Nonparametric Regression Estimation using Weak Separability

Joris Pinkse[0]

This version: November 2001

In this paper I propose three new estimators of nonparametric regression functions subject to *weak separability* (WS). The use of WS reduces the *curse of dimensionality*. WS nests other separability concepts such as (generalized) additive separability ((G)AS). The advantage of WS over (G)AS is that WS allows for interactions between regressors whereas (G)AS does not permit any interactions. The estimators use marginal integration and are shown to have a limiting normal distribution and a convergence rate which is the same as that of an unconstrained nonparametric estimator of a regression function of lower dimension. An attractive and unusual feature of two of my estimators is that regressors can have arbitrary convex support and that the integration regions can depend on the values of the remaining variables. The estimators can be iterated and I show that under strong assumptions further asymptotic efficiency improvements are possible. The computation of the estimators is simple. The performance of one of the estimators is studied in a simulation study.

[0] This research was financially supported by the Social Sciences and Humanities Research Council of Canada (SSHRC). I thank the co–editor, Richard Blundell, three anonymous referees, Don Andrews, Chuck Blackorby, Craig Brett, Erwin Diewert, John Galbraith, David Green, Nancy Heckman, Joel Horowitz, Guido Imbens, Oliver Linton, Rosa Matzkin, Ariel Pakes, Peter Robinson, Margaret Slade, Thanasis Stengos and seminar participants at the University of British Columbia (statistics and economics), the London School of Economics, University College London, Yale University, Royal Holloway College, Arizona State University, the universities of Groningen, Chicago, California at Berkeley, Los Angeles and Davis, Northwestern University, the MIT/Harvard econometrics workshop, the 2000 Canadian Economics Association meetings, the 2000 World Congress of the Econometric Society and the 2000 Canadian Econometrics Study Group at Guelph for useful suggestions.

1 Introduction

This paper is concerned with the estimation of the (conditional mean) regression function

$$a(x) = E(Y_1|X_1 = x), \tag{1}$$

for an i.i.d. sequence $\{(X_i, Y_i)\}$ with $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^d$. There is a plethora of estimation methods for a. Fully parametric estimation methods offer fast convergence under the assumption of a having a prespecified parametric form. At the other extreme, fully nonparametric methods allow for the estimation of a subject to minimal conditions but the precision of unconstrained nonparametric estimators deteriorates rapidly as d increases. In fact, the loss in precision relative to estimators of lower–dimensional functions increases with the sample size. This is due to the *curse of dimensionality* (Bellman, 1961, p.97, see also Fan and Gijbels, 1996, p.264).

Since many regression functions commonly used in economics feature many regressors, unconstrained nonparametric regression is often not an option. It is hence necessary to impose restrictions on a. One possibility is to allow a to be nonparametric only in a subset of the regressors and to specify a parametric form for the remaining ones. Examples are the partial linear model of Robinson (1988) and single index models.

An alternative possibility is to impose *separability* conditions. Strong or *additive separability*[1] (AS) assumes that a takes the form

$$a(x) = \sum_{j=1}^{D} g^j(x^j), \tag{2}$$

with $x^1, \ldots, x^D$ nonoverlapping subvectors of $x \in \mathbb{R}^d$ and $D \leq d$. Let x_0 denote the point at which a is to be estimated. Then AS reduces the dimensionality because for the estimation of g^j, 'close' pertains to the distance between the subvectors x_0^j and X_i^j, not between the entire vectors x_0 and X_i. The standard assumption in the literature is that the dimension of x^j, d_j, equals one for all $j = 1, \ldots, D$, i.e. the x^j are scalars. The dimensionality of the problem is then reduced from d to 1. Additively separable models can be estimated by nonparametric series regression estimation (e.g. Andrews, 1991), by *backfitting* (Friedman and Stützle, 1981, Breiman and Friedman,1985) and by *marginal integration* (MI) (Linton and Nielsen, 1995 and Tjøstheim and Auestad, 1994).

MI estimators work by integrating an unconstrained nonparametric estimator $\hat{a}$ over all dimensions not pertinent to a particular (sub)function. For instance, to estimate g^j one can integrate

[1] See in this context Stone, 1985.

$\hat{a}$ over x^{-j}, a vector whose elements are those of x except those in x^j. Although MI estimators of AS models are themselves not fully efficient, their efficiency can be improved (e.g. Fan, Härdle and Mammen, 1998). Linton (1997) has shown that using MI followed by a single backfitting step achieves full efficiency.

Generalized additive separability (GAS) (e.g. Hastie and Tibshirani, 1980 and Stone, 1986)[2] assumes that

$$a(x) = m\left(\sum_{j=1}^{D} g^j(x^j)\right). \tag{3}$$

To economists GAS is a particularly attractive alternative to AS in the context of limited dependent variable models. Consider for instance the binary choice model,

$$Y_i^* = a_*(X_i) - U_i, \quad Y_i = I(Y_i^* \geq 0), \tag{4}$$

with I the *indicator function*. Then $a(x) = m\big(a_*(x)\big)$, with m the distribution function of the error U_i and $a_*(x) = \sum_{j=1}^{D} g^j(x^j)$.

If the link function m is known then a can be estimated subject to (3) using the estimation method of Linton and Härdle (1998). Linton (2000) proposes a fully efficient estimator of an important subclass of such GAS models. Horowitz (2001) derives an MI estimator of a subject to the GAS assumption which does not assume that m is known.

A third separability assumption, *weak separability*, is used in this paper. The term 'weak separability' is due to Goldman and Uzawa (1964), although the concept was introduced by Leontief (1947).[3] My definition, as worded in definition 1, imposes some monotonicity conditions in addition to the weak separability restriction per se.

Definition 1 *A function a is weakly separable (WS) in $x^0, x^1, \ldots, x^D$ if scalar–valued functions $m, g^1, \ldots, g^D$ exist such that for all $x \in \mathbb{R}^d$,*

$$a(x) = a(x^0, x^1, \ldots, x^D) = m\big(x^0, g^1(x^1), \ldots, g^D(x^D)\big), \tag{5}$$

where m is increasing in $g^1, \ldots, g^D$ and each g^j is increasing in its first argument x_1^j. The subvectors x^j, $j = 0, \ldots, D$ cannot overlap and have dimension d_j, with $d_0 \geq 0, d_1, \ldots, d_D \geq 2, D + d_0 \geq 2$.

WS hence reduces the dimensionality of the problem to $d_m = \max(d_0 + D, d_1, \ldots, d_D)$, as opposed to 1 if a is additive in individual regressors. Therefore the dimensionality of the problem

[2] Generalized additive separability has a much longer history in economics, albeit without the qualifier 'generalized'.

[3] Another early reference is Strotz (1957).

increases no slower than $\sqrt{d}$. It is possible to construct a nested version of (5) which reduces the dimensionality to 2, regardless of d, but the estimators in this paper are only appropriate for functions a that satisfy (5). The requirement that the groups be nonoverlapping is restrictive. A still weaker forms of separability, which has not been used in this context, is *latent separability* (Blundell and Robin, 2000).

Under WS the role of regressors depends on the group to which they belong, whereas under (G)AS they are treated symmetrically. In order to exploit the benefits of WS fully, one hence needs to have some prior information to create a reasonable grouping of regressors.

The most obvious limitation of (G)AS is that it does not allow for interactions between regressors. WS does allow for such interactions, albeit subject to restrictions.[4] Interactions between regressor variables can be important as the following four examples demonstrate. The first example relates to returns to education. In the most narrow returns to education model (Mincer, 1974, chapter 2), the difference in expected log earnings between two individuals with the same level of experience x^0 depends only on differences in characteristics $z = [x^1, \ldots, x^D]^T$ such as schooling and demographics, but not on the experience level itself.[5] Then returns to education would be additively separable in x_0, z, i.e. $a(x) = m(x^0) + g(z)$. With AS, g itself must moreover be additively separable in the various schooling and demographic variables. For GAS, additive separability must hold for a transformation of expected log earnings. However, empirical research (e.g. Lazear 1977) has shown that the way that earnings vary with experience differs across educational backgrounds. WS allows x^0 to interact freely with one or more indices of the schooling and demographic variables.

Now consider the case in which a is a multiproduct cost function. There are *economies of scope* (Baumol et al., 1982) if the cost of producing multiple products is less than the sum of the production cost of each individual good. AS allows for no economies of scope and (G)AS only for very specific ones.

Another potential application is that of *hedonic pricing* (Court, 1939) models. Cannaday (1994) explains the prices of Chicago appartments in terms of a number of appartment characteristics including the living area, number of bedrooms and bathrooms, amenities, location, view and restrictions on pet ownership. Although Cannaday uses a regular linear regression model, which is trivially additively separable, there are arguments for a WS structure. Under AS the value of an extra bathroom is independent of the number of bedrooms. Moreover, the value of bedrooms and

[4]See Blackorby et al. (1991) for a lucid discussion of the various forms of separability.

[5]I thank David Green for this example.

bathrooms is under AS assumed independent of the location of the appartment and its size. WS assumes the existence of indices — perhaps ones for size, amenities, location and regulations — which can freely interact with one another.

Such limitations are not restricted to models with continuous dependent variables. In binary choice models, for instance, the same issues arise. If the dependent variable is the mode of transport chosen to get to work (as in e.g. Train, 1980, and Horowitz, 1993) then the regressors can include variables relating to the time and inconvenience a particular mode of transport entails, its cost, the availability of autos and the number of drivers in a household, and the respondent's household income.[6] If, for instance, the disutility of changing buses twice is the same if the total travel time is 10 minutes as when it is an hour, then the variables representing the number of bus changes and total travel time can be additively separated from one another. If not, then a_* in (4) cannot be AS and a is then not GAS. Moreover, even if a_* is AS, a is only GAS if the U_i's are homoskedastic. WS, on the other hand, does allow for some restrictive forms of heteroskedasticity.

It is possible to introduce some interaction into (G)AS models by allowing the x^j to be vector–valued. Variables that one wants to interact can then be gathered in the same x^j vector. However, most theoretical results do not support vector–valued x^j. Moreover, the dimensionality of the problem is then the same as that of the maximum of the dimensions of the x^j vectors, which may not be less than the dimensionality under WS. Finally, under the vector form of (G)AS no interactions would be allowed between elements in different x^j–vectors, which may necessitate the use of bigger groups than in the case of WS and hence a smaller degree of dimension–reduction than could be achievable under WS.

The choice of the type of separability to impose, if any, ultimately involves a trade–off between bias due to misspecification and a greater variance because of the greater dimensionality. On this scale WS is located somewhere between GAS and an unconstrained estimator. Since (2) and (3) imply (5), the estimators proposed in this paper will consistently estimate any regression functions a which satisfy (2) or (3). However, if the model is truly (G)AS then estimators that are specifically designed to estimate (G)AS models are likely to be more accurate.

The proposed estimation methods use marginal integration. Unlike most estimators in this literature two of the three estimators I propose allow for arbitrary, possibly infinite, convex support. General convex support could be useful because in many instances particular combinations of re-

[6]Train studies a multinomial choice problem rather than a binary one. Horowitz uses a semiparametric estimator of a single index model with a more limited set of variables.

gressor values cannot occur. Indeed, I doubt that there are any 300 square foot appartments with 4 bedrooms. My approach uses *comparison functions* which allow the integration regions to depend on the values of variables which are not integrated over. Instead of using my approach, one may be able to circumvent the support problem by trimming out observations to make the support of X_1 the Carthesian product of the supports of elements in the X_1–vector. However, unless the support of X_1 is close to a hypercube, such a procedure can be very inefficient.

Like other work in this area I assume that the regressors are continuous. In most economic applications continuity of all regressors is not a reasonable assumption. Since my estimator does not use derivative estimators, it is probably possible to extend the results to allow for discrete regressors (see Delgado and Mora, 1995). Equally problematic for many empirical economic applications is that none of these methods (mine included) allow for endogenous regressors. Allowing for endogenous regressors in nonparametric models is difficult and can only be achieved under strong assumptions (see e.g. Florens and Renault, 2000, Newey and Powell, 1990, Newey, Powell and Vella, 1997, Pinkse and Ng, 1998, and Pinkse, 2000).

As mentioned earlier, with (G)AS it is possible to improve asymptotic efficiency by the use of a multi–stage procedure (Linton 1997, 2000). Indeed, it is then possible to achieve, what Linton calls, 'full oracle efficiency', i.e. asymptotic efficiency is as good as if the remaining components were observed. Under strong conditions a similar result applies under some circumstances for one of the WS estimators proposed here. The procedure involves combining iterated and noniterated WS estimators.

There are other uses for iteration. Because of the separability constraints, a can be estimated consistently in some regions outside of the support of X_1 provided that separability holds globally. It may hence be possible to use estimators that require strong support restrictions in a second step after a is first estimated using my method, provided that a is known to satisfy WS.

The outline of this paper is as follows. Section 2 introduces the three new estimators. The main result is contained in section 3, which is followed by a discussion of ways to recover the structural components of a and the benefits of iterating the procedure. A simulation study of the properties of the estimators is in section 5. The appendix contains all proofs and derivations.

2 Estimation Methods

2.1 $d_0 \geq 1$, $D = 1$

I propose three closely related estimation methods which I for ease of exposition introduce for the case $d_0 \geq 1, D = 1$. The formal results are in section 3 and apply to any regression function satisfying definition 1. Thus,

$$a(x) = m\big(x^0, g^1(x^1)\big). \tag{6}$$

Recall from definition 1 that a is monotonic in g^1 and g^1 is monotonic in x_1^1. Then for any positive function λ^* for which the integral exists,

$$\underline{g}^1(x^1) = \int a(x)\lambda^*(x^0)dx^0 = \int m\big(x^0, \upsilon^*\big(g^1(x^1)\big)dx^0,$$

for some monotonic function υ^*. Since m and g^1 cannot be separately identified from (6), one needs to impose identifying restrictions in order to estimate g^1, m individually, but not in order to estimate a. Separate identification of m, g^1 can be achieved by fixing λ^*. This issue is discussed in section 4.1. Here the focus is on the estimation of a at some point $x_0 = (x_0^0, x_0^1)$.

Note that $a(x) = a(x_0)$ whenever $x^0 = x_0^0$ and $g^1(x^1) = g^1(x_0^1)$ or equivalently when $x^0 = x_0^0$ and $\underline{g}^1(x^1) - \underline{g}^1(x_0^1) = 0$. A possible estimator could then use a comparison function $\underline{\chi}^1$ like

$$\underline{\chi}^1(x^1) = \underline{\chi}^1(x^1, x_0^1) = \underline{g}^1(x_0^1) - \underline{g}^1(x^1) = \int \big(a(x^0, x_0^1) - a(x)\big)\lambda^*(x^0)dx^0. \tag{7}$$

$\underline{\chi}^1$ can be estimated by replacing a in (7) with a fully unconstrained *Nadaraya–Watson* (NW) (Nadaraya, 1965, and Watson, 1965) kernel regression estimator. The NW estimator requires the practitioner to choose a *kernel* k, i.e. an even function which integrates to one, and a *bandwidth* h, whose choice depends on the sample size.[7] If the argument of k is a vector ξ then $k(\xi) = \prod_{j=1}^{d_\xi} k(\xi_j)$, where d_ξ denotes the dimension of ξ. Let moreover $K(t) = k(t/h)/h$ and $K_i(x) = K(x - X_i)$. Then the NW estimator is

$$\hat{a}(x) = \frac{\sum_{i=1}^n K_i(x)Y_i}{\sum_{i=1}^n K_i(x)}, \tag{8}$$

with n the number of observations. $\underline{\chi}^1$ can then be estimated by

$$\hat{\underline{\chi}}^1(x^1) = \int \big(\hat{a}(x^0, x_0^1) - \hat{a}(x)\big)\lambda^*(x^0)dx^0, \tag{9}$$

[7] For ease of notation the bandwidth is chosen to be the same in every dimension. This is not necessary.

