$$

where the deterministic quantity $E(\widetilde{M}_n) = f_{g,W}(g(x), w)M + O_p(b_*)$. It follows that: $e_0^\top \widehat{M}_n^{-1}\widehat{U}_{n1} = e_0^\top M^{-1}\widetilde{U}_{n1}^{**} + o_p(\delta_n^*)$, $e_0^\top \widehat{M}_n^{-1}\widehat{U}_{n2} = e_0^\top M^{-1}\widetilde{U}_{n2}^{**} + o_p(\delta_n^*)$, and $e_0^\top \widehat{M}_n^{-1}\widehat{U}_{n3} = O_p(b_*^p)$, where $\widetilde{U}_{n1}^{**} = O_p(n^{-1/2}b_*^{-(d_w+1)/2})$ and $\widetilde{U}_{n2}^{**} = O_p(n^{-1/2}b^{-(d_x-1)/2})$. When $d_x - 1 > d_w + 1$, the term $\widetilde{U}_{n2}^{**}$ is of larger order in probability than $\widetilde{U}_{n1}^{**}$ and is the same order in probability as $\widehat{g}(x) - g(x)$, while if $d_x - 1 < d_w + 1$, the term $\widetilde{U}_{n1}^{**}$ dominates both $\widetilde{U}_{n2}^{**}$ and $\widehat{g}(x) - g(x)$.

Therefore,

$$\widehat{h}(\widehat{c}) - h(\widehat{c}) = \frac{1}{f_{g,W}(g(x), w)}\left(e_0^\top M^{-1}\widetilde{U}_{n1}^{**} + e_0^\top M^{-1} m\widetilde{U}_{n2,0}^{**}\right) + O(b^p) + O(b_*^p) + o_p(\delta_n^*),$$

where:

$$\begin{aligned}
\operatorname{var}\left[\frac{e_0^\top M^{-1}\widetilde{U}_{n1}^{**}}{f_{g,W}(g(x), w)}\right] &= \frac{1}{nb_*^{d_w+1}}\varphi(K)\frac{E\left[f_Z^2(Z)a^2(Z)\sigma^2(Z)|g(X) = g(x), W = w\right]}{f_{g,W}(g(x), w)} \\
\operatorname{var}\left[\frac{e_0^\top M^{-1}\widetilde{U}_{n2}^{**}}{f_{g,W}(g(x), w)}\right] &= \frac{1}{nb^{d_x-1}}\psi(K)\left[\frac{\partial h(g(x), w)}{\partial \gamma}\right]^2 \times \\
&\quad g^2(x)\int \frac{\sigma^2(\rho, \theta_0, w)a^2(\rho, \theta_0, w)f_{V,\widetilde{W}}^2(\rho/\rho_0, w)f_{Z*}(\rho, \theta(x), w)}{[\frac{\partial h}{\partial \gamma}(\rho/\rho_0, w)]^2\rho^2}d\rho dw \\
&= \frac{1}{nb^{d_x-1}}\psi(K)\left[\frac{\partial h(g(x), w)}{\partial \gamma}\right]^2 g^2(x)I(x_0).
\end{aligned}$$

The two random variables $\widetilde{U}_{n1}^{**}, \widetilde{U}_{n2,0}^{**}$ are asymptotically mutually uncorrelated.

■

A.3 Lemmas

In Lemma 1 we derive a uniform asymptotic expansion for $\widehat{s}(r,x,w)$ and its derivatives, while in Lemma 2 we state a version of the well-known lemma which replaces the sums in (35) by integrals.

LEMMA 1. *We have*

$$\sup_{z\in\Psi_Z}\sup_{r\in\Psi_r(w)}|\widehat{s}(r,x,w)-s(r,x,w)|=o_p(\delta_n^{1/2}). \tag{56}$$

$$\sup_{z\in\Psi_Z}\sup_{r\in\Psi_r(w)}\left|\frac{\partial\widehat{s}}{\partial r}(r,x,w)-\frac{\partial s}{\partial r}(r,x,w)\right|=o_p(\delta_n^{1/2}). \tag{57}$$

Furthermore,

$$\widehat{s}(r,x,w)-s(r,x,w)=-\frac{\widehat{r}(x\cdot s(r,x,w),w)-r}{\frac{\partial h}{\partial\gamma}(s(r,x,w)\cdot g(x),w)\times g(x)}+R_s(r,x,w), \tag{58}$$

where $\sup_{z\in\Psi_Z}\sup_{r\in\Psi_r(w)}|R_s(r,x,w)|=o_p(\delta_n)$.