Because $\hat{a}$ is integrated in several dimensions in (9), the dimensionality of the problem of estimating $\underline{\chi}^1$ is d_1 as opposed to the dimensionality of the problem of estimating a, which is d. As a result, $\underline{\hat{\chi}}^1$ converges faster than does $\hat{a}$. In fact, the convergence rate of $\underline{\hat{\chi}}^1$ is generally the same as that of a Nadaraya–Watson estimator of dimension d_1 using the same input parameters k, h. It is then possible to construct an estimator of $a(x_0)$ that is more efficient than $\hat{a}(x_0)$, namely

$$\hat{a}^*_{S\chi}(x_0) = \frac{\sum_{i=1}^n K_0(X_i^0) K\big(\underline{\hat{\chi}}^1(X_i^1)\big)\Lambda_i Y_i}{\sum_{i=1}^n K_0(X_i^0) K\big(\underline{\hat{\chi}}^1(X_i^1)\big)\Lambda_i}, \tag{10}$$

with $K_0(X_i^0) = K(x_0^0 - X_i^0)$ and where the Λ_i–factors allow the practitioner to trim out or give less weight to particular observations. If the $\underline{\chi}^1$–function were known, $\hat{a}^*_{S\chi}$ would simply be a Nadaraya–Watson style estimator of a (d_0+1)–dimensional regression function. Here the convergence rate of $\hat{a}^*_{S\chi}$ is that of a Nadaraya–Watson estimator of a $\max(d_0+1, d_1)$–dimensional function using the same input parameters. Since $\max(d_0+1, d_1) < d_0 + d_1 = d$ a dimension reduction is achieved. $\hat{a}^*_{S\chi}$ uses *generated regressors*. Other examples of the use of generated regressors in a nonparametric context are Ahn (1997), Horowitz (2001) and Rilstone (1996).

The estimator $\hat{a}^*_{S\chi}$ is straightforward to implement, although it does require numerical integration. However, $\hat{a}^*_{S\chi}$ has two problems. It does not allow for an infinite integration region (i.e. λ^* must be zero outside some bounded set) and, more importantly, it assumes that the integration region is independent of the variables that are not integrated out. The problem with a constant integration region is that $\hat{a}$ does not estimate a consistently outside of the support of X_1. If the support of X_1 is not the product of the supports of X_1^0 and X_1^1, then the integration region has to be limited to values x^0 for which $x = (x^0, x^1)$ is in the support of X_1 for all x^1 in the support of X_1^1.

The convex support problem can be fixed by using a weight function $\tilde{\lambda}$ (in lieu of λ^* in (7)) which can depend on both x^1 and x_0^1, i.e.

$$\begin{aligned}
\chi^1(x^1) &= \int \big(a(x^0, x_0^1) - a(x)\big)\tilde{\lambda}^1(x^0, x^1)\tilde{\lambda}^1(x^0, x_0^1)dx^0, \\
\hat{\chi}^1(x^1) &= \int \big(\hat{a}(x^0, x_0^1) - \hat{a}(x)\big)\tilde{\lambda}^1(x^0, x^1)\tilde{\lambda}^1(x^0, x_0^1)dx^0.
\end{aligned} \tag{11}$$

With (11), a practitioner can choose $\tilde{\lambda}^1$ in such a way that the integration region for each (x^1, x_0^1)–pair is the set of values x^0 for which both (x^0, x^1) and (x^0, x_0^1) belong to the support of X_1. I have however been unable to show that my results hold for

$$\hat{a}_{S\chi} = \frac{\sum_{i=1}^n K_0(X_i^0) K\big(\hat{\chi}^1(X_i^1)\big)\Lambda_i Y_i}{\sum_{i=1}^n K_0(X_i^0) K\big(\hat{\chi}^1(X_i^1)\big)\Lambda_i}, \tag{12}$$

when the integration region in (11) is infinite.[8] Moreover, even when $\tilde{\lambda}^1$ is positive only on a bounded set, the estimator $\hat{a}_{S\gamma}$ proposed below has the same asymptotic properties as $\hat{a}_{S\chi}$, albeit for a different choice of λ–function. I nevertheless provide theoretical results for $\hat{a}_{S\chi}$, also, since it allows for individual estimation of the g^1 and m–functions subject to identification restrictions, unlike the two estimators proposed below.

I now propose two estimators which do allow for infinite integration regions. They use comparison functions γ^1, π^1 defined by

$$\gamma^1(x^1) = \int \big(a(x^0, x_0^1) - a(x^0, x^1)\big) f(x^0, x^1) f(x^0, x_0^1) \lambda^1(x^0, x^1) \lambda^1(x^0, x_0^1) dx^0,$$
$$\pi^1(x^1) = \gamma^1(x^1)/\delta^1(x^1), \text{ with } \delta^1(x^1) = \int f(x^0, x^1) f(x^0, x_0^1) \lambda^1(x^0, x^1) \lambda^1(x^0, x_0^1) dx^0,$$

where f denotes the density of X_1. γ^1 and π^1 can be estimated by

$$\hat{\gamma}^1(x^1) = \int \big(\hat{a}(x^0, x_0^1) - \hat{a}(x)\big) \hat{f}(x^0, x^1) \hat{f}(x^0, x_0^1) \lambda^1(x^0, x^1) \lambda^1(x^0, x_0^1) dx^0, \tag{13}$$
$$\hat{\pi}^1(x^1) = \hat{\gamma}^1(x^1)/\hat{\delta}^1(x^1), \text{ with } \hat{\delta}^1(x^1) = \int \hat{f}(x^0, x^1) \hat{f}(x^0, x_0^1) \lambda^1(x^0, x^1) \lambda^1(x^0, x_0^1) dx^0. \tag{14}$$

Unlike $\hat{\gamma}^1$, $\hat{\pi}^1$ corrects for the fact that integration regions are different for different (x_0^1, x^1)–values. This is important because γ^1 can be close to zero, not only because $a(x^0, x^1)$ is close to $a(x^0, x_0^1)$ but also when $\delta^1(x^1)$ is small. The second step estimator will then give too much weight to observations in the tails of the distribution.

$\hat{\pi}^1$ corrects this problem, but has the problem that it may have a greater (small sample) variance than does $\hat{\gamma}^1$ because of the denominator term. $\hat{a}_{S\gamma}, \hat{a}_{S\pi}$ are identical to $\hat{a}_{S\chi}$ defined in (12) except that $\hat{\chi}^1$ is to be replaced with $\hat{\gamma}^1, \hat{\pi}^1$.

A related problem with all three estimation procedures a is that for some values of x^1, x_0^1, δ^1 is very close to or equals zero. This occurs for instance when the support of X_1^0 and the first element of X_1^1 is a diagonal strip and x_1^1 and x_{01}^1 are far apart. Using observations i for which $\delta^1(X_i^1)$ is close to zero can make the second stage estimator unstable. I therefore use the Λ_i's to trim out such observations in the second stage.

2.2 General Case

The estimators for the general case are similar to that for the case $d_0 \geq 1, D = 1$, but are notationally more cumbersome. Let as before x^j denote the j–th subvector of x and recall that x^{-j} denotes the

[8] This is because showing uniform consistency of $\hat{a}$ on an infinite support is difficult under general conditions.

vector x without x^j. Further, $z = x^{-0}$, $z^j = x^j$ and z^{-j} is the z–vector without x^j. For any function $\omega : \mathbb{R}^d \to \mathbb{R}$, ω^j is the same as ω, except that the arguments are rearranged, i.e.

$$\forall x \in \mathbb{R}^d : \omega^j(x^{-j}, x^j) = \omega(x).$$

It is possible both to use the same function λ throughout and to use different functions for different values of j. From hereon I will assume there is one function λ, but all results go through identically when different λ–functions are used.

Let $\gamma^j(x^j) = \int \big(a^j(x^{-j}, x_0^j) - a^j(x^{-j}, x^j)\big) f^j(x^{-j}, x_0^j) f^j(x^{-j}, x^j) \lambda^j(x^{-j}, x_0^j) \lambda^j(x^{-j}, x^j) dx^{-j}$ and let δ^j, π^j be similar generalizations of δ^1, π^1.

Equations (11), (13) and (14) generalize to

$$\begin{aligned}
\hat{\chi}^j(x^j) &= \int \big(\hat{a}^j(x^{-j}, x_0^j) - \hat{a}^j(x^{-j}, x^j)\big) \tilde{\lambda}^j(x^{-j}, x^j) \tilde{\lambda}^j(x^{-j}, x_0^j) dx^{-j}, \text{ setting } \lambda = \tilde{\lambda}/f, \\
&= \int \big(\hat{a}^j(x^{-j}, x_0^j) - \hat{a}^j(x^{-j}, x^j)\big) f(x^{-j}, x^j) f^j(x^{-j}, x_0^j) \lambda^j(x^{-j}, x^j) \lambda^j(x^{-j}, x_0^j) dx^{-j} \quad (15) \\
\hat{\gamma}^j(x^j) &= \int \big(\hat{a}^j(x^{-j}, x_0^j) - \hat{a}^j(x^{-j}, x^j)\big) \hat{f}^j(x^{-j}, x^j) \hat{f}^j(x^{-j}, x_0^j) \lambda^j(x^{-j}, x^j) \lambda^j(x^{-j}, x_0^j) dx^{-j}, \quad (16) \\
\hat{\pi}^j(x^j) &= \hat{\gamma}^j(x^j)/\hat{\delta}^j(x^j), \\
&\quad \text{with } \hat{\delta}^j(x^j) = \int \hat{f}^j(x^{-j}, x_0^j) \hat{f}^j(x^{-j}, x^j) \lambda^j(x^{-j}, x^j) \lambda^j(x^{-j}, x_0^j) dx^{-j}. \quad (17)
\end{aligned}$$

The estimators of the WS function a are then

$$\hat{a}_{S\chi} = \frac{\sum_{i=1}^n K_0(X_i^0, \hat{\chi}_i) \Lambda_i Y_i}{\sum_{i=1}^n K_0(X_i^0, \hat{\chi}_i) \Lambda_i}, \quad \hat{a}_{S\gamma} = \frac{\sum_{i=1}^n K_0(X_i^0, \hat{\gamma}_i) \Lambda_i Y_i}{\sum_{i=1}^n K_0(X_i^0, \hat{\gamma}_i) \Lambda_i}, \quad \hat{a}_{S\pi} = \frac{\sum_{i=1}^n K_0(X_i^0, \hat{\pi}_i) \Lambda_i Y_i}{\sum_{i=1}^n K_0(X_i^0, \hat{\pi}_i) \Lambda_i}, \quad (18)$$

with $\hat{\chi}_i = [\hat{\chi}_i^1, \ldots, \hat{\chi}_i^D]^T$, $\hat{\gamma}_i = [\hat{\gamma}_i^1, \ldots, \hat{\gamma}_i^D]^T$, $\hat{\gamma}_i^j = \hat{\gamma}^j(X_i^j)$, $\hat{\pi}_i = [\hat{\pi}_i^1, \ldots, \hat{\pi}_i^D]^T$, $\hat{\pi}_i^j = \hat{\pi}^j(X_i^j)$, $\Lambda_i = \prod_{j=1}^D \Lambda_i^j$ and $\Lambda_i^j = \Lambda^j(X_i^j)$.

3 Main Results

I now state my assumptions for asymptotic normality of the estimators $\hat{a}_{S\chi}, \hat{a}_{S\gamma}, \hat{a}_{S\pi}$ defined in (18). Partition the vectors $x^j = [x^{j1}, (x^{j2})^T]^T$, with x^{j1} a scalar and x^{j2} possibly vector–valued.

Assumption A $\{(Y_i, X_i)\}$ *is an i.i.d. sequence with for some* $\epsilon_\mu > 0$, $E|Y_1|^{4+\epsilon_\mu} < \infty$. *The distribution of* X_1 *is absolutely continuous. The conditional mean function* a, *defined in (1), satisfies definition 1 and is increasing in* $x^{11}, \ldots, x^{D1}$.

The existence of moments greater than four is strong but not unusual. Let $\mathcal{S} = \prod_{j=1}^D \mathcal{S}^j$ and let $\mathcal{S}^0 \subset \mathbb{R}^{d_0}$.

Assumption B *x_0 is an element of $\mathcal{S}^* = \mathcal{S}^0 \times \mathcal{S}$ with $\mathcal{S}^j$, $j = 0, \ldots, D$ open, convex and bounded. For any $j = 1, \ldots, D$ and some practitioner–chosen nonnegative function λ,*
$\delta^j(x^j) = 0 \Rightarrow x^j \notin \bar{\mathcal{S}}^j$, with $\bar{\mathcal{S}}^j$ the closure of $\mathcal{S}^j$.

The support restriction in assumption B is weak. It requires that the probability that $\big(X_1^0, g(Z_1)\big)$ lies in a small neighborhood of $\big(x_0^0, g(z_0)\big)$ is positive. If this restriction were not satisfied, it would also not be possible to use the unconstrained estimator. Note that $\mathcal{S}^*$ is *not* the support of X_1. Indeed, the support of X_1 can be $\mathbb{R}^d$ and the regions of integration in the first step can be infinite, also.

Assumption C *$\Lambda : \mathbb{R}^{d-d_0} \to \mathbb{R}$ is nonnegative and positive only on $\mathcal{S}$.*

Definition 2 *$\mathcal{W}_{d_\xi, r}$ is the class of functions $\omega : \mathbb{R}^{d_\xi} \to \mathbb{R}$ for which ω's r–th partial derivatives exist and are bounded and continuous.*

Let $\nu = af$.

Assumption D *For some $r \geq 2$, f, ν, Λ are boundedly integrable, λ is bounded and*

$$\frac{\partial^2 \omega}{\partial (x^{j1})^2} \in \mathcal{W}_{d,r}, \quad \omega = f, \nu, \Lambda, \lambda;\ j = 1, \ldots, D.$$

The variance function $\sigma^2(x) = V(Y_1 | X_1 = x)$ is once partially differentiable.

In most cases, the existence of four derivatives in most directions and six in some is sufficient for $\hat{a}_{S\bullet}$, $\bullet = \gamma, \pi, \chi$, to have the same rate of convergence as an unconstrained nonparametric kernel estimator of a regression function with d_m regressors using the same choice of kernel. When fewer derivatives exist, then $\hat{a}_{S\bullet}$ usually still converges faster than the fully unconstrained estimator $\hat{a}$, but the degree of dimension reduction attainable is then less.

It is possible to choose a function Λ that is many times differentiable and positive only on an open convex set. An example is the function $\Lambda(t) = \bar{\Phi}\big(1/(1 - ||t||^2)\big) I(||t|| < 1)$ with I the indicator function and $\bar{\Phi}$ the standard normal distribution function.

In section 2 I mentioned that for the results relating to $\hat{a}_{S\chi}$ it is necessary to restrict λ. This condition is expressed in assumption E, which does not apply to $\hat{a}_{S\gamma}, \hat{a}_{S\pi}$.

Assumption E *λ is positive only on a bounded set, on which f is bounded away from zero.*

Assumption F *The kernel k is a product kernel, i.e., $k(\xi) = \prod_{t=1}^{d_\xi} k(\xi_t)$ for some $r+2$ times differentiable r–th order kernel k with exponentially decreasing tails, i.e. $\int k(t)dt = 1$, $\int k(t)t^s dt = 0$, $s = 1, \dots, r-1$, and $\int |k^{(u)}(t)t^s|dt < \infty$ for any $0 \leq s < \infty, u = 0, \dots, r+2$.*

Assumption F imposes strong conditions on the choice of kernel. Since the kernel is chosen by the practitioner, strong conditions on its choice do not limit the range of potential applications. For any r, there are kernels that satisfy assumption F. In particular, for $r = 4$, the scalar–argument kernel

$$k(t) = (3 - t^2)\phi(t)/2, \tag{19}$$

with ϕ the standard normal density, satisfies assumption F.[9]
Let for some $\varepsilon > 0$,

$$c_h = \max\big(2d - \min(d_1, \dots, d_D), d_m + 4\big) + \varepsilon.$$

Assumption G

$$n^{-1}h^{-c_h} = o(1), \quad nh^{2r+d_m} = O(1). \tag{20}$$

Assumption G is easy to satisfy. To obtain convergence at the optimal rate, one should choose the bandwidth $h \sim n^{-1/(2r+d_m)}$ such that the conditions in (20) hold when $r > (c_h - d_m)/2$. For instance, when $d = 5, D = 2, d_0 = 1, d_1 = d_2 = 2$, $d_m = 3$ and $(c_h - d_m)/2 = (8 + \varepsilon - 3)/2 = (5 + \varepsilon)/2$. Since ε can be chosen arbitrarily close to zero, $r = 4$ suffices.

The same bandwidth is used in both stages of the estimation procedure. This is not essential and indeed generally not advisable.

I now proceed with the statement of the main result. Let q be one of γ, π, χ and let $f_\bullet^+$ denote the joint density of $\big(X_1^0, q(Z_1)\big)$ conditional on $\Lambda_1 > 0$. Let for any function ω

$$\Delta_\bullet(\omega) = E\big(\omega(X_1)\Lambda(X_1)|X_1^0 = x_0^0, g(Z_1) = g(z_0), \Lambda(X_1) > 0\big) f_\bullet^+(x_0^0, 0)p_0, \tag{21}$$

and $\Delta_\bullet = \Delta_\bullet(1)$, with $p_0 = P(\Lambda_1 > 0)$. Further, let for any function $\omega : \mathbb{R}^d \to \mathbb{R}$,

$$\Delta_{\bullet\omega}^j(x) = \Delta_\bullet \left(\frac{\frac{\partial a}{\partial x^{j1}}}{\frac{\partial q^j}{\partial x^{j1}}} (\omega^j \cdot \lambda^j)(x^{-j}, \cdot)\Psi_\bullet^j \right) / \Delta_\bullet, \tag{22}$$

[9] There is evidence that higher order kernels do not work well in samples of moderate size. An alternative technique, which may be preferable in practice, is that of local polynomial estimation.

where $\Psi^j_\chi(x^j) = \Psi^j_\gamma(x^j) = 1$, $\Psi^j_\pi(x^j) = 1/\delta^j(x^j)$ and $(\omega \cdot \lambda)(x) = \omega(x)\lambda(x)$. Let further $\Delta^j_{\bullet\omega t} = \Delta^j_{\bullet\omega}(X_t)$. Finally, let $\rho_n = n^{-1/2}h^{-d_m/2}$, $\ell = \lim_{n\to\infty} \rho_n^{-1}h^r$ and let $\mathcal{B}_\gamma, \mathcal{B}_\pi, \mathcal{B}_\chi$ be some finite numbers.

Theorem 1 *Let assumptions A–D and F–G hold. Then*

$$\rho_n^{-1}\big(\hat{a}_{S\gamma}(x_0) - a(x_0)\big) \xrightarrow{\mathcal{L}} N\big(\ell\mathcal{B}_\gamma, \mathcal{V}_\gamma\big), \quad \rho_n^{-1}\big(\hat{a}_{S\pi}(x_0) - a(x_0)\big) \xrightarrow{\mathcal{L}} N\big(\ell\mathcal{B}_\pi, \mathcal{V}_\pi\big), \tag{23}$$

and for $\bullet = \gamma, \pi$,

$$\mathcal{V}_\bullet = \kappa^{d_m}\left(I(d_0 + D = d_m)\mathcal{V}^0_\bullet + \sum_{j=1}^{D} I(d_j = d_m)\mathcal{V}^j_\bullet\right), \tag{24}$$

with $\kappa = \int k^2(t)dt$, *and*

$$\mathcal{V}^0_\bullet = \Delta_\bullet(\sigma^2\Lambda)/\Delta^2_\bullet(1), \quad \mathcal{V}^j_\bullet = f_{x^j}(x^j_0)E\Big(\big(\sigma_1\lambda_1\Delta^j_{\bullet f1}\big)^2|X^j_1 = x^j_0\Big), \tag{25}$$

with f_{x^j} *the density of* X^j_1 *and* $\sigma^2_1 = \sigma^2(X_1)$ *where* $\sigma^2(x) = V(Y_1|X_1 = x)$.

If assumptions A–G hold then moreover

$$\rho_n^{-1}\big(\hat{a}_{S\chi}(x_0) - a(x_0)\big) \xrightarrow{\mathcal{L}} N\big(\ell\mathcal{B}_\chi, \mathcal{V}_\chi\big), \tag{26}$$

with $\mathcal{V}_\chi = \mathcal{V}_\gamma$.

To achieve the optimal rate of convergence, $h \sim n^{-1/(2r+d_m)}$, which results in a convergence rate of $n^{-r/(2r+d_m)}$. If the bandwidth sequence is thus chosen, the asymptotic bias is nonzero. The asymptotic bias can be removed by undersmoothing, i.e. choosing a bandwidth sequence which goes to zero at a rate faster than the optimal rate. Note also that if $d_0 + D > \max_{j=1,\dots,D} d_j$, then the $\hat{a}_{S\bullet}$–estimators have the same asymptotic distribution as the NW estimator with known $g^1, \dots, g^D$. In other cases the $\mathcal{V}^j_\bullet$–terms contribute to the asymptotic variance.

The asymptotic variance $\mathcal{V}_\bullet$ is estimable, as the following theorem shows. Let

$$\hat{\Delta}_\bullet(\omega) = n^{-1}\sum_{i=1}^{n} K_0(X_i, \hat{q}_i)\Lambda_i\omega_i, \tag{27}$$

and

$$\hat{P}^j_t(x) = K_0(X^0_t, \hat{q}^{-j}_t)K_0'(\hat{q}^j_t)\Lambda_t\lambda^j(x^{-j}, X^j_t). \tag{28}$$

$$\hat{R}^j_{\omega t}(x) = \hat{P}^j_t(x)\hat{\Psi}^j_{\bullet t}\frac{Y_t - \hat{a}(x_0)}{\hat{\Delta}_\bullet}\hat{\omega}^j(x^{-j}, X^j_t), \quad \text{for } \omega = \nu, f, \tag{29}$$

where $\hat{\Psi}^j_{\bullet t} = \hat{\Psi}^j_\bullet(X^j_t)$, with $\hat{\Psi}^j_\chi(x^j) = \hat{\Psi}^j_\gamma(x^j) = 1$ and $\hat{\Psi}^j_\pi(x^j) = 1/\hat{\delta}^j(x^j)$, and $\hat{\nu}$ is the numerator in (8) divided by n.

Theorem 2 *Let assumptions A–D, F–G hold.* $\mathcal{V}_\bullet$ *in (24) is consistently estimated by*

$$\hat{\mathcal{V}}_\bullet = \kappa^{d_m}\left(I(d_0+D=d_m)\hat{\mathcal{V}}^0_\bullet + \sum_{j=1}^{D} I(d_j=d_m)\hat{\mathcal{V}}^j_\bullet\right), \tag{30}$$

where $\hat{\mathcal{V}}^0_\bullet$ *and* $\hat{\mathcal{V}}^j_\bullet$ *are estimators of* $\mathcal{V}^0_\bullet$ *and* $\mathcal{V}^j_\bullet$ *defined in (25) and are given by*

$$\hat{\mathcal{V}}^0_\bullet = n^{-1}\sum_{i=1}^{n} K_0(X^0_i,\hat{q}_i)\Lambda_i^2\big(Y_i-\hat{a}(x_0)\big)^2/\hat{\Delta}^2_\bullet(1), \tag{31}$$

and

$$\hat{\mathcal{V}}^j_\bullet = \kappa^{-d_j}h^{d_j}n^{-1}\sum_{i=1}^{n}\left(n^{-1}\sum_{t=1}^{n}\big(R^j_{ft}(X_i)Y_i - R^j_{\nu t}(X_i)\big)\right)^2 \tag{32}$$

If in addition assumption E holds, then $\mathcal{V}_\chi$ *is consistently estimated even when* λ^j *in (32) and (28) is replaced with* $\tilde{\lambda}^j/\hat{f}^j$.

4 Further Issues

4.1 Separate Identification of m and $g^1,\ldots,g^D$

It can be of interest to obtain separate estimates of m and $g^1,\ldots,g^D$. In section 2 I mentioned that it was possible to achieve such identification by fixing the choice of λ^* when using $\hat{a}^*_{S\chi}$. Doing so will result in estimators of g^j, m which converge at rates $O_p(n^{-1/2}h^{-d_j/2})$ and $O_p(n^{-1/2}h^{-(d_0+D)/2})$. However, achieving identification by choice of an input parameter may be undesirable and can be avoided. Here I propose two identification conditions which neither depend on the choice of an input parameter nor on the distribution of the random variables. The two identification conditions are motivated by the estimation of a cost function and can be replaced by similar conditions if such alternative conditions are deemed more appropriate for a particular application.

The cost of producing x^0 units of output when the vector of input prices is x^1 is $m\big(x^0,g^1(x^1)\big)$, where g^1 is the unit cost function. A natural identification condition, therefore, is

$$\forall g^1 : m(1,g^1) = g^1, \tag{33}$$

i.e. the production of one unit of output costs g^1. Since $a(1,x^1) = m\big(1,g^1(x^1)\big) = g^1(x^1)$, g^1 can be estimated by

$$\hat{g}^1(x^1) = \hat{a}_{S\bullet}(1,x^1).$$

An estimate of m, $\hat{m}$, can then be obtained by nonparametrically regressing Y_i on $X_i^0, \hat{g}^1(X_i^1)$. It can be shown that $\hat{m}$ converges at a rate of $O_p(n^{-1/2}h^{-(d_0+D)/2})$.[10] Irrespective of the dimensions, $\hat{m}$ hence converges at the same rate as when the g–functions are known. The convergence rate of $\hat{g}^1$ is however slower than optimal if $d_1 < d_m$.

An alternative identification condition is

$$\forall x^{11} : g^1(x^{11}, 1) = x^{11}. \tag{34}$$

Condition (34) does not afford a straightforward interpretation in the context of cost functions, but could be replaced with $g^1(t, \dots, t) = t$, which is implied by $g(1, \dots, 1) = 1$ and homogeneity of the unit cost function.[11] Two implications of (34) are that $m(x^0, g^1) = a\big(x^0, (g^1, 1)\big)$ and that for any function $\underline{g}^1$ which is a monotonic transformation of g^1,

$$g^1(x^1) = \underline{g}_{(1)}^{-1}\big(\underline{g}^1(x^1)\big),$$

with $\underline{g}_{(1)}^{-1}$ the inverse of the function $\underline{g}_{(1)}(x^{11}) = \underline{g}^1(x^{11}, 1)$. m can hence be estimated by

$$\hat{m}(x^0, g^1) = \hat{a}_{S\bullet}\big(x^0, (g^1, 1)\big).$$

If the integration region is finite and independent of the values of the variables which are not integrated over, then a monotonic transformation of g^1 is $\underline{g}^1(x^1) = \int a(x^0, x^1)\lambda^*(x^0)dx^0$, as we saw in section 2.[12] $\underline{g}^1$ can then be consistently estimated by (see Pinkse, 1999)

$$\hat{\underline{g}}^1(x^1) = \int \hat{a}(x^0, x^1)\lambda^*(x^0)dx^0. \tag{35}$$

Let $\hat{\underline{g}}_{(1)}(x^{11}) = \hat{\underline{g}}^1(x^{11}, 1)$. Assume without loss of generality that $g^1(x^1) \in [0, 1]$ and that $\underline{g}_{(1)}$ is increasing in a neighborhood of $[0, 1]$. Let $\underline{g}_{(1)}^{-1}$ be the inverse of $\underline{g}_{(1)}$ and let $\hat{\underline{g}}_{(1)}^{-}$ be some function which satisfies (i) $\hat{\underline{g}}^-(t) = 0$, $t < \hat{\underline{g}}_{(1)}(0)$, (ii) $\hat{\underline{g}}^-(t) = 1$, $t > \hat{\underline{g}}_{(1)}(1)$, and (iii) for any $\hat{\underline{g}}_{(1)}(0) \leq t \leq \hat{\underline{g}}_{(1)}(1)$ there exists some s^* such that $\hat{\underline{g}}_{(1)}^{-}(t) = s^*$ and for which $\hat{\underline{g}}_{(1)}(s^*) = t$. Then

$$\hat{g}^1(x^1) = \hat{\underline{g}}_{(1)}^{-}\big(\hat{\underline{g}}^1(x^1)\big),$$

converges to g^1 at a rate of $O_p(n^{-1/2}h^{-d_1/2})$. A proof of this result is in lemma 26 in the appendix.

[10] To establish this result would entail a tedious repetition of the same arguments as were made in the proof of Theorem 1; to conserve space I have not done so.

[11] Note that if homogeneity itself is imposed efficiency achievements in addition to those possible with weak separability are feasible. See Tripathi (1997) for a discussion of imposing such conditions in nonparametric estimation.

[12] The assumption can be relaxed by using an iterative procedure, i.e. by replacing $\hat{a}$ in (35) with $\hat{a}_{S\bullet}$. See section 4.2 for a discussion of iterative procedures.

The convergence rates with (34) are hence complementary to those with (33). If $d_1 = d_m$ condition (33) yields estimators which converge at an optimal rate and when $d_0 + D = d_m$ it is (34) which yields such estimators. Finally, both identification conditions above go through similarly when $D > 1$.

4.2 Iteration

It is possible to iterate the estimator and, under certain circumstances, to improve the asymptotic efficiency of the procedure by combining uniterated and iterated estimators. I demonstrate the methodology using the simplest possible case, i.e. when $q = \chi$, $D = 1$, $\Lambda = 1$, $\tilde{\lambda}$ does not depend on x^1 and X_1^1 has compact support.[13]

The iterated estimator of a_0 is then $\hat{\hat{a}}_{S\chi}$, which is defined as $\hat{a}_{S\chi}$ with $\hat{\chi}^1$ replaced with

$$\hat{\hat{\chi}}^1(x^1) = \int \big(\hat{a}_{S\bullet}(x^0, x_0^1) - \hat{a}(x^0, x^1)\big)\lambda^*(x^0)dx^0, \tag{36}$$

Note that I use the unconstrained estimator $\hat{a}(x^0, x^1)$ in (36) because the fact that $a(x^0, x^1)$ is estimated is immaterial for the asymptotic distribution of $\hat{a}_{S\bullet}$. Appendix A.11 contains a somewhat heuristic derivation which shows that

$$\hat{\hat{a}}_{S\chi}(x_0) - \hat{a}_I(x_0) \approx -\big(\hat{a}_{S\chi}(x_0) - \hat{a}_I(x_0)\big), \tag{37}$$

where $\hat{a}_I$ is the NW estimator with regressors X_i^0, χ_i^1 and $\approx$ means 'up to terms which converge faster than $\hat{\hat{a}}_{S\chi}(x_0) - a(x_0)$ and $\hat{a}_{S\chi}(x_0) - a(x_0)$. Hence

$$\hat{a}_{C\chi}(x_0) = \big(\hat{\hat{a}}_{S\chi}(x_0) + \hat{a}_{S\chi}(x_0)\big)/2 \approx \hat{a}_I(x_0).$$

Therefore if $d_0 + 1 = d_1$ combining iterated and noniterated estimators in the above–described fashion removes the generated regressor component and the estimator subject to separability is then asymptotically as efficient as if χ^1 were observed. When $d_0 + 1 > d_1$ this had always been the case and when $d_0 + 1 < d_1$, there are additional terms which impact on the limiting distribution.