LEMMA 2. *Let* $\omega(v,w)$ *be a continuous function. Then for any* $x^*\in\Psi_X$

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\omega(V_i,\widetilde{W}_i)\cdot[\widehat{r}(V_i\cdot x^*,\widetilde{W}_i)-r(V_i\cdot x^*,\widetilde{W}_i)]\\ =&\int\omega(v,w)\cdot[\widehat{r}(v\cdot x^*,w)-r(v\cdot x^*,w)]f_{V,\widetilde{W}}(v,w)dvdw+o_p(\delta_n).\end{aligned} \tag{59}$$

PROOF OF LEMMA 1. Note that our estimator $\widehat{s}(r,x,w)$ is defined for all $r\in\mathbb{R}$. Define for all $r\in\mathbb{R}$: $Q(v;x,w,r)=[r(q_0(x)\cdot v,w)-r]^2$ and $\widehat{Q}(v;x,w,r)=[\widehat{r}(q_0(x)\cdot v,w)-r]^2$. Both functions are continuous in r over $\mathbb{R}$ and x,w over Ψ_X,Ψ_W. We interpret the value $s^\dagger=s^\dagger(r,x,w)$ as the unique minimizer of $Q(v;x,w,r)$ over $v\in\Psi_V$ and let $s=s(r,x,w)=s^\dagger(r,x,w)/v_0(x)$. Then take $\widehat{s}^\dagger(r,x,w)$ to be any approximate minimizer of $\widehat{Q}(v;x,w,r)$ over $v\in\Psi_V$, and $\widehat{s}(r,x,w)=\widehat{s}^\dagger(r,x,w)/v_0(x)$ as defined in (13). Because $\widehat{Q}(v;x,w,r)$ is a continuous function of v over the compact set Ψ_V, a minimum exists and hence so does $\widehat{s}(r,x,w)$. Furthermore, $|\widehat{s}(r,x,w)|\leq\left(\sup_{v\in\Psi_V}v\right)/v_0(x)$, and is bounded provided $\inf_{x\in\Psi_X}v_0(x)>0$.

We first prove consistency of the un-normalized quantity $\widehat{s}^\dagger(r,x,w)$. The proof follows the main steps of similar results in the literature for optimization estimators (see for example Pötscher and Prucha (1991, Lemmas 3.1 and 4.2) or Andrews (1995, Lemma A1)). It relies on the use of the

identification assumption and a uniform weak law of large numbers. We have

$$\begin{aligned}
\sup_{v\in\Psi_V}|\widehat{Q}(v;x,w,r)-Q(v;x,w,r)| &\leq 2\sup_{v\in\Psi_V}|\widehat{r}(q_0(x)\cdot v,w)-r(q_0(x)\cdot v,w)|\times|r(q_0(x)\cdot v,w)-r| \\
&\quad+\sup_{v\in\Psi_V}|\widehat{r}(q_0(x)\cdot v,w)-r(q_0(x)\cdot v,w)|^2 \\
&\leq 2\sup_{x\in\Psi_X}|\widehat{r}(x,w)-r(x,w)|\times|r(x,w)-r| \qquad (60)\\
&\quad+\sup_{x\in\Psi_X}|\widehat{r}(x,w)-r(x,w)|^2 \\
&= o_p(1). \qquad (61)
\end{aligned}$$

This result is true for any given $r\in\mathbb{R}$ and $x\in\Psi_X, w\in\Psi_{\widetilde{W}}$ by the uniform consistency of $\widehat{r}(x,w)$ and the compactness of the relevant sets. Since $s^\dagger(r,x,w)$ is the unique minimizer of the continuous function $Q(v;x,w,r)$, we obtain the consistency of $\widehat{s}^\dagger(r,x,w)$ by Theorem 2.1 of Newey and McFadden (1994), i.e.,

$$\widehat{s}^\dagger(r,x,w)=s^\dagger(r,x,w)+o_p(1). \qquad (62)$$

The result (61) holds uniformly over $r\in\Psi_r(w)$ and $x\in\Psi_X, w\in\Psi_{\widetilde{W}}$ because of the uniform consistency of $\widehat{r}(x,w)$ and the compactness of the sets $\Psi_r(w)$, Ψ_X, $\Psi_{\widetilde{W}}$.