The result does extend to the case when $D > 1$, but not to $\hat{a}_{S\gamma}, \hat{a}_{S\pi}$ regardless of the choice of λ and it does not work for $\hat{a}_{S\chi}$ when Λ is not constant or indeed when $\tilde{\lambda}^j$ depends on x^j. It is possible that the methodology can be generalized to cover these cases, also, but such a procedure would be complicated.

[13] These conditions do not quite match those of theorem 1, but the argument and derivations simplify considerably.

5 Simulations

The simulation study compares $\hat{a}_{S\pi}$ to the unconstrained estimator $\hat{a}$ using the six models listed below.[14] I have chosen these particular models to highlight performance issues rather than concen-

$$
\begin{array}{ll}
\text{Linear} & Y_i = \sum_{t=1}^{d} X_{it} + U_i, \\
\text{Probit–Like} & Y_i = \bar{\Phi}\Big(\sum_{t=1}^{d} X_{it}\Big) + U_i, \\
\text{Product of Logs} & Y_i = \Big(\sum_{t=1}^{d_0}(X_{it}^0)^2\Big)\prod_{j=1}^{D}\Big(\operatorname{sgn}(X_{i1}^j)\log_{10}\big((X_{i1}^j)^4 + \sum_{t=1}^{d_j}(X_{it}^j)^2\big)\Big), \text{ if } d_0 > 0, \\
 & Y_i = \prod_{j=1}^{D}\Big(\arctan\big(\sum_{t=1}^{d_j} X_{it}^1\big)/\pi + 0.51\Big), \text{ if } d_0 = 0, \\
\text{Probit} & Y_i = I(\sum_{t=1}^{d} X_{it} + U_i \geq 0), \\
\text{Flat} & Y_i = 1, \\
\text{Arctan–Power} & Y_i = \Big(\sum_{t=1}^{d_0}(X_{it}^0)^2\Big)\left(1.5*\Big(\arctan\big(\sum_{t=1}^{d_1} X_{it}^1\big)/\pi + 0.51\Big)\right)^{\cdots}, \text{ if } d_0 > 0, \\
 & Y_i = \Big(\arctan\big(\sum_{t=1}^{d_1} X_{it}^1\big)/\pi + 0.51\Big)\left(1.5*\Big(\arctan\big(\sum_{t=1}^{d_2} X_{it}^2\big)/\pi + 0.51\Big)\right)^{\cdots}, \text{ if } d_0 = 0,
\end{array}
$$

trating on models with obvious empirical relevance.

I used the Mersenne–Twister random number generator (Matsumoto and Nishimura, 1998) because of its exceptional properties.[15] In all cases are the regressors independent and have $N(0,1)$–distributions. The errors are always independent of the regressors and are $N(0,\sigma^2)$–distributed for $\sigma = 1, 2$. The linear and probit models are standard. The probit–like model was included to compare performance when the systematic component of the model is bounded and indeed small relative to the variation in the errors. The product of logs and arctan models were included to simulate interaction terms. The arctan–power model does not reflect any model of interest in economics, but is introduced to assess the sensitivity of the results to the choice of relatively familiar stylized models.

In all cases λ was chosen equal to one and Λ was the function

$$\Lambda(x) = \left(1 + \exp\Big(e^{-1/\left(1-||x||^2/(dC_\Lambda^2)\right)}\Big)\right),$$

where $C_\Lambda = 4$.[16] The choice of C_λ was motivated by the results of preliminary experiments. A good choice of C_Λ is dependent on the scaling and distribution of the regressors.

[14] I attempted to make a comparison with the Horowitz (2001) estimator, but the computer program I wrote was insufficiently fast to conduct large scale simulations.

[15] It apparently has a proven period of $2^{19937} - 1$, excellent distributional properties and is exceptionally fast.

[16] In the experiments, the choice of Λ was based on the entire vectors X_i, not just on Z_i, as in the proofs.

I used the Mersenne–twister random number generator. The number of observations was either 100 or 200, the number of regressors 3, 4 or 9.[17]

When the number of regressors was 3, $d_0 = D = 1, d_1 = 2$, when $d = 4$, $d_0 = 0, d_1 = d_2 = D = 2$ and when $d = 9$, $d_0 = 0, d_1 = d_2 = d_3 = D = 3$. The number of replications was always 1,000. The regression function was estimated at X_i, $i = 1, \ldots, n$, and the results were aggregated across all i, and across replications.

The MSE entry in the tables is 1,000 times the average mean square error across all observations and all replications, MSE99 is the average mean square error over all replications, dropping the worst 1% in each replication and MDAE is the average median absolute error, where the average is taken over all observations.

The aggregate results show that $\hat{a}_{S\pi}$ can be unstable. This is not surprising since I chose Λ, h to be the same throughout and the estimator at points X_i at which $\Lambda = 0$ is then a weighted average over observations that are far away from X_i. Moreover, at such points the denominator of $\hat{a}_{S\pi}$ is likely to be close to zero and can even be negative because of the use of higher order kernels. Nevertheless, if even 1% of points is dropped at each replication, $\hat{a}_{S\pi}$ frequently performs better than does $\hat{a}$, particularly when the error variance is large relative to the variance of the structural part of the model. The results for $d = 3$ versus $d = 4, 9$ for the 'product of logs' and 'arctan' models are quite different given that the nature of the models for $d = 3$ is different from the other two.

Aggregate results obscure the strengths and weaknesses of $\hat{a}_{S\pi}$. As it turns out, $\hat{a}_{S\pi}$ performs particularly well (relative to $\hat{a}$) at those points x at which data are somewhat sparse but for which there are relatively many i for which $a(X_i)$ is close to $a(x)$. It performs poorly at points that are in the tail of the distribution, particularly so when there are few observations with similar regression function values. In regions with many data points, both estimators perform well.

To illustrate the variation in performance across points, consider figures 1a–1d. In all cases I used the linear specification and $d = 3, d_0 = D = 1$. Data were generated as described earlier with $n = 100$. I ran 1,000 simulations and estimated the regression function using $\hat{a}_{S\pi}$ at four different points, each point corresponding to one of the four graphs. Along with the density of $\hat{a}_{S\pi} - a$ I have plotted a normal density with zero mean and variance equal to the estimated variance of $\hat{a}_{S\pi} - a$. From figures 1a–1d, it is apparent that $\hat{a}_{S\pi}(x_0)$ is biased downward when most of the data would have regression function values that are less than $a(x_0)$ and upward if the converse is true. It is

[17]Speed of computation is not generally a constraint for individual data sets. Kim et al. (1999) propose a method which improves the computational efficiency of AS estimator. I have not investigated the possibility of devising a similar method for my WS estimators.

likely that this problem would be mitigated when a version of $\hat{a}_{S\pi}$ using local polynomial estimators instead of NW estimators would be used.[18] The results also suggest that even in moderate samples $\hat{a}_{S\pi}$ is approximately normal, although the results may be more favorable than would have been the case had I used a highly nonlinear model for these experiments.

The most interesting aspect of this experiment, however, is a comparison of the variances. The estimator variance is lowest at $(0, 0, 0)$ since the regressor density f is highest there. Normally, the estimator variances at $(-1, 1, -1)$ and $(1, 1, 1)$ would be comparable since $f(-1, 1, -1) = f(1, 1, 1)$. Note however that there are many more observations i for which $a(X_i) = X_{i1} + X_{i2} + X_{i3}$ is close to $a(-1, 1, -1) = -1$ than there are such observations which are close to $a(1, 1, 1) = 3$. The estimator variance at $(-1, 1, -1)$ is in fact only a little larger than that at $(0, 1, 0)$, despite the fact that $f(0, 1, 0)/f(-1, 1, -1) = e$. This suggests that $\hat{a}_{S\pi}$ does succeed in achieving the desired variance reduction.

6 Conclusions

I have proposed three estimators of regression functions which are weakly separable. Theoretical results show that the convergence rate of these estimators is comparable to that of the unconstrained Nadaraya–Watson kernel regression estimator of regression functions with fewer regressors. Simulation results in this paper suggest that in many, but not all, cases accuracy improvements indeed arise, provided that the separability assumption is correct. The asymptotic distribution of the estimators is normal. Computer simulations indicate that the approximate normality already obtains in small samples. However, although the bias does not feature in the asymptotic distribution it does have an effect in samples of moderate size. This problem could potentially be remedied by using a variant of the estimator which uses local polynomial estimators instead of Nadaraya–Watson ones. Nevertheless, the method proposed in this paper does reduce the estimator variance as expected.

I show that it is possible to identify and estimate the structural components of the separable structure individually. Doing so, however, results in a more complicated estimation procedure. I also show that it could be beneficial to iterate the estimator. Indeed, under strong conditions the use of generated regressors instead of observed ones is of no asymptotic consequence.

[18] The theoretical results assume the use of NW type estimators. There is no reason to believe that the results of this paper could not be obtained for local polynomial estimators, but proofs would be longer.

7 References Cited

Ahn, H. (1997), "Semiparametric estimation of a single–index model with nonparametrically generated regressors," Econometric Theory 12, 3–31.

Andrews, D.W.K. (1991), "Asymptotic normality of series estimators for nonparametric and semiparametric regression models," Econometrica 59, 307–345.

Baumol, W.J., J.C. Panzar and R.D. Willig (1982), "Contestable markets and the theory of industry structure," Harcourt Brace Jovanovich (New York).

Bellman, R.E. (1961), "Adaptive control processes," Princeton University Press.

Blackorby, C., R. Davidson and W. Schworm (1991), "Implicit separability: characterisation and implications for consumer demands," Journal of Economic Theory 55, 364–399.

Blundell, R. (1988), "Consumer behaviour: theory and empirical evidence – a survey," Economic Journal 98, 16–65.

Blundell, R. and J.–M. Robin (2000), "Latent separability: grouping goods without weak separability,", Econometrica 68, 53–84.

Breiman, L. and ,J. H. Friedman(1985). "Estimating Optimal Transformations for Multiple Regression and Correlation." Journal of the American Statistical Association, 80, 580–619.

Cannaday, R.E. (1994), "Condominium covenants: cats yes, dogs no," Journal of Urban Economics 35, 71–82.

Court, A.T. (1939), "Hedonic price indexes with automotive examples," in "The dynamics of automobile demand," General Motors, New York, 98–119.

Darolles, S, Florens, J.P. and E. Renault (2000), "Nonparametric instrumental regression," CREST working paper.

Delgado, M. and J. Mora (1995), "Nonparametric and semiparametric estimation with discrete regressors," Econometrica 63, 1477–1482.

Fan, J. and I. Gijbels (1996), "Local polynomial modelling and its applications," Chapman and Hall (London).

Fan, Y. and Q. Li (1996), "Consistent model specification tests: omitted variables and semiparametric functional forms,", Econometrica 64, 865–890.

Friedman, J.H. and W. Stützle (1981), "Projection pursuit regression,", Journal of the American Statistical Association 76, 817–823.

Goldman, S.M. and H. Uzawa (1964), "A note on separability in demand analysis," Econometrica

32, 387–398.

Gozalo, P.L. and O.B. Linton (2001), "Testing additivity in generalized nonparametric regression models," Journal of Econometrics 104, 1–48.

Hastie, T.J. and R. Tibshirani (1990), "Generalized additive models," Chapman and Hall, London.

Hoeffding, W. (1948), "A nonparametric test of independence," Annals of Mathematical Statistics 19, 546–557.

Horowitz, J.L. (1993), "Semiparametric estimation of a work–trip mode choice model," Journal of Econometrics 58, 49–70.

Horowitz, J.L. (1999), "Nonparametric estimation of a generalized additive model with an unknown link function," University of Iowa working paper.

Horowitz, J.L. (2001), "Nonparametric estimation of a generalized additive model with an unknown link function," Econometrica 69, 499–513.

Kim, W., O.B. Linton, and N. Hengartner (1999), "A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals," Journal of Computational and Graphical Statistics 8, 278–297.

Lazear, E. (1997), "Education: consumption or production," Journal of Political Economy 85, 569–598.

Leontief, W.W. (1947), "Introduction to a theory of the internal structure of functional relationships," Econometrica 15, 361–373.

Linton, O.B. (1997), "Efficient estimation of additive nonparametric regression models," Biometrika 84, 469–474.

Linton, O.B. (2000), "Efficient estimation of generalized additive nonparametric regression models," Econometric Theory 16, 502–523.

Linton, O.B. and J.P. Nielsen (1995), "Estimating structured nonparametric regression by the kernel method," Biometrika 82, 93–101.

Linton, O.B. and W. Härdle (1996), "Estimating additive regression models with known links," Biometrika 83, 529–540.

Matsumoto, M. and T. Nishimura (1998), "Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator," ACM Transactions on Modeling and Computer Simulation 8, 3–30.

Mincer, J. (1974), *Schooling, experience and earnings,* Columbia University Press (New York).

Nadaraya, E.A. (1964), "On estimating regression," Theory of Probability and its Applications 9,

141–142.
Newey, W.K. and J.L. Powell (1990), "Nonparametric instrumental variables estimation," Princeton University working paper.
Newey, W.K., Powell, J.L. and F. Vella (1999), "Nonparametric estimation of triangular simultaneous equations models," Econometrica 67, 565–603.
Parzen, E. (1962), "On estimation of a probability density function and mode," Annals of Mathematical Statistics 33, 1065–1076.
Pinkse, J. (1999), "Nonparametric misspecification testing," UBC working paper.
Pinkse, J. (2000), "Nonparametric two–step regression estimation when regressors and error are dependent ," Canadian Journal of Statistics 28, 289–300.
Pinkse, J and S. Ng (1998), "Nonparametric two–step regression estimation when regressors and error are dependent ," Canadian Journal of Statistics 28, 289–300.
Prakasa Rao, B.L.S. (1983), "Nonparametric functional estimation," Academic Press (New York).
Rilstone, P. (1996), "Nonparametric estimation of models with generated regressors," International Economic Review 37, 299–313.
Robinson, P.M. (1988), "Root–N–consistent semiparametric regression," Econometrica 56, 931–954.
Serfling, R. (1980), "Approximation theorems of mathematical statistics,", Wiley (New York).
Sheather, S.J. and M.C. Jones, "A reliable data–based bandwidth selection method for kernel density estimation," Journal of the Royal Statistical Society B 53, 683–690.
Stone, C.J. (1985), "Additive regression and other nonparametric models," Annals of Statistics 13, 689–705.
Stone, C.J. (1986), "The dimensionality reduction principle for generalized additive models," Annals of Statistics 14, 590–606.
Strotz (1957), "The empirical implications of a utility tree," Econometrica 25, 269–280.
Tjøstheim, D. and B. Auestad (1994), "Nonparametric identification of nonlinear time series: projections," Journal of the American Statistical Association 89, 1398–1409.
Train, K. (1980), "A structured model of automobile ownership and mode choice," Review of Economic Studies 47, 257–370.
Tripathi, G. (1997), "Semiparametric efficiency bounds and estimating models with shape restrictions," Ph.D. thesis, Northwestern university.
Watson, G.S. (1964), "Smooth regression analysis," Sankhyā A26, 359–372.

A Proofs

A.1 Some technical lemmas

The following lemma was inspired by theorem 2.1.1 of Prakasa Rao (1983) and uses a method developed by Parzen (1962).