As a consequence of B2(b) we have that for all x,w,r, and for all $\overline{\eta}\geq\eta>0$ there exists $\epsilon(x,w,r)\geq\epsilon>0$ such that

$$\inf_{v:|v-s(x,w,r)|>\eta}Q(v;x,w,r)\geq\epsilon(x,w,r). \qquad (63)$$

This follows because

$$\begin{aligned}
\frac{\partial Q}{\partial v}(v;x,w,r) &= 2\left[r(q_0(x)\cdot v,w)-r\right]\frac{\partial r}{\partial x^\top}(q_0(x)\cdot v,w)\times q_0(x) \\
&= 0 \text{ at } v=s(x,w,r)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 Q}{\partial v^2}(v;x,w,r) &= 2\left[\frac{\partial r}{\partial x^\top}(q_0(x)\cdot v,w)\times q_0(x)\right]^2+2\left[r(q_0(x)\cdot v,w)-r\right]q_0^\top(x)\frac{\partial^2 r}{\partial x\partial x^\top}(q_0(x)\cdot v,w)q_0(x) \\
&= 2\left[\frac{\partial r}{\partial x^\top}(q_0(x)\cdot s(x,w,r),w)\times q_0(x)\right]^2 \text{ at } v=s(x,w,r).
\end{aligned}$$

Since g is homogenous of degree one, it satisfies Euler's Law whence $g(x)=x^\top\partial g(x)/\partial x$ for any x, so that since $g(x)\neq 0$, we have

$$q_0^\top(x)\frac{\partial r}{\partial x}(q_0(x)\cdot s(x,w,r),w)=\frac{1}{v_0(x)}\frac{\partial h}{\partial\gamma}(g(x),w)x^\top\frac{\partial g}{\partial x}(x)=\frac{1}{v_0(x)}\frac{\partial h}{\partial\gamma}(g(x),w)g(x)\neq 0,$$

and (63) follows by uniform continuity.

Therefore, using the triangle inequality

$$\begin{aligned}
& \Pr\left[\exists(x,w,r) : |\widehat{s}(x,w,r) - s(x,w,r)| \geq \eta\right] \\
\leq\ & \Pr\left[\exists(x,w,r) : Q(\widehat{s}(x,w,r);x,w,r) \geq \epsilon(x,w,r)\right] \\
\leq\ & \Pr\left[\exists(x,w,r) : \widehat{Q}(\widehat{s}(x,w,r);x,w,r) \geq \epsilon/2\right] \\
& + \Pr\left[\sup_{v,x,w,r}\left|\widehat{Q}(v;x,w,r) - Q(v;x,w,r)\right| \geq \epsilon/2\right] \\
\leq\ & \Pr\left[\exists(x,w,r) : \widehat{Q}(\widehat{s}(x,w,r);x,w,r) \geq \epsilon/2\right] + o(1) \\
\leq\ & \Pr\left[\exists(x,w,r) : \widehat{Q}(s(x,w,r);x,w,r) \geq \epsilon/2\right] + o(1) \\
\leq\ & \Pr\left[\sup_{v,x,w,r}\left|\widehat{Q}(v;x,w,r) - Q(v;x,w,r)\right| \geq \epsilon/2\right] + o(1) = o(1).
\end{aligned}$$

It follows that

$$\sup_{z\in\Psi_Z}\sup_{r\in\Psi_r(w)}|\widehat{s}(r,x,w) - s(r,x,w)| = o_p(1).$$

We next establish the uniform asymptotic expansion for $\widehat{s}^\dagger(r,x,w)$. Define $\widehat{M}(v;x,w,r) = \widehat{r}(q_0(x)\cdot v,w) - r$ and its derivative

$$\frac{\partial\widehat{M}(v;x,w,r)}{\partial v} = \sum_{j=1}^{d_x} q_{0j}(x)\frac{\partial\widehat{r}}{\partial x_j}(q_0(x)\cdot v,w).$$

We know that

$$\widehat{M}(\widehat{s}^\dagger(r,x,w);x,w,r) = o_p(\delta_n).$$

Define also the corresponding population moment condition $M(v;r,x,w) = r(q_0(x)\cdot v,w) - r$ and its derivative

$$\frac{\partial M(v;x,w,r)}{\partial v} = \frac{\partial h}{\partial\gamma}(v\cdot g(q_0(x)),w)\cdot g(q_0(x)).$$

By a Taylor expansion any consistent sequence $\widehat{s}^\dagger(r,x,w)$ satisfies

$$\begin{aligned}
o_p(\delta_n) &= \widehat{M}(\widehat{s}^\dagger(r,x,w);x,w,r) \qquad (64)\\
&= \widehat{M}(s^\dagger(r,x,w);x,w,r) + \left.\frac{\partial\widehat{M}(v;x,w,r)}{\partial v}\right|_{v=\overline{s}^\dagger(r,x,w)} v_0(x)(\widehat{s}(r,x,w) - s(r,x,w)),
\end{aligned}$$

where $\overline{s}^{\dagger}(r,x,w)$ are intermediate values. Since $\widehat{s}(r,x,w)$ is consistent, with probability tending to one we have $x\overline{s}(r,x,w) \in \Psi_X$. Therefore, for some $\epsilon_n \to 0$ we have with probability tending to one

$$\left|\frac{\partial \widehat{M}(\overline{s}^{\dagger}(r,x,w);x,w,r)}{\partial v} - \frac{\partial M(s^{\dagger}(r,x,w);x,w,r)}{\partial v}\right|$$

$$\begin{aligned}
&\leq \sup_{|s-s^{\dagger}(r,x,w)|\leq \epsilon_n}\left|\frac{\partial \widehat{M}(s;x,w,r)}{\partial v} - \frac{\partial M(s;x,w,r)}{\partial v}\right| \\
&\quad + \left|\frac{\partial M(\overline{s}^{\dagger}(r,x,w);x,w,r)}{\partial v} - \frac{\partial M(s^{\dagger}(r,x,w);x,w,r)}{\partial v}\right| \\
&\leq \sup_{x\in\Psi_X}\left|\sum_{j=1}^{d_x} q_{0j}(x)\left\{\frac{\partial \widehat{r}}{\partial x_j}(x,w) - \frac{\partial r}{\partial x_j}(x,w)\right\}\right| + o_p(1) = o_p(1).
\end{aligned}$$

Since $\partial M(s(r,x,w);x,w,r)/\partial v > 0$, it follows that

$$\widehat{s}(r,x,w) - s(r,x,w) = -\left[\frac{\partial M(s^{\dagger}(r,x,w);x,w,r)}{\partial v}\right]^{-1}\frac{1}{v_0(x)}\widehat{M}(s^{\dagger}(r,x,w);x,w,r)[1+o_p(1)],$$

i.e.,

$$\widehat{s}(r,x,w) - s(r,x,w) = -\frac{\widehat{r}(x\cdot s(r,x,w),w) - r}{\frac{\partial h}{\partial \gamma}(s(r,x,w)\cdot g(x),w)\times g(x)}[1+o_p(1)]. \tag{65}$$

This is true for all r,x,w and the error term in (65) is uniform over r,x,w under our assumptions. This concludes the uniform asymptotic expansion for $\widehat{s}(r,x,w)$.

We now give the proof of (57). Differentiate the first order condition (64) with respect to r and use the chain rule to obtain

$$\begin{aligned}
o_p(\delta_n) &= \left.\frac{\partial \widehat{M}(v;x,w,r)}{\partial r}\right|_{\downarrow v=\widehat{s}^{\dagger}(r,x,w)} + \frac{\partial \widehat{M}(\widehat{s}^{\dagger}(r,x,w);x,w,r)}{\partial \widehat{s}}\frac{\partial \widehat{s}^{\dagger}(r,x,w)}{\partial r} \\
&= -1 + \sum_{j=1}^{d_x} x_j \frac{\partial \widehat{r}}{\partial x_j}(x\cdot \widehat{s}(r,x,w),w)\frac{\partial \widehat{s}(r,x,w)}{\partial r},
\end{aligned}$$

from which it follows that

$$\frac{\partial \widehat{s}(r,x,w)}{\partial r} = \frac{1}{\sum_{j=1}^{d_x} x_j \frac{\partial \widehat{r}}{\partial x_j}(x\cdot \widehat{s}(r,x,w),w)}.$$