Lemma 1 *For any* $\omega \in \mathcal{W}_{d_\xi,r}$,

$$\sup_{\xi_0 \in \mathbb{R}^{d_\xi}} \left| \int K(t-\xi_0)\omega(t)dt - \omega(\xi_0) \right| = O(h^r).$$

Proof: Use the substitution $t \leftarrow (t-\xi_0)/h$ to establish that

$$\int K(t-\xi_0)\omega(t)dt - \omega(\xi_0) == \int k(t)\big(\omega(\xi_0+th) - \omega(\xi_0)\big)dt.$$

Let $c_1, \ldots, c_{d_\xi}$ be constants in $[0,1]$ and let t_j, ξ_{0j} denote the j–th elements of the t, ξ_0–vectors. Note that by assumption F, $\int k(t)t_j^s dt = 0$ for $s = 1, \ldots, r-1$. Then applying the mean value theorem in each direction in turn and using the Laplace transformation for the remainder term, it follows that

$$\begin{aligned} & h^{-r} r! \int k(t)\big(\omega(\xi_0+th) - \omega(\xi_0)\big)dt \qquad (38) \\ = & \int k(t) t_1^r \frac{\partial^r \omega}{\partial \xi_1^r}(\xi_{01} + c_1 t_1 h, \xi_{02} + t_2 h, \ldots, \xi_{0d_\xi} + t_{d_\xi} h)dt \\ & + h^{-r} r! \int k(t)\big(\omega(\xi_{01}, \xi_{02} + t_2 h, \ldots, \xi_{0d_\xi} + t_{d_\xi} h) - \omega(\xi_0)\big)dt \\ = & \sum_{j=1}^{d_\xi} \int k(t) t_j^r \frac{\partial^r \omega}{\partial \xi_j^r}(\xi_{01}, \ldots, \xi_{0,j-1}, \xi_{0j} + c_j t_j h, \xi_{0,j+1} + t_{j+1} h, \ldots, \xi_{0d_\xi} + t_{d_\xi} h)dt. \qquad (39) \end{aligned}$$

The RHS in (39) is $O(1)$, uniformly in ξ_{01} by assumptions D and F. □

Lemma 2 *Let* $\{\xi_i\}$, $\xi_1 \in \mathbb{R}^{d_\xi}$ *be some i.i.d. sequence of random vectors with continuous density. Let* $\varsigma \in \mathbb{N}^{d_\xi}$, *and let*

$$K^{(\varsigma)} = \frac{\partial^{|\varsigma|} K}{\prod_{j=1}^{d_\xi} \partial \xi_j^{\varsigma_j}}, \qquad (40)$$

with $|\varsigma| = \sum_{j=1}^{d_\xi} \varsigma_j$. *If* $\{\omega_i\}$ *is an i.i.d. sequence of random vectors such that for some integer* $p^* > 0$ *and some* $C > \infty$ $\forall \xi : E(||\omega_1||^{p^*} | \xi_1 = \xi) \leq C$, *then*

$$E||K^{(\varsigma)}(\xi_1)\omega_1||^{p^*} = O(h^{(1-p^*)d_\xi - p^*|\varsigma|}), \qquad (41)$$

$$n^{-1} \sum_{i=1}^{n} ||K_0^{(\varsigma)}(\xi_i)\omega_i||^{p^*} = O_p(h^{(1-p^*)d_\xi - p^*|\varsigma|}). \qquad (42)$$

Proof: Taking expectations on the left hand side (LHS) in (42) gives (41), whose LHS is bounded by

$$CE|K^{(\varsigma)}(\xi_1)|^{p^*} = Ch^{-p^*(d_\xi+|\varsigma|)}E|k^{(\varsigma)}(\xi_1/h)|.$$

Since $|k^{(\varsigma)}|$ is itself a continuous, albeit usually not differentiable, function the same substitution can be applied as in lemma 1 and hence $E|k^{(\varsigma)}(\xi_1/h)| = O(h^{d_\xi})$. □

The following is a general lemma concerning V–statistics, which entails a small modification of the theorem on page 188 of Serfling (1980).[19]

Lemma 3 *Let $\zeta : \mathbb{R}^{2d} \to \mathbb{R}$ be a possibly asymmetric V–statistic kernel. If*

$$E\zeta_{12} = 0, \quad E|\zeta_{11}| = o(nc_n), \quad E\zeta_{12}^2 = o(n^2c_n^2), \tag{43}$$

then

$$n^{-2}\sum_{t=1}^{n}\sum_{i=1}^{n}\zeta_{it} = n^{-1}\sum_{t=1}^{n}(E_t\zeta_{it} + E_t\zeta_{ti}) + o_p(c_n) \tag{44}$$

Moreover, provided that the limit is finite and positive,

$$c_n^{-1}n^{-2}\sum_{t=1}^{n}\sum_{i=1}^{n}\zeta_{it} \xrightarrow{\mathcal{L}} N\Big(0, \lim_{n\to\infty} n^{-1}c_n^{-2}E\big((\zeta_{21}+\zeta_{12})(\zeta_{31}+\zeta_{13})\big)\Big) \tag{45}$$

If in addition to (43), $E\big((\zeta_{21}+\zeta_{12})(\zeta_{31}+\zeta_{13})\big) = o(nc_n^2)$, then

$$n^{-2}\sum_{t=1}^{n}\sum_{i=1}^{n}\zeta_{it} = o_p(c_n). \tag{46}$$

Proof: First (44). Let $\zeta^*(x,\tilde{x}) = \big(\zeta(x,\tilde{x}) + \zeta(\tilde{x},x)\big)/2$. Then

$$n^{-2}\sum_{t=1}^{n}\sum_{i=1}^{n}\zeta_{it} = n^{-2}\sum_{t=1}^{n}\sum_{i\neq t}^{n}\zeta_{it}^* + n^{-2}\sum_{i=1}^{n}\zeta_{ii}^*. \tag{47}$$

Apply the theorem on page 188 of Serfling (1980) to the first RHS term in (47). The second RHS term in (47) is also $o_p(c_n)$ because $E|\zeta_{11}| = o(nc_n)$ by (43). (45) and (46) follow from the fact that the first RHS term in (44) is an i.i.d. sum. □

[19] The result is originally due to Hoeffding (1948).

A.2 Approximation to V–statistic

When convenient I will use the alternative notation $\mathcal{P}_n\xi = n^{-1}\sum_{i=1}^n \xi$, $\mathcal{P}\xi = E\xi_1$. Let

$$\psi_n^* = h^r + n^{-1/2}h^{-d/2}\log n, \quad \psi_{nj} = h^r + n^{-1/2}h^{-d_j/2}\log n, \quad j = 1, \dots, D, \tag{48}$$

such that by assumption G

$$\psi_{nj} = O(\psi_n^*), \quad (\psi_n^*)^2 = o(\psi_{nj}), \quad j = 1, \dots, D. \tag{49}$$

Finally, let $\psi_n = \max_{j=1,\dots,D}\psi_{nj} = h^r + n^{-1/2}h^{-d_*/2}\log n$, with $d_* = \max(d_1, \dots, d_D)$.

Lemma 4 is required to transform $\hat{\chi}$ into form similar to that of $\hat{\gamma}, \hat{\pi}$.

Lemma 4 *In this lemma I use the shorthand notation $\tilde{\lambda}_0^j = \tilde{\lambda}^j(x^{-j}, x_0^j)$, $\tilde{\lambda}^j = \tilde{\lambda}^j(x^{-j}, x^j)$ and similar notation applies to other symbols. Let*

$$\bar{\chi}^j(x^j) = \int \big(\hat{\nu}_0^j\hat{f}^j - \hat{f}_0^j\hat{\nu}^j - (a_0^j - a^j)\hat{f}^j\hat{f}_0^j\big)\lambda^j\lambda_0^j dx^{-j} + \gamma^j(x^j), \quad j = 1, \dots, D.$$

$$\text{Then } \sup_{x^j}|\hat{\chi}^j(x^j) - \bar{\chi}^j(x^j)| = o_p\big((\psi_n^*)^2\big). \tag{50}$$

Proof: Horowitz (1999), lemma 1, provides a uniform consistency result for the numerator and denominator of kernel regression estimators. Here, together with lemma 1, it implies that $\sup_x|\hat{f}(x) - f(x)| = o_p(\psi_n^*)$, $\sup_x|\hat{\nu}(x) - \nu(x)| = o_p(\psi_n^*)$. Hence, recalling that $\lambda^j = \tilde{\lambda}^j/f^j$,

$$\begin{aligned}
\sup_{x^j}|\hat{\chi}^j - \bar{\chi}^j| &= \sup_{x^j}\left|\int \frac{\hat{\nu}_0^j\hat{f}^j - \hat{\nu}^j\hat{f}_0^j}{\hat{f}^j\hat{f}_0^j}\tilde{\lambda}^j\tilde{\lambda}_0^j dx^{-j} - \bar{\chi}^j\right| \\
&= \sup_{x^j}\left|\int (\hat{\nu}_0^j\hat{f}^j - \hat{f}_0^j\hat{\nu}^j)\left(1 + \frac{f^jf_0^j - \hat{f}^j\hat{f}_0^j}{\hat{f}^j\hat{f}_0^j}\right)\lambda^j\lambda_0^j dx^{-j} - \bar{\chi}^j\right| \\
&= \sup_{x^j}\left|\int (\hat{\nu}_0^j\hat{f}^j - \hat{f}_0^j\hat{\nu}^j)\left(\frac{2f^jf_0^j - \hat{f}^j\hat{f}_0^j}{f^jf_0^j}\right)\lambda^j\lambda_0^j dx^{-j} - \bar{\chi}^j\right| + o_p\big((\psi_n^*)^2\big) \\
&= \sup_{x^j}\left|\int \big(\hat{\nu}_0^j\hat{f}^j - \hat{f}_0^j\hat{\nu}^j + (a_0^j - a^j)(f^jf_0^j - \hat{f}^j\hat{f}_0^j)\big)\lambda^j\lambda_0^j dx^{-j} - \bar{\chi}^j\right| + o_p\big((\psi_n^*)^2\big) \\
&= o_p\big((\psi_n^*)^2\big),
\end{aligned}$$

where the third equality follows from

$$1 + \frac{f^jf_0^j - \hat{f}^j - \hat{f}_0^j}{\hat{f}^j\hat{f}_0^j} = 1 + \frac{f^jf_0^j - \hat{f}^j - \hat{f}_0^j}{f^jf_0^j} + \frac{(f^jf_0^j - \hat{f}^j\hat{f}_0^j)^2}{f^jf_0^j\hat{f}^j\hat{f}_0^j},$$

noting that f^j, f_0^j are bounded away from zero whenever $\lambda^j, \lambda_0^j > 0$ by assumption E. □

Let $\{A_t\}$, $\{B_t\}$ be sequences such that all elements in a sequence either equal 1 a.s. or equal Y_t. For some functions $\omega, \omega^* : \mathbb{R}^d \to \mathbb{R}$, let

$$\begin{aligned}
\eta_{A;\omega}(x) &= \lambda(x)\omega(x)f(x)E(A_1|X_1 = x), \quad \hat{\eta}_{A;\omega}(x) = \lambda(x)\omega(x)\mathcal{P}_n\big(K(x)A\big), \\
\alpha^j_{AB;\omega\omega^*}(x^j, \tilde{x}^j) &= \int \eta_{A;\omega}(x)\eta^j_{B;\omega^*}(x^{-j}, \tilde{x}^j)dx^{-j}, \quad \hat{\alpha}^j_{AB;\omega\omega^*}(x^j, \tilde{x}^j) = \int \hat{\eta}_{A;\omega}(x)\hat{\eta}^j_{B;\omega^*}(x^{-j}, \tilde{x}^j)dx^{-j}, \\
\hat{\beta}^j_{AB;\omega\omega^*}(x^j, \tilde{x}^j) &= \int \hat{\eta}_{A;\omega}(x)\eta^j_{B;\omega^*}(x^{-j}, \tilde{x}^j)dx^{-j}, \\
\hat{b}^j_{AB;\omega\omega^*}(x^j, \tilde{x}^j) &= n^{-1}\sum_{t=1}^n K_t(x^j)(\lambda^j \cdot \omega^j)(X_t^{-j}, x^j)\eta^j_{B;\omega^*}(X_t^{-j}, \tilde{x}^j)A_t.
\end{aligned}$$

In particular, let $\eta_A(x) = \eta_{A;1}(x)$ and similarly for $\hat{\eta}, \alpha, \hat{\alpha}, \hat{\beta}, \hat{b}$. Thus,

$$\begin{aligned}
\hat{\gamma}^j(x^j) = \hat{\alpha}^j_{1Y}(x^j, x^j_0) - \hat{\alpha}^j_{Y1}(x^j, x^j_0), \quad \gamma^j(x^j) = \alpha^j_{1Y}(x^j, x^j_0) - \alpha^j_{Y1}(x^j, x^j_0), \\
\hat{\delta}^j(x^j) = \hat{\alpha}^j_{11}(x^j, x^j_0), \quad \delta^j(x^j) = \alpha^j_{11}(x^j, x^j_0), \\
\bar{\chi}^j(x^j) = \hat{\alpha}^j_{1Y}(x^j, x^j_0) - \hat{\alpha}^j_{Y1}(x^j, x^j_0) + \hat{\alpha}^j_{11;a1}(x^j, x^j_0) - \hat{\alpha}^j_{11;1a}(x^j, x^j_0) + \chi^j(x^j).
\end{aligned} \tag{51}$$

I now establish uniform convergence of $\hat{\chi}^j, \hat{\gamma}^j, \hat{\pi}^j$ to χ^j, γ^j, π^j. Because of symmetry, lemmas 5 and 6 below equally apply when the arguments (x^j and x^j_0) are swapped.

Lemma 5 *For any functions ω, ω^* for which $\omega\lambda$, $\omega^*\lambda$ satisfy the conditions imposed on λ in assumption D and for $j = 1, \ldots, D$,*

$$\begin{aligned}
\sup_{x^j}\Big|&\big(\hat{\alpha}^j_{AB;\omega\omega^*}(x^j, x^j_0) - \alpha^j_{AB;\omega\omega^*}(x^j, x^j_0)\big) \\
&- \big(\hat{b}^j_{AB;\omega\omega^*}(x^j, x^j_0) - E\hat{b}^j_{AB;\omega\omega^*}(x^j, x^j_0)\big) - \big(\hat{b}^j_{BA;\omega^*\omega}(x^j_0, x^j) - E\hat{b}^j_{BA;\omega^*\omega}(x^j_0, x^j)\big) \\
&- \big((E\hat{\beta}^j_{AB;\omega\omega^*}(x^j, x^j_0) - \alpha^j_{AB;\omega\omega^*}(x^j, x^j_0)\big) - \big(E\hat{\beta}^j_{BA;\omega^*\omega}(x^j_0, x^j) - \alpha^j_{BA;\omega^*\omega}(x^j_0, x^j)\big)\Big| = o_p\big((\psi^*_n)^2\big).
\end{aligned}$$

Proof: Assume without loss of generality that $\omega = \omega^* = 1$. I establish that the following sufficient conditions hold.

$$\sup_{x^j} |\hat{\alpha}^j_{AB}(x^j, x^j_0) - \hat{\beta}^j_{AB}(x^j, x^j_0) - \hat{\beta}^j_{BA}(x^j_0, x^j) + \alpha^j_{AB}(x^j, x^j_0)| = o_p\big((\psi^*_n)^2\big), \tag{52}$$

$$\sup_{x^j}\Big|\big(\hat{\beta}^j_{AB}(x^j, x^j_0) - E\hat{\beta}^j_{AB}(x^j, x^j_0)\big) - \big(\hat{b}^j_{AB}(x^j, x^j_0) - E\hat{b}^j_{AB}(x^j, x^j_0)\big)\Big| = o_p\big((\psi^*_n)^2\big), \tag{53}$$

$$\sup_{x^j}\Big|\big(\hat{\beta}^j_{AB}(x^j_0, x^j) - E\hat{\beta}^j_{AB}(x^j_0, x^j)\big) - \big(\hat{b}^j_{AB}(x^j_0, x^j) - E\hat{b}^j_{AB}(x^j_0, x^j)\big)\Big| = o_p\big((\psi^*_n)^2\big), \tag{54}$$

First (52). The expression in absolute values is

$$\int \big(\hat{\eta}_A(x) - \eta_A(x)\big)\big(\hat{\eta}^j_B(x^{-j}, x^j_0) - \eta^j_B(x^{-j}, x^j_0)\big)dx^{-j}. \tag{55}$$