Using again the fact that $1/a - 1/b = -(a-b)/ab$, we have

$$\frac{\partial \widehat{s}(r,x,w)}{\partial r} - \frac{\partial s(r,x,w)}{\partial r} = \frac{-\sum_{j=1}^{d_x} x_j\left[\frac{\partial \widehat{r}}{\partial x_j}(x\cdot \widehat{s}(r,x,w),w) - \frac{\partial r}{\partial x_j}(x\cdot s(r,x,w),w)\right]}{\sum_{j=1}^{d_x} x_j \frac{\partial \widehat{r}}{\partial x_j}(x\cdot \widehat{s}(r,x,w),w)\sum_{j=1}^{d_x} x_j \frac{\partial r}{\partial x_j}(x\cdot s(r,x,w),w)} \tag{66}$$

and it suffices to ensure that the denominator of (66) is bounded away from zero with probability tending to one and that the numerator is small. We have with probability tending to one

$$\left|\frac{\partial \widehat{r}}{\partial x}(x\cdot\widehat{s}(r,x,w),w)-\frac{\partial r}{\partial x}(x\cdot s(r,x,w),w)\right|$$

$$\begin{aligned}
&= \left|\frac{\partial r}{\partial x}(x\cdot\widehat{s}(r,x,w),w)-\frac{\partial r}{\partial x}(x\cdot s(r,x,w),w)\right.\\
&\quad \left.+\frac{\partial \widehat{r}}{\partial x}(x\cdot\widehat{s}(r,x,w),w)-\frac{\partial r}{\partial x}(x\cdot\widehat{s}(r,x,w),w)\right|\\
&\leq \left|\frac{\partial^2 r}{\partial x\partial x^\top}(x\cdot\overline{s}(r,x,w),w)\times x\right|\times|\widehat{s}(r,x,w)-s(r,x,w)|\\
&\quad +\sup_{|s-s(r,x,w)|\leq\epsilon_n}\left|\frac{\partial \widehat{r}}{\partial x}(x\cdot s,w)-\frac{\partial r}{\partial x}(x\cdot s,w)\right|\\
&\leq \sup_{x\in\Psi_X}\sup_{w\in\Psi_{\widetilde{W}}}\left|\frac{\partial^2 r}{\partial x\partial x^\top}(x,w)\times x\right|\times|\widehat{s}(r,x,w)-s(r,x,w)|\\
&\quad +\sup_{x\in\Psi_X}\sup_{w\in\Psi_{\widetilde{W}}}\left|\frac{\partial \widehat{r}}{\partial x}(x,w)-\frac{\partial r}{\partial x}(x,w)\right|=o_p(\delta_n^{1/2}),
\end{aligned}$$

which gives the required bound on the numerator. Using the same result and the positivity of $\sum_{j=1}^d x_j\partial r(x\cdot s(r,x,w),w)/\partial x_j$, the denominator of (66) is bounded away from zero with probability tending to one.

■

Proof of Lemma 2. Define an empirical process $\nu_n(\cdot)$ by

$$\begin{aligned}
\nu_n(\tau) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \pi(V_i,\widetilde{W}_i,\tau)-E\pi(V_i,\widetilde{W}_i,\tau),\ \text{where} \qquad (67)\\
\pi(V_i,\widetilde{W}_i,\tau) &= \omega(V_i,\widetilde{W}_i)\cdot\tau(V_i,\widetilde{W}_i),
\end{aligned}$$

and $\tau\in\mathcal{T}$ for some pseudo-metric space $\mathcal{T}$ with pseudo-metric $\rho_{\mathcal{T}}(\cdot,\cdot)$ defined by

$$\rho_{\mathcal{T}}(\tau_1,\tau_2)=\left[\int\{\pi(v,w,\tau_1)-\pi(v,w,\tau_2)\}^2\,dvdw\right]^{1/2}. \qquad (68)$$

The function $\tau:\mathbb{R}^{d_w+1}\to\mathbb{R}$. Suppose $\widehat{\tau}$ is an estimator of $\tau_0\in\mathcal{T}$. It is well known that (see, for example, Andrews (1994, p. 2257))

$$\nu_T(\widehat{\tau})-\nu_T(\tau_0)\xrightarrow{p}0 \qquad (69)$$

if: (i) $\Pr(\widehat{\tau} \in \mathcal{T}) \to 1$, (ii) $\rho_{\mathcal{T}}(\widehat{\tau},\, \tau_0) \xrightarrow{p} 0$, and (iii) $\{\nu_T(\cdot)\ :\ T \geq 1\}$ is stochastically equicontinuous at τ_0. We take

$$\begin{aligned}\widehat{\tau}(v,w) &= [\widehat{r}(v \cdot x^*, w) - r(v \cdot x^*, w)] \\ \tau_0(v,w) &= 0\end{aligned}$$

and verify the conditions (i)–(iii) above.