Again using lemma 1 of Horowitz (1999) together with lemma 1 implies that $\sup_x |\hat{\eta}_A(x) - \eta_A(x)| = o_p(h^r + n^{-1/2}h^{-d/2}\log n) = o_p(\psi_n^*)$ (and hence the same applies for $\hat{\eta}_B - \eta_B$). Therefore, (55) (and hence (52)) is $o_p\big((\psi_n^*)^2\big)$. Now (53).

$$\text{Let} \quad \xi_t(x^j) = A_t K_t(x^j)\left(\int K_t(x^{-j})\lambda^j(x^{-j},x^j)\eta_B^j(x^{-j},x_0^j)dx^{-j} - \lambda^j(X_t^{-j},x^j)\eta_B^j(X_t^{-j},x_0^j)\right).$$

Since $\hat{\beta}_{AB}^j(x^j,x_0^j)$ can alternatively be expressed as $\mathcal{P}_n\big(AK(x^j)\int K(x^{-j})\lambda^j(x^{-j},x^j)\eta_B^j(x^{-j},x_0^j)dx^{-j}\big)$, (53) is $(\mathcal{P}_n - \mathcal{P})\xi$. Since $\{\xi_t\}$ is i.i.d., lemma 1 of Horowitz (1999) implies that $(\mathcal{P}_n - \mathcal{P})\xi = o_p(n^{-1/2}\sqrt{E\xi_1^2}\log n)$. But by lemma 1

$$\sup_{\tilde{x}^{-j},x^j}\left|\int K(\tilde{x}^{-j}-x^{-j})\lambda^j(x^{-j},x^j)\eta_B^j(x^{-j},x_0^j)dx^{-j} - \lambda^j(\tilde{x}^{-j},x^j)\eta_B^j(\tilde{x}^{-j},x_0^j)\right| = O(h^r),$$

and because, again by lemma 2, $\sup_{x^j} E\big(A_1K_1(x^j)\big)^2 = O(h^{-d_j})$, it follows that $E\xi_1^2 = O(h^{2r-d_j})$. So (53) is $o_p(n^{-1/2}h^{r-d_j/2}\log n) = o_p(\psi_{nj}) = o_p(\psi_n^*)$ by (49). Condition (54) follows similarly. □

Let

$$\begin{aligned}
\hat{\gamma}^{*j}(x^j) &= \hat{\beta}_{1Y}^j(x^j,x_0^j) + \hat{\beta}_{Y1}^j(x_0^j,x^j) - \hat{\beta}_{Y1}^j(x^j,x_0^j) - \hat{\beta}_{1Y}^j(x_0^j,x^j) - \gamma^j(x^j),\\
\hat{\delta}^{*j}(x^j) &= \hat{\beta}_{11}^j(x^j,x_0^j) + \hat{\beta}_{11}^j(x_0^j,x^j) - \delta^j(x^j),\\
\hat{\pi}^{*j}(x^j) &= \pi^j(x^j) + \big(\hat{\gamma}^{*j}(x^j) - \pi^j(x^j)\hat{\delta}^{*j}(x^j)\big)/\delta^j(x^j),\\
\hat{\chi}^{*j}(x^j) &= \hat{\beta}_{1Y}^j(x^j,x_0^j) + \hat{\beta}_{Y1}^j(x_0^j,x^j) - \hat{\beta}_{Y1}^j(x^j,x_0^j) - \hat{\beta}_{1Y}^j(x_0^j,x^j) + \chi^j(x^j)\\
&\quad + \hat{\beta}_{11;a1}^j(x^j,x_0^j) + \hat{\beta}_{11;1a}^j(x_0^j,x^j) - \hat{\beta}_{11;1a}^j(x^j,x_0^j) - \hat{\beta}_{11;a1}^j(x_0^j,x^j),\\
\hat{\gamma}^{\bullet j}(x^j) &= \hat{b}_{1Y}^j(x^j,x_0^j) + \hat{b}_{Y1}^j(x_0^j,x^j) - \hat{b}_{Y1}^j(x^j,x_0^j) - \hat{b}_{1Y}^j(x_0^j,x^j) - \gamma^j(x^j),\\
\hat{\delta}^{\bullet j}(x^j) &= \hat{b}_{11}^j(x^j,x_0^j) + \hat{b}_{11}^j(x_0^j,x^j) - \delta^j(x^j)\\
\hat{\pi}^{\bullet j}(x^j) &= \pi^j(x^j) + \big(\hat{\gamma}^{\bullet j}(x^j) - \pi^j(x^j)\hat{\delta}^{\bullet j}(x^j)\big)/\delta^j(x^j),\\
\hat{\chi}^{\bullet j}(x^j) &= \hat{b}_{1Y}^j(x^j,x_0^j) + \hat{b}_{Y1}^j(x_0^j,x^j) - \hat{b}_{Y1}^j(x^j,x_0^j) - \hat{b}_{1Y}^j(x_0^j,x^j)\\
&\quad + \hat{b}_{11;a1}^j(x^j,x_0^j) + \hat{b}_{11;1a}^j(x_0^j,x^j) - \hat{b}_{11;1a}^j(x^j,x_0^j) - \hat{b}_{11;a1}^j(x_0^j,x^j) + \chi^j(x^j)\\
&= \hat{b}_{Y1}^j(x_0^j,x^j) - \hat{b}_{Y1}^j(x^j,x_0^j) + \hat{b}_{11;a1}^j(x^j,x_0^j) - \hat{b}_{11;a1}(x_0^j,x^j) + \chi^j(x^j),
\end{aligned} \tag{56}$$

where the last equality in (56) holds because $\hat{b}_{11;1a}^j = \hat{b}_{1Y}^j$.

Now, from lemmas 4 and 5 it follows that for $j = 1,\ldots,D$,

$$\sup_{x^j}\big|\big(\hat{\gamma}^j(x^j) - \gamma^j(x^j)\big) - \big(\hat{\gamma}^{\bullet j}(x^j) - E\hat{\gamma}^{\bullet j}(x^j)\big) - \big(E\hat{\gamma}^{*j}(x^j) - \gamma^j(x^j)\big)\big| = o_p\big((\psi_n^*)^2\big), \tag{57}$$

$$\sup_{x^j}\big|\big(\hat{\delta}^j(x^j) - \delta^j(x^j)\big) - \big(\hat{\delta}^{\bullet j}(x^j) - E\hat{\delta}^{\bullet j}(x^j)\big) - \big(E\hat{\delta}^{*j}(x^j) - \delta^j(x^j)\big)\big| = o_p\big((\psi_n^*)^2\big), \tag{58}$$

$$\sup_{x^j}\big|\big(\hat{\chi}^j(x^j) - \chi^j(x^j)\big) - \big(\hat{\chi}^{\bullet j}(x^j) - E\hat{\chi}^{\bullet j}(x^j)\big) - \big(E\hat{\chi}^{*j}(x^j) - \chi^j(x^j)\big)\big| = o_p\big((\psi_n^*)^2\big), \tag{59}$$

using (51). A similar result for $\hat{\pi}^j$ requires lemma 6 below, and is given in (67).

Lemma 6 *Let ω, ω^* be as in lemma 5. Then*

$$\sup_{x^j} |\hat{\alpha}^j_{AB;\omega\omega^*}(x^j, x^j_0) - \alpha^j_{AB;\omega\omega^*}(x^j, x^j_0)| = O_p(\psi_{nj}) \tag{60}$$

Proof: I again assume that $\omega = \omega^* = 1$ without loss of generality. Lemma 5 and (48) imply that the following four conditions are sufficient for (60).

$$\sup_{x^j} |\hat{b}^j_{AB}(x^j, x^j_0) - E\hat{b}^j_{AB}(x^j, x^j_0)| = o_p(\psi_{nj}),\ \sup_{x^j} |\hat{b}^j_{AB}(x^j_0, x^j) - E\hat{b}^j_{AB}(x^j_0, x^j)| = o_p(\psi_{nj}), \tag{61}$$

$$\sup_{x^j} |E\hat{\beta}^j_{AB}(x^j, x^j_0) - \alpha^j_{AB}(x^j, x^j_0)| = O(h^r),\ \sup_{x^j} |E\hat{\beta}^j_{AB}(x^j_0, x^j) - \alpha^j_{AB}(x^j_0, x^j)| = O(h^r). \tag{62}$$

First (61). I show the first result where the second follows similarly. Note that $\hat{b}^j_{AB}(x^j, x^j_0) - E\hat{b}^j_{AB}(x^j, x^j_0) = (\mathcal{P}_n - \mathcal{P})\xi$ with $\xi_t = K_t(x^j)\lambda^j(X^{-j}_t, x^j)\eta^j_B(X^{-j}_t, x^j_0)A_t$. Note that $E\xi_1^2 = O(h^{-d_j})$ by lemma 2 ($\varsigma = 0$) and again by lemma 1 of Horowitz (1999), $(\mathcal{P}_n - \mathcal{P})\xi = o_p(n^{-1/2}\sqrt{E\xi_1^2}\log n) = o_p(n^{-1/2}h^{-d_j/2}\log n) = o_p(\psi_{nj})$. Finally (62). I again only show the first result. In the proof of lemma 5, we expressed $\hat{\beta}^j_{AB}(x^j, x^j_0)$ as (rearranging terms) $\int \lambda(x)\eta^j_B(x^{-j}, x^j_0)\mathcal{P}_n\big(AK(x)\big)dx^{-j}$, such that

$$\begin{aligned}
&\sup_{x^j} |E\hat{\beta}^j_{AB}(x^j, x^j_0) - \alpha^j_{AB}(x^j, x^j_0)| \\
&\quad = \sup_{x^j} \left| \int \lambda(x)\eta_B(x^{-j}, x^j_0)\big(E(K_1(x)A_1) - E(A_1|X_1 = x)f(x)\big) \right| \\
&\qquad\qquad \leq \sup_{x^j} \int \big|\lambda(x)\eta^j_B(x^{-j}, x^j_0)\big|dx^{-j} \sup_x \big|E\big(K_1(x)A_1\big) - E(A_1|X_1 = x)f(x)\big| = O(h^r)
\end{aligned}$$

by lemma 1. □

Lemmas 4 and 6 have the following consequences (it may help to refer to (51)).

$$\sup_{x^j} |\hat{\gamma}^j(x^j) - \gamma^j(x^j)| = O_p(\psi_{nj}), \tag{63}$$

$$\sup_{x^j} |\hat{\delta}^j(x^j) - \delta^j(x^j)| = O_p(\psi_{nj}), \tag{64}$$

$$\sup_{x^j} |\hat{\chi}^j(x^j) - \chi^j(x^j)| = O_p(\psi_{nj}), \tag{65}$$

$$\sup_{x^j \in \bar{\mathcal{S}}^j} |\hat{\pi}^j(x^j) - \pi^j(x^j)| = O_p(\psi_{nj}), \tag{66}$$

$$\sup_{x^j \in \bar{\mathcal{S}}^j} \big|\big(\hat{\pi}^j(x^j) - \pi^j(x^j)\big) - \big(\hat{\pi}^{\bullet j}(x^j) - E\hat{\pi}^{\bullet j}(x^j)\big) - \big(E\hat{\pi}^{*j}(x^j) - \pi^j(x^j)\big)\big| \quad = o_p\big((\psi^*_n)^2\big). \tag{67}$$

Result (66) may be less obvious than (63)–(65). But since $\hat{\pi} - \pi = \big(\hat{\gamma} - \gamma - \pi(\hat{\delta} - \delta)\big)/\delta + (\hat{\pi} - \pi)(\delta - \hat{\delta})/\delta$, (63), (64) and because δ^j is bounded away from zero on $\bar{\mathcal{S}}^j$ (66) holds. Finally, (67) follows similarly from (57) and (58).

Let $\mathcal{Q}^j = \{x^j \in \mathcal{S}^j : \gamma^j(x^j) = 0\}$ and $\bar{\mathcal{Q}}^j = \{x^j \in \bar{\mathcal{S}}^j : \gamma^j(x^j) = 0\}$. The following lemma establishes that γ^j and π^j are strictly decreasing in x^{j1} in a neighborhood of the set $\mathcal{Q}^j$.

Lemma 7 *For any $j = 1, \ldots, D$,*

$$\max_{x^j \in \bar{\mathcal{Q}}^j} \frac{\partial \gamma^j}{\partial x^{j1}}(x^j) < 0, \quad \max_{x^j \in \bar{\mathcal{Q}}^j} \frac{\partial \pi^j}{\partial x^{j1}}(x^j) < 0.$$

Proof: Note that $\max_{x^j \in \bar{\mathcal{Q}}^j} \frac{\partial \gamma^j}{\partial x^{j1}}(x^j) = -\min_{x^j \in \bar{\mathcal{Q}}^j} \int \left(\frac{\partial a^j}{\partial x^{j1}} \cdot f^j \cdot \lambda^j \right)(x^{-j}, x^j)(f^j \cdot \lambda^j)(x^{-j}, x_0^j) dx^{-j} < 0,$

because $\bar{\mathcal{S}}^j$ is assumed compact (which implies that $\bar{\mathcal{Q}}^j$ is compact also) and a is assumed to be strictly increasing in x^{j1}. The result for π^j follows from the fact that (omitting arguments)

$$\frac{\partial \pi^j}{\partial x^{j1}} = \left(\frac{\partial \gamma^j}{\partial x^{j1}} - \pi^j \frac{\partial \delta^j}{\partial x^{j1}} \right) / \delta^j.$$

But for any $x^j \in \bar{\mathcal{Q}}^j$, $\pi^j(x^j) = 0$ and δ^j is positive and bounded away from zero on $\bar{\mathcal{S}}^j$. □

Recall that q is one of γ, π, χ. Let $\hat{q}$, $\hat{q}^\bullet$ and any other symbols related to q be the corresponding symbol in terms of γ, π, χ. The problem with the functions q^j is that they can be zero even when $g(x^j) \neq g(x_0^j)$. I therefore approximate q by a new function $\tilde{q}$.

Let $\tilde{q}$ be a function with the same smoothness properties as q, for which for all $j = 1, \ldots, D$, $\tilde{q}^j$ is everywhere monotonically decreasing in x^{j1} and identical to q^j in an open neighborhood $\mathcal{N}^j$ of $\mathcal{Q}^j$. Such a function $\tilde{q}$ exists by lemma 7.

In lemmas 8–10 I derive a first expansion of $\hat{a}_{S\bullet}(x_0) - a(x_0)$.