We take $\mathcal{T}$ to be the class of functions with bounded Sobolev norm of order p^* where $p^* > (d_w + 1)/2$; specifically, for some large $C < \infty$, let

$$\mathcal{T} = \left\{ \tau(\cdot) : \left(\sum_{|\alpha| \leq p^*} \int \left(D^\alpha \tau(v,w)\right)^{1/2} dv dw \right)^{1/2} \leq C \right\}. \tag{70}$$

Note that $\partial\widehat{\tau}(v,w)/\partial v = \sum_{j=1}^{d_x} x_j[\partial\widehat{r}(v \cdot x^*, w)/\partial x_j - \partial r(v \cdot x^*, w)/\partial x_j]$ and likewise with the higher order partials.

Note that $\tau_0(\cdot)$ lies in $\mathcal{T}$ by assumption. To show that (i) $\Pr(\widehat{\tau} \in \mathcal{T}) \to 1$, it suffices to show that $\widehat{\tau}(v,w)$ has partial derivatives of order p^* that are bounded uniformly over the support with probability tending to one. Note that the latter holds by the uniform consistency of the derivatives of $\widehat{r}$ assumed in B4. With the pseudo-metric defined in (68), the condition (ii) $\rho_{\mathcal{T}}(\widehat{\tau}, \tau_0) \xrightarrow{p} 0$ holds trivially. Then condition (iii) is satisfied by Andrews (1994). ■

The proofs of (46) and (49) are similar to the proof of Lemma 2. Instead of (67) we have a process

$$\begin{aligned}\nu_n(\tau) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \pi(Z_i, \tau) - E\pi(Z_i, \tau), \text{ where} \\ \pi(Z_i,\, \tau) &= w_n(Z_i) \cdot \tau(Z_i),\end{aligned}$$

and in the first case take

$$\begin{aligned}w_n(Z_i) &= \frac{1}{b^{*(d_w+1)/2}} k\left(\frac{g(x) - g(X_i)}{b^*}\right) \prod_{l=1}^{d_w} k\left(\frac{w_l - W_{li}}{b^*}\right) \\ \widehat{\tau}(Z_i) &= \widehat{r}(Z_i) - r(Z_i) \\ \tau_0(Z_i) &= 0,\end{aligned}$$

while in the second case take

$$\begin{aligned}w_n(Z_i) &= \frac{-1}{b^{*(d_w+1)/2}} k\left(\frac{g(x) - g(X_i)}{b^*}\right) \prod_{l=1}^{d_w} k\left(\frac{w_l - W_{li}}{b^*}\right) \frac{\partial h}{\partial \gamma}(g(X_i), W_i) \\ \widehat{\tau}(Z_i) &= \widehat{g}(X_i) - g(X_i) \\ \tau_0(Z_i) &= 0.\end{aligned}$$

In both cases w_n has a square integrable envelope and $\widehat{\tau}, \tau_0 \in \mathcal{T}$.

References

[1] Ahn, H. (1995), "Nonparametric two-stage estimation of conditional choice probabilities in a binary choice model under uncertainty," *Journal of Econometrics*, 67, 337-378.

[2] An, M. Y. (2000), "A Semiparametric Distribution for Willingness to Pay and Statistical Inference with Dichotomous Choice CV Data," *American Journal of Agricultural Economics*, forthcoming.

[3] Andrews, D. W. K., (1991), "Asymptotic Normality of Series Estimators For Nonparametric and Semiparametric Regression Models," *Econometrica*, 59, 307-345.

[4] Andrews, D. W. K., (1994), "Empirical Process Methods in Econometrics," *The Handbook of Econometrics*, vol. IV, eds. D.F. McFadden and R.F. Engle III, 2247-2294, Amsterdam: North Holland.