Lemma 8 *For any i.i.d. sequence $\{\xi_i\}$ with $E|\xi_1|^{1+\epsilon_\xi}$ for some $\epsilon_\xi > 0$,*

$$n^{-1} \sum_{i=1}^n K_0(X_i^0) \xi_i \Lambda_i \left(K_0(\hat{q}_i) - K_0(\tilde{q}_i) - \sum_{j=1}^D K_0(\tilde{q}_i^{-j}) K_0'(\tilde{q}_i^j)(\hat{q}_i^j - q_i^j) \right) = o_p(\rho_n). \tag{68}$$

Proof: The following two results are sufficient for (68).

$$n^{-1} \sum_{i=1}^n K_0(X_i^0) \xi_i \Lambda_i \left(K_0(\hat{q}_i) - K_0(q_i) - \sum_{j=1}^D K_0(q_i^{-j}) K_0'(q_i^j)(\hat{q}_i^j - q_i^j) \right) = o_p(\rho_n), \tag{69}$$

$$n^{-1} \sum_{i=1}^n K_0(X_i^0) \xi_i \Lambda_i \left(\left(K_0(q_i) - K_0(\tilde{q}_i) \right) + \sum_{j=1}^D \left(K_0(q_i^{-j}) K_0'(q_i^j)(\hat{q}_i^j - q_i^j) - K_0(\tilde{q}_i^{-j}) K_0'(\tilde{q}_i^j)(\hat{q}_i^j - q_i^j) \right) \right) = o_p(\rho_n). \tag{70}$$

First (69). Below ς is always a D–vector of nonnegative integers and $|\varsigma|$ denotes its largest element. Choose some Φ, with $3 < \Phi \leq r+2$. Let $\iota_\varsigma \in \mathbb{R}^D$ be such that $\iota_{\varsigma_j} = I(\varsigma_j = \Phi)$ with I the indicator function. Further, let exponents in parentheses denote derivative order. Then by the mean value theorem (order Φ in each direction) the LHS in (69) is

$$\sum_{2\leq|\varsigma|\leq\Phi} n^{-1}\sum_{i=1}^{n} K_0(X_i^0)\xi_i\Lambda_i \prod_{j=1}^{D}\left(\frac{(\hat{q}_i^j - q_i^j)^{\varsigma_j}}{\varsigma_j!}\right)\prod_{j=1}^{D}\left(\left(K_0^{(\varsigma_j)}(q_i^j)\right)^{1-\iota_{\varsigma_j}}\left(K_0^{(\Phi)}(\cdot)\right)^{\iota_{\varsigma_j}}\right), \tag{71}$$

where $(\cdot)$ denotes some quantity between q_i^j and $\hat{q}_i^j$. Now, by (63), (65) and(66), $\sup_{x^j}|\hat{q}^j(x^j) - q^j(x^j)| = O_p(\psi_n)$. Hence

$$\max_{i=1,\dots,n}\left|\prod_{j=1}^{D}\left(\frac{(\hat{q}_i^j - q_i^j)^{\varsigma_j}}{\varsigma_j!}\right)\right| = O_p(\psi_n^{|\varsigma|}). \tag{72}$$

Further, $\sup_{q^j}|K_0^{(\Phi)}(q^j)| = O(h^{-\Phi-1})$ since $k^{(\Phi)}$ is assumed bounded. Let $\bar{k} = |k|$ and let $\bar{K}$ be defined accordingly. Thus, for any $2 \leq |\varsigma| \leq \Phi$, the LHS summand in (71) is

$$O_p\left(\psi_n^{|\varsigma|}h^{-|\iota_\varsigma|}n^{-1}\sum_{i=1}^{n}\left|K_0(X_i^0)\xi_i\Lambda_i\prod_{j=1}^{D}\left(\bar{K}_0^{(\varsigma_j)}(q_i^j)\right)^{1-\iota_{\varsigma_j}}\right|\right) = O_p(\psi_n^{|\varsigma|}h^{-|\varsigma|-|\iota_\varsigma|}), \tag{73}$$

by lemma 2. Convergence is slowest, either when $|\varsigma| = 2$ or when $|\varsigma| = \Phi$ and $|\iota_\varsigma| = 1$, which results in a convergence rate of $\psi_n^2 h^{-2} + \psi_n^\Phi h^{-\Phi-1}$. Since $\psi_n = O(\rho_n \log n)$,

$$\psi_n^2 h^{-2} + \psi_n^\Phi h^{-\Phi-1} = O\Big(\rho_n\big(\rho_n h^{-2}\log^2 n) + \rho_n^{\Phi-1}h^{-\Phi-1}\log^\Phi n\big)\Big) = o(\rho_n),$$

since $\rho_n h^{-2}\log^2 n = h^{r-2}\log^2 n + n^{-1/2}h^{-d_m/2-2}\log^2 n = o(1)$ and $\rho_n^{\Phi-1}h^{-\Phi-1}\log^\Phi n = h^{\Phi(r-1)-r-1}\log^\Phi n + \left(n^{-1/2}h^{-d_m/2-1}\right)^{\Phi-1}h^{-2}\log^\Phi n = o(1)$, which follows from assumption G because $r > 2$, $nh^{d_m+4}/\log^2 n \to \infty$, and $\Phi > 3$.

Now (70). I will show that for any $s > 0$, for any integer $0 \leq t \leq \Phi$, and for any $j = 1, \dots, D$,

$$\sup_{x^j}\left|K^{(t)}\big(q^j(x^j)\big) - K^{(t)}\big(\tilde{q}^j(x^j)\big)\right|\Lambda^j(x^j) = o(n^{-s}), \tag{74}$$

i.e. the LHS expression converges at a rate faster than any power of n. Condition (74) implies (70). First, for $x^j \notin \bar{\mathcal{S}}^j$ or $x^j \in \mathcal{N}^j$, the LHS in (74) is zero by the definitions of Λ^j and $\tilde{q}^j$. But both q^j and $\tilde{q}^j$ are nonzero on $\bar{\mathcal{S}}^j\backslash\mathcal{N}^j$ and, since $\bar{\mathcal{S}}^j\backslash\mathcal{N}^j$ is compact, they are in fact bounded away from zero on $\bar{\mathcal{S}}^j\backslash\mathcal{N}^j$. Since $k^{(t)}$ has exponentially declining tails by assumption F and because h decreases faster than some power of n, for any $c_q > 0$, $k^{(t)}(c_q/h)$ decreases exponentially in n. □

Lemma 9 *For any function ω for which $\omega^2 f$ is boundedly differentiable, if $\hat{\Delta}_\bullet(\omega)$ is as defined in (27) then*

$$\hat{\Delta}_\bullet(\omega) - \Delta_\bullet(\omega) = o_p(1).$$

Proof: Write

$$\hat{\Delta}_\bullet(\omega) - \Delta_\bullet(\omega) = \hat{\Delta}_\bullet - n^{-1}\sum_{i=1}^{n} K_0(X_i^0)\Lambda_i\omega_i\left(K_0(\hat{q}_i) - K_0(\tilde{q}_i) - \sum_{j=1}^{D} K_0(\tilde{q}_i^{-j})K_0'(\tilde{q}_i^j)(\hat{q}_i^j - q_i^j)\right) \quad (75)$$

$$+ n^{-1}\sum_{i=1}^{n} K_0(X_i^0)\Lambda_i\omega_i\left(\sum_{j=1}^{D} K_0(\tilde{q}_i^{-j})K_0'(\tilde{q}_i^j)(\hat{q}_i^j - q_i^j)\right) \quad (76)$$

$$+ n^{-1}\sum_{i=1}^{n}\big(K_0(X_i^0)\Lambda_i\omega_i K_0(\tilde{q}_i) - \Delta_\bullet\big). \quad (77)$$

Lemma 8 establishes that the RHS in (75) is $o_p(\rho_n) = o_p(1)$. By (63), (65) and (66), $\hat{q}_i^j - q_i^j = o_p(\psi_n)$, uniformly in i, j and by lemma 2, (76) is hence $o_p(\psi_n h^{-1}) = o_p(1)$.[20] Finally, (77) is the sample mean of a sequence of i.i.d. random variables. I show that the variances of the elements of the sequence are $o(n)$ and their means are $o(1)$. Let $\bar{k}(t) = k^2(t)/\int k^2(t)dt$, which is an even second order kernel. Then the variance of a summand of (77) is bounded by

$$E\big(K_0(X_1^0, \tilde{q}_1)\Lambda_1\omega_1\big)^2 = O(h^{-d_0-D}) = O(h^{-d_m}) = o(n),$$

by lemma 2, where the last equality follows from assumption G. For the bias, observe that

$$E\big(K_0(X_1^0, \tilde{q}_1)\Lambda_1\omega_1\big) - \Delta_\bullet(\omega)$$
$$= E\big(K_0(X_1^0, \tilde{q}_1)\Lambda_1\omega_1|\Lambda_1 > 0\big)p_0 - E\big(\Lambda_1\omega_1|\Lambda_1 > 0, X_1^0 = x_0^0, \tilde{q}_1 = 0\big)f_{\bullet+}(x_0^0, 0)p_0.$$

Apply lemma 1. □

Let $M_i = (Y_i - a)/\Delta_\bullet$ and $\mu^{\bullet j}(x^j) = E\hat{q}^{\bullet j}(x^j)$, $\mu^{*j}(x^j) = E\hat{q}^{*j}(x^j)$, and

$$F_i = K_0(X_i^0, \tilde{q}_i)\Lambda_i M_i, \quad \hat{F}_i = K_0(X_i^0, \hat{q}_i)\Lambda_i M_i, \quad i = 1, \ldots, n. \quad (78)$$

Lemma 10 *Let $T_i^j = K_0(X_i^0, \tilde{q}_i^{-j})\Lambda_i M_i K_0'(\tilde{q}_i^j)$. Then*

$$\big(\hat{a}_{S\bullet}(x_0) - a(x_0)\big) = n^{-1}\sum_{i=1}^{n} F_i + \sum_{j=1}^{D} n^{-1}\sum_{i=1}^{n} T_i^j(\hat{q}_i^{\bullet j} - \mu_i^{\bullet j}) + \sum_{j=1}^{D} n^{-1}\sum_{i=1}^{n} T_i^j(\mu_i^{*j} - q_i^j) + o_p(\rho_n). \quad (79)$$

[20] See the treatment of the higher order terms in lemma 8 for a more elaborate discussion of a similar expression.

Proof: Let $\hat{\Gamma} = \hat{a}_{S\bullet}\hat{\Delta}_\bullet$, $\Gamma = a/\Delta_\bullet$ such that (omitting arguments)

$$\hat{a}_{S\bullet} - a = \frac{\hat{\Gamma} - a\hat{\Delta}_\bullet}{\Delta_\bullet} - (\hat{a}_{S\bullet} - a)\frac{\hat{\Delta}_\bullet - \Delta_\bullet}{\Delta_\bullet}. \tag{80}$$

By lemma 9 (choose $\omega = 1$), the second RHS term in (80) converges faster than does the first. The first RHS term in (80) is $n^{-1}\sum_{i=1}^n \hat{F}_i$. Apply lemma 8 with $\xi_i = M_i$, to establish that

$$n^{-1}\sum_{i=1}^n \hat{F}_i = n^{-1}\sum_{i=1}^n F_i + \sum_{j=1}^D n^{-1}\sum_{i=1}^n T_i^j(\hat{q}_i^j - q_i^j) + o_p(\rho_n). \tag{81}$$

The first RHS term in (81) is the first RHS term in (79). I now establish that the second RHS term in (81) is the same as the sum of the second and third RHS terms in (79) except for an asymptotically negligble term. By (57), (59), and (67),

$$\max_{i=1,\dots,n}\left|(\hat{q}_i^j - q_i^j) - (\hat{q}_i^{\bullet j} - \mu_i^{\bullet j}) - (\mu_i^{*j} - q_i^j)\right| = o_p\big((\psi_n^*)^2\big), \quad j = 1, \dots, D. \tag{82}$$

So for $j = 1, \dots, D$,

$$\begin{aligned}
&\left|n^{-1}\sum_{i=1}^n K_0(X_i, \tilde{q}_i^{-j})K_0'(\tilde{q}_i^j)\Lambda_i M_i\big((\hat{q}_i^j - q_i^j) - (\hat{q}_i^{\bullet j} - E\hat{q}_i^{\bullet j}) - (E\hat{q}_i^{*j} - q_i^j)\big)\right| \\
&\quad \le \max_{i=1,\dots,n}\left|(\hat{q}_i^j - q_i^j) - (\hat{q}_i^{\bullet j} - \mu_i^{\bullet j}) - (\mu_i^{*j} - q_i^j)\right| \times n^{-1}\sum_{i=1}^n \left|K_0(X_i, \tilde{q}_i^{-j})K_0'(\tilde{q}_i^j)\Lambda_i M_i\right| \\
&\quad = o_p\big((\psi_n^*)^2 h^{-1}\big) = o_p(\rho_n),
\end{aligned}$$

by (82), lemma 2 and assumption G. □

In the remainder of the proof of theorem 1, I introduce new symbols $\hat{q}_{Ii}^{\bullet j}, \hat{q}_{IIi}^{\bullet j}, \mu_{Ii}^{\bullet j}, \mu_{IIi}^{\bullet j}, \mu_{\delta i}^{\bullet j}$ below, which allow me to rewrite (79) as

$$\hat{a}_{S\bullet}(x_0) - a(x_0) = n^{-1}\sum_{i=1}^{n}(F_i - EF_1) \tag{83}$$

$$+ \sum_{j=1}^{D} n^{-1}\sum_{i=1}^{n} T_i^j \Psi_{\bullet i}^j(\hat{q}_{Ii}^{\bullet j} - \mu_{Ii}^{\bullet j}) \tag{84}$$

$$- \sum_{j=1}^{D} n^{-1}\sum_{i=1}^{n} T_i^j \Psi_{\bullet i}^j(\hat{q}_{IIi}^{\bullet j} - \mu_{IIi}^{\bullet j}) \tag{85}$$

$$+ \sum_{j=1}^{D} n^{-1}\sum_{i=1}^{n}\Big(T_i^j(\mu_i^{*j} - q_i^j) - E\big(T_1^j(\mu_1^{*j} - q_1^j)\big)\Big) \tag{86}$$

$$+ I(q=\pi)\sum_{j=1}^{D} n^{-1}\sum_{i=1}^{n} T_i^j \pi_i^j(\hat{\delta}_i^{\bullet j} - \mu_{\delta i}^j) \tag{87}$$

$$+ EF_1 + \sum_{j=1}^{D} E\big(T_1^j(\mu_1^{*j} - q_1^j)\big) \tag{88}$$

$$+ o(\rho_n). \tag{89}$$

I now introduce the new notation, after which I provide a brief outline of the remainder of the proofs. Let $K_{ts}^j = K(X_t^j - X_s^j), K_{t0}^j = K(X_t^j - K_0^j)$, but for all symbols not pertaining to kernels, let $\lambda_{ts}^j = \lambda^j(X_t^{-j}, X_s^j)$, $\lambda_{t0}^j = \lambda^j(X_t^{-j}, x_0^j)$ and similarly for other symbols. Let further

$$G_{tsi}^j = K_{ts}^j \lambda_{ts}^j \lambda_{ti}^j f_{ti}^j, \quad j = 1, \ldots, D;\ i, t, s = 0, \ldots, n, \tag{90}$$

and for $i = 1, \ldots, n; u = 0, i$,

$$\hat{q}_{Iiu}^j = n^{-1}\sum_{i=1}^{n} G_{t0i}^j(Y_t - a_{tu}^j), \quad \hat{q}_{IIiu}^j = n^{-1}\sum_{i=1}^{n} G_{ti0}^j(Y_t - a_{tu}^j), \tag{91}$$

with $a_{ti}^j = a^j(X_t^{-j}, X_i^j)$ and $a_{t0}^j = a^j(X_t^{-j}, x_0^j)$.
Then each of the $\hat{b}$–symbols in (56) can be expanded, e.g.

$$\hat{b}_{Y1}^j(x_0^j, X_i^j) = n^{-1}\sum_{t=1}^{n} G_{t0i}^j Y_t,$$

such that

$$\hat{\gamma}_i^{\bullet j} = n^{-1}\sum_{t=1}^{n} G_{t0i}^j(Y_t - a_{ti}^j) - n^{-1}\sum_{t=1}^{n} G_{ti0}^j(Y_t^j - a_{t0}^j) - \gamma_i^j = \hat{q}_{Iii}^j - \hat{q}_{IIi0}^j - \gamma_i^j, \tag{92}$$

$$\hat{\delta}_i^{\bullet j} = n^{-1}\sum_{t=1}^{n}(G_{t0i}^j + G_{ti0}^j) - \delta_i^j, \tag{93}$$

$$\hat{\chi}_i^{\bullet j} = n^{-1}\sum_{t=1}^{n} G_{t0i}^j(Y_t^j - a_{t0}^j) - n^{-1}\sum_{t=1}^{n} G_{ti0}^j(Y_t^j - a_{ti}^j) + \chi_i^j, = \hat{q}_{Ii0}^j - \hat{q}_{IIii}^j + \chi_i^j, \tag{94}$$

$$\hat{\pi}_i^{\bullet j} = \hat{\gamma}_i^{\bullet j}/\delta_i^j - \pi_i^j(\hat{\delta}_i^{\bullet j} - \delta_i^j)/\delta_i^j = (\hat{q}_{Iii}^j - \hat{q}_{IIi0}^j)/\delta_i^j - \pi_i^j - \pi_i^j(\hat{\delta}_i^{\bullet j} - \delta_i^j) \tag{95}$$

Set

$$\hat{q}_{Ii}^{\bullet j} = \begin{cases} \hat{q}_{Iii}^j, & q = \gamma, \pi, \\ \hat{q}_{Ii0}^j, & q = \chi. \end{cases} \qquad \hat{q}_{IIi}^{\bullet j} = \begin{cases} \hat{q}_{IIi0}^j, & q = \gamma, \pi, \\ \hat{q}_{IIii}^j, & q = \chi. \end{cases}$$

Then from (92)–(95) it follows that

$$\hat{q}_i^{\bullet j} = \Psi_{\bullet i}^j(\hat{q}_{Ii}^{\bullet j} - \hat{q}_{IIi}^{\bullet j}) + I(q = \pi)\pi_i^j\hat{\delta}_i^{\bullet j}.$$

Now let the symbols $\mu_{Ii}^{\bullet j}, \mu_{IIi}^{\bullet j}, \mu_{\delta i}^{\bullet j}, \mu_{Iii}^j, \mu_{Ii0}^j$ be to $\hat{q}_{Ii}^{\bullet j}, \hat{q}_{IIi}^{\bullet j}, \hat{\delta}_i^{\bullet j}, \hat{q}_{Iii}^j, \hat{q}_{Ii0}^j$, what $\mu_i^{\bullet j}$ is to $\hat{q}_i^{\bullet j}$.[21] Then

$$\mu_i^{\bullet j} = \Psi_{\bullet i}^j(\mu_{Ii}^{\bullet j} - \mu_{IIi}^{\bullet j}) + I(q = \pi)\pi_i^j\mu_{\delta i}^{\bullet j}.$$

The expansion in (83)–(89) then follows from (79).