[5] Andrews, D. W. K., (1995), "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–596.

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

[7] Blundell, R. and J. L. Powell (2000), "Endogeneity in Nonparametric and Semiparametric Regression Models," unpublished manuscript.

[8] Blundell, R. and J. L. Powell (2001), "Endogeneity in Semiparametric Binary Response Models," unpublished manuscript.

[9] Breiman, L. and J. H. Friedman, (1985), "Estimating Optimal Transformations For Multiple Regression and Correlation," *Journal of the American Statistical Association,* 80, 580-598.

[10] Chen, H. and A. Randall (1997): "Semi-nonparametric Estimation of Binary Response Models With an Application to Natural Resource Valuation, *Journal of Econometrics*, 76, 323-340.

[11] Creel, M., and J. Loomis (1997): "Semi-nonparametric Distribution-free Dichotomous Choice Contingent Valuation," *Journal of Environmental Economics and Management*, 32, 341-358.

[12] Chesher, A. (2001), "Quantile Driven Identification of Structural Derivatives," Unpublished manuscript.

[13] Chiang, A. C. (1984), *Fundamental Methods of Mathematical Economics.* New York: McGraw Hill.

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

[15] Friedman, J. H. and W. Stutzle, (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association,* 76, 817-823.

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

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

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

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

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

[21] 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.

[22] Hengartner, N. W. and S. Sperlich, (2002), "Rate Optimal Estimation with the Integration Method in the Presence of Many Covariates," Working Paper 01-69, Carlos III de Madrid.

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

[24] Horowitz, J., and Mammen, E. (2002), "Nonparametric Estimation of an Additive Model With A Link Function," Working paper.

[25] Imbens, G. W. and W. K. Newey, (2001), "Identification and Estimation of Triangular Simultaneous Systems Without Additivity," Unpublished manuscript.

[26] Jefferson, G., A.G.Z. Hu, X. Guan, and X. Yu, (2002), "Ownership, Performance, and Innovation in China's Large and Medium-Size Industrial Enterprise Sector," *China Economic Review*, forthcoming.

[27] Lewbel, A., O. Linton, and D. L. McFadden (2001), "Estimating Features of a Distribution From Binomial Data" Unpublished Manuscript.

[28] Linton, O. and W. Härdle (1996), "Estimating additive regression models with known links," *Biometrika* 83, 529-540.

[29] Linton, O. and J.P. Nielsen (1995), "A kernel method of estimating structured nonparametric regression based on marginal integration," *Biometrika*, 82, 93-100.

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

[31] Masry, E., (1996b), "Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and their Applications* 65, 81-101.

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

[33] 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.

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

[35] McFadden, D. L. (1999), "Computing Willingness-to-Pay in Random Utility Models," in J. Moore, R. Riezman, and J. Melvin (eds.), Trade Theory, and Econometrics: Essays in Honour of John S. Chipman, Routledge.

[36] Müller, H.G. (1989): "Adaptive nonparametric peak estimation," *The Annals of Statistics* 17, 1053-1069.

[37] Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[38] Newey, W. K., and R. L. Matzkin (1993), "Kernel Estimation of Nonparametric Limited Dependent Variable Models," unpublished manuscript.

[39] Newey, W. K., J. L. Powell, and F. Vella (1999), "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 567-603.

[40] Pinkse, J., (2001), "Nonparametric Regression Estimation Using Weak Separability," unpublished manuscript.

[41] Pötscher, B. M. and I. R. Prucha (1991), "Basic Structure of the Asymptotic Theory in Dynamic Nonlinear Econometric Models," Reprinted in *Nonlinear Models*, 1997, Bierens, Herman J., and Gallant, A. Ronald, eds., Elgar Reference Collection. International Library of Critical Writings in Econometrics, vol. 8. Cheltenham, U.K. and Lyme, N.H.: Elgar, 333-497.

[42] 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.

[43] Primont, D. and D. Primont, (1994), "Homothetic Non-parametric Production Models," *Economics Letters,* 45, 191-195.

[44] Robinson, P. M. (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

[45] Romano, J.P., (1988): "On weak convergence and optimality of kernel density estimates of the mode," *The Annals of Statistics* 16, 629-647.

[46] Shephard, R. W. (1953), *Cost and Production Functions*, Princeton: Princeton University Press.

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

[48] Stone, C.J. (1986), "The Dimensionality Reduction Principle For Generalized Additive Models," *Annals of Statistics,* 14, 590-606.

[49] Tjøstheim, D. and B. H. Auestad, (1994), "Nonparametric Identification of Nonlinear Time Series: Projections," *Journal of the American Statistical Association,* 89, 1398-1409.

[50] Tripathi, G. and W. Kim, (2001), "Nonparametric Estimation of Homogeneous Functions," unpublished manuscript.

[51] Zellner, A. and H. Ryu (1998), "Alternative Functional Forms For Production, Cost and Returns to Scale," *Journal of Applied Econometrics*, 13, 101-127.

TABLE 1: Monte Carlo fit criteria for $g(x)$ estimates

$\frac{\sigma_x}{\sigma_x+\sigma_\varepsilon}$		0.75			0.5			0.25		
n		100	300	500	100	300	500	100	300	500
IMSE	h_1	0.0494	0.0436	0.0416	0.1280	0.1006	0.0867	0.4004	0.3193	0.2797
	h_2	0.0189	0.0138	0.0122	0.0462	0.0369	0.0322	0.1622	0.1184	0.1025
	h_3	0.0172	0.0102	0.0090	0.0306	0.0225	0.0191	0.1021	0.0843	0.0661
IMAE	h_1	0.0292	0.0251	0.0246	0.0775	0.0576	0.0555	0.2402	0.1978	0.1604
	h_2	0.0116	0.0087	0.0077	0.0289	0.0224	0.0203	0.0952	0.0749	0.0661
	h_3	0.0104	0.0063	0.0056	0.0186	0.0138	0.0118	0.0638	0.0521	0.0427
PMSE	h_1	0.0487	0.0432	0.0244	0.1237	0.0855	0.0781	0.3976	0.3180	0.2204
	h_2	0.0198	0.0133	0.0109	0.0492	0.0357	0.0271	0.1910	0.1152	0.1046
	h_3	0.0183	0.0098	0.0077	0.0294	0.0213	0.0158	0.1023	0.7483	0.0592
PMAE	h_1	0.0302	0.0270	0.0173	0.0779	0.0541	0.0473	0.2238	0.1972	0.1435
	h_2	0.0116	0.0081	0.0066	0.0294	0.0215	0.0178	0.0979	0.0709	0.0632
	h_3	0.0111	0.0060	0.0049	0.0177	0.0128	0.0103	0.0638	0.0587	0.0392

TABLE 2: Monte Carlo fit criteria for $h(g)$ estimates

$\frac{\sigma_x}{\sigma_x+\sigma_\varepsilon}$		0.75			0.5			0.25		
n		100	300	500	100	300	500	100	300	500
IMSE	h_1	0.1103	0.0953	0.0918	0.2140	0.1724	0.1711	0.4911	0.3807	0.3525
	h_2	0.1284	0.0946	0.0828	0.1703	0.1279	0.1150	0.3597	0.2702	0.2481
	h_3	0.2101	0.1551	0.1357	0.2242	0.1638	0.1442	0.3378	0.2645	0.2183
IMAE	h_1	0.0765	0.0624	0.0579	0.1570	0.1218	0.1144	0.3740	0.2834	0.2581
	h_2	0.0894	0.0603	0.0511	0.1275	0.0898	0.0801	0.2771	0.1991	0.1803
	h_3	0.1552	0.1045	0.0884	0.1686	0.1150	0.0988	0.2614	0.1678	0.1625
PMSE	h_1	0.0963	0.0706	0.0537	0.1972	0.1451	0.1440	0.4527	0.3500	0.3523
	h_2	0.0908	0.0568	0.0418	0.1376	0.0913	0.0800	0.3310	0.2462	0.2146
	h_3	0.1528	0.0968	0.0760	0.1733	0.1083	0.0826	0.3124	0.2214	0.1662
PMAE	h_1	0.0678	0.0490	0.0397	0.1382	0.1057	0.0989	0.3558	0.2600	0.2597
	h_2	0.0653	0.0329	0.0251	0.1116	0.0666	0.0593	0.2607	0.1917	0.1609
	h_3	0.1088	0.0580	0.0479	0.1286	0.0708	0.0566	0.2457	0.1842	0.1201