In appendix A.4 I deal with (88), appendix A.5 covers (83) and (84) and appendix A.6 establishes that (85)–(87) are asymptotically negligble. First, appendix A.3 provides some useful lemmas on generated regressors which are used in subsequent proofs.

A.3 Generated Regressors

For $j = 1, \ldots, D$, let $\tilde{\tau}^j$ be such that $\tilde{\tau}^j\left(\tilde{q}^j(x^j), x^{j2}\right) = x^{j1}$ for all x^j. By construction $\tilde{q}^j$ is strictly monotonic in x^{j1} and hence $\tilde{\tau}^j$ is well–defined. Let $\mathcal{W}_r^{*j}$ be the class of functions $\omega : \mathbb{R}^{d_j} \to \mathbb{R}$ for which $\partial^r\omega/\partial\xi_1^r$ is continuous, where ξ_1 denotes the first argument of ω. The lemmas below deal with generated regressors and are used in subsequent appendices.

Lemma 11

$$\frac{\partial\tilde{\tau}^j}{\partial q^j} \in \mathcal{W}_r^{*j}, \quad j = 1, \ldots, D.$$

[21] See lemma 10.

Proof: Note that

$$\frac{\partial \tilde{\tau}^j}{\partial \tilde{q}^j} = 1/\frac{\partial \tilde{q}^j}{\partial x^{j1}}.$$

Higher order partial derivatives of $\tilde{\tau}^j$ with respect to $\tilde{q}^j$ thus have higher order derivatives of $\tilde{q}^j$ with respect to x^{j1} in the numerator and powers of the first partial in the denominator. The first partial of $\tilde{q}^j$ with respect to x^{j1} is bounded away from zero and the $r+1$–st derivative is bounded by construction. □

For any function $\omega : \mathbb{R}^{2d} \to \mathbb{R}$, let $\Delta^*_{(t)}(\omega) = \Delta^*_{(t)}(\omega, x^0_0, 0)$ with

$$\Delta^*_{(t)}(\omega, x^0, q) = \int \frac{\omega\big(x^0, \tilde{\tau}^1(q^1, x^{12}), x^{12}, \ldots, \tilde{\tau}^D(q^D, x^{D2}), x^{D2}, t\big)}{\prod_{j=1}^D \frac{\partial \tilde{q}^j}{\partial x^{j1}}\big(\tilde{\tau}^j(q^j, x^{j2}), x^{j2}\big)} dx^{12} \cdots dx^{D2}, \tag{96}$$

$$\Delta_{(t)}(\omega) = \Delta_\bullet\big(\omega(\cdot, t)\big). \tag{97}$$

Lemma 12 *For any function $\omega : \mathbb{R}^{2d} \to \mathbb{R}$ and vectors q, t for which both sides exist,* $\Delta^*_{(t)}(\omega\Lambda f) = \Delta_{(t)}(\omega)$.

Proof: Let $\omega^* = \omega\Lambda$ and let $\mathcal{C}(x)$ be shorthand for $\big(x^0_0, \tilde{\tau}^1(0, x^{12}), x^{12}, \ldots, \tilde{\tau}^D(0, x^{D2}), x^{D2}\big)$. Note that from (96) and (97)

$$\begin{aligned}
\Delta_{(t)}(\omega) &= E\big(\omega(X_1, t)\Lambda_1 | X^0_1 = x^0_0, g_1 = g(z_0), \Lambda_1 > 0\big) f^+_\bullet(x^0_0, 0) p_0 \\
&= E\Big(\omega^*\big(\mathcal{C}(X_1), t\big) | X^0_1 = x^0_0, \tilde{q}(Z_1) = 0, \Lambda_1 > 0\Big) f^+_\bullet(x^0_0, 0) p_0 \\
&= p_0 \int \omega^*\big(\mathcal{C}(x), t\big) f_{X^0, X^{12}, \ldots, X^{D2}, q | \Lambda > 0}(x^0_0, x^{12}, \ldots, x^{D2}, 0) dx^{12} \cdots dx^{D2} \\
&= p_0 \int \frac{\omega^*\big(\mathcal{C}(x), t\big)}{\prod_{j=1}^D \frac{\partial \tilde{q}^j}{\partial x^{j1}}\big(\tilde{\tau}^j(0, x^{j2}), x^{j2}\big)} f_{X|\Lambda>0}\big(\mathcal{C}(x)\big) dx^{12} \cdots dx^{D2} \\
&= \int \frac{\omega^*\big(\mathcal{C}(x), t\big)}{\prod_{j=1}^D \frac{\partial \tilde{q}^j}{\partial x^{j1}}\big(\tilde{\tau}^j(0, x^{j2}), x^{j2}\big)} f\big(\mathcal{C}(x)\big) dx^{12} \cdots dx^{D2} = \Delta^*_{(t)}(\omega^* f) = \Delta^*_{(t)}(\omega\Lambda f),
\end{aligned} \tag{98}$$

where the fourth equality follows from substitution of $x^{j1} = \tilde{\tau}(q^1, x^{j2})$. □

Let $\bar{\mathcal{W}}_{2d,r}$ be the class of functions $\omega : \mathbb{R}^{2d} \to \mathbb{R}$, i.e. $(\omega(x, t))$, for which the r–th partial derivatives with respect to any element of the x–vector are continuous in $x^0, x^1, \ldots, x^D$, uniformly in t.

Lemma 13 *For any function $\omega \in \bar{\mathcal{W}}_{2d,r}$,*

$$\sup_t \Big| E\big(K_0(X^0_1) K_0(\tilde{q}_1) \omega(X_1, t) \Lambda_1\big) - \Delta_{(t)}(\omega\Lambda f) \Big| = O(h^r). \tag{99}$$

Proof: By substitution of $\tilde{q}^j = \tilde{q}^j(x^j)$ and lemma 12, the LHS in (99) is

$$\sup_t \left| \int K_0(x^0) K_0(\tilde{q}) \Delta^*_{(t)}(\omega \Lambda f, x^0, \tilde{q}) dx^0 d\tilde{q} - \Delta^*_{(t)}(\omega \Lambda f, x_0^0, 0) \right| \tag{100}$$

Lemma 1 requires $\Delta^*_{(t)}(\omega \Lambda f, \cdot, \cdot) \in \mathcal{W}_{d_0+D,r}$ for (100) to be $O(h^r)$. Let $\mathcal{G} = \{q : \exists z \in \bar{\mathcal{S}} : \tilde{q}(z) = q\}$, which is compact since $\bar{\mathcal{S}}$ is compact. Note that $\Delta^*_{(t)}(\omega \Lambda f, x^0, \tilde{q}) = 0$ whenever $(x^0, \tilde{q}) \notin \mathbb{R}^{d_0} \times \mathcal{G}$. But $\Delta^*_{(t)}(\omega \Lambda f, \cdot, \cdot) \in \mathcal{W}_{d_0+D,r}$ by assumption D and the assumption that $\omega \in \mathcal{W}^*_{2d,r}$. □

Lemma 14 *For any function ω for which $\partial \omega / \partial x^{j1} \in \bar{\mathcal{W}}_{2d,r}$,*

$$\sup_t \left| E\big(T_1^j \omega(X_1, t)\big) + \Delta_{(t)} \left(\frac{\frac{\partial \epsilon^j}{\partial x^{j1}} \omega}{\frac{\partial \tilde{q}^j}{\partial x^{j1}}} \right) \right| = O(h^r), \tag{101}$$

where $\epsilon^j(x) = E(M_1 | X_1 = x) \Psi^j_\bullet(x^j) = \Psi^j_\bullet(x^j)\big(a(x) - a(x_0)\big)/\Delta_\bullet$.

Proof: Let

$$\bar{\omega}(x,t) = -\frac{\partial \frac{\omega \epsilon^j \Lambda f}{\partial \tilde{q}^j / \partial x^{j1}}}{\partial x^{j1}}(x,t)/f(x).$$

Then

$$\begin{aligned} E\big(T_1^j \omega(X_1,t)\big) &= E\Big(K_0(X_1^0, \tilde{q}_1^{-j}) K_0'(\tilde{q}_1^j) \omega(X_1,t) \epsilon_1^j \Lambda_1\Big) \\ &= \int K_0\big(x^0, \tilde{q}^{-j}(z^{-j})\big) K_0'\big(\tilde{q}^j(x^j)\big) \omega(x,t) \epsilon^j(x) \Lambda(z) f(x) dx \\ &= \int K_0\big(x^0, \tilde{q}(z)\big) \bar{\omega}(x,t) f(x) dx = E\big(K_0(X_1^0, \tilde{q}_1) \bar{\omega}(X_1,t)\big). \end{aligned} \tag{102}$$

By lemma 13, the RHS in (102) is $\Delta^*_{(t)}(\bar{\omega} f) + O(h^r)$, uniformly in t. Since $\epsilon^j(x) = 0$ whenever $x^0 = x_0^0, \tilde{q} = 0$,

$$\Delta^*_{(t)}(\bar{\omega} f) = \Delta^*_{(t)} \left(\frac{\partial \epsilon^j}{\partial x^{j1}} \frac{\omega \Lambda f}{\partial \tilde{q}^j / \partial x^{j1}} \right) = \Delta_{(t)} \left(\frac{\partial \epsilon^j}{\partial x^{j1}} \frac{\omega}{\partial \tilde{q}^j / \partial x^{j1}} \right),$$

where the first equality follows from the fact that $\epsilon^j(x) = 0$ whenever $x^0 = x_0^0, \tilde{q}(z) = 0$ and the second equality from lemma 12. □

A.4 Bias

Recall the definition of F_i in (78).

Lemma 15

$$EF_1 = O(h^r), \tag{103}$$

$$E\big(T_1^j(\mu_1^{*j} - q_1^j)\big) = O(h^r), \quad j = 1, \ldots, D. \tag{104}$$

Proof: (103) follows directly from lemma 1, hence I concentrate on (104). Note that from (56) it follows that $\mu_1^{*j} - q_1^j$ is a finite sum of terms $v^j(X_1)$ with

$$v^j(x^j) = \big(E\hat{\beta}^j_{AB;\omega\omega^*}(x^j, x_0^j) - \alpha^j_{AB;\omega\omega^*}(x^j, x_0^j)\big)\bar{\Psi}^j(x^j) \text{ or}$$
$$v^j(x^j) = \big(E\hat{\beta}^j_{AB;\omega\omega^*}(x_0^j, x^j) - \alpha^j_{AB;\omega\omega^*}(x_0^j, x^j)\big)\bar{\Psi}^j(x^j),$$

with $\bar{\Psi}^j(x^j)$ one of $1, \pi^j(x^j)/\delta^j(x^j), 1/\delta^j(x^j)$. Assume without loss of generality (as before) that $\omega = \omega^* = 1$. I will establish the result for the first choice of v^j, where the result for the second possibility follows similarly. Let $\mathcal{A}(x) = E(M_1|X_1 = x)\Lambda(x)$. Then it suffices to show that

$$E\big(K_0(X_1^0, \tilde{q}_1^{-j})K_0'(\tilde{q}_1^j)\mathcal{A}(X_1)v^j(X_1^j)\big) = O(h^r). \tag{105}$$

The LHS in (105) is by integration by parts[22] equal to (omitting arguments)

$$-\int K_0 \left(\frac{\frac{\partial(\mathcal{A}f)}{\partial x^{j1}}v^j + \mathcal{A}f\frac{\partial v^j}{\partial x^{j1}}}{\frac{\partial \tilde{q}^j}{\partial x^{j1}}} - \frac{\mathcal{A}f v^j \frac{\partial^2 \tilde{q}^j}{\partial (x^{j1})^2}}{\left(\frac{\partial \tilde{q}^j}{\partial x^{j1}}\right)^2} \right).$$

It hence suffices to show that

$$\sup_{x^j \in \bar{\mathcal{S}}^j} |v^j(x^j)| = O(h^r), \quad \sup_{x^j \in \bar{\mathcal{S}}^j} \left| \frac{\partial v^j}{\partial x^{j1}}(x^j) \right| = O(h^r),$$

or alternatively that

$$\sup_{x^j \in \bar{\mathcal{S}}^j} |E\hat{\beta}^j_{AB}(x^j, x_0^j) - \alpha^j_{AB}(x^j, x_0^j)| = O(h^r), \quad \sup_{x^j \in \bar{\mathcal{S}}^j} \left| E\left(\frac{\partial \hat{\beta}^j_{AB}}{\partial x^{j1}}(x^j, x_0^j) - \frac{\partial \alpha^j_{AB}}{\partial x^{j1}}(x^j, x_0^j) \right) \right| = O(h^r). \tag{106}$$

The first condition in (106) is (62) and is hence satisfied. For the second condition in (106) it is sufficient to show that

$$\sup_{x \in \bar{\mathcal{S}}} \left| E\frac{\partial \hat{\eta}_A}{\partial x^{j1}}(x) - \frac{\partial \eta_A}{\partial x^{j1}}(x) \right| = O(h^r).$$

Let $\mathcal{A}^*(x) = E(A_1|X_1 = x)f(x)$. Then by integration by parts

$$\int \frac{\partial K}{\partial x^{j1}}(x - t)\iota^*(t)dt - \frac{\partial \iota^*}{\partial x^{j1}}(x) = \int K(x - t)\frac{\partial \iota^*}{\partial x^{j1}}(t)dt - \frac{\partial \iota^*}{\partial x^{j1}}(x).$$

Apply lemma 1. □

[22] Similar to the first few steps in lemma 14.

A.5 Asymptotic Normality

I now derive the limiting distribution of ρ_n^{-1} times the sum of (83) and (84). Let for $i = 1, \ldots, n$, $u = 0, i$,

$$U_t^j \;\; = K_{t0}^j \lambda_{t0}^j \Delta_{\bullet ft}^j (Y_t - a_{t0}^j), \;\; W_{tiu}^j \;\; = T_i^j \Psi_{\bullet i}^j G_{t0i}^j (Y_t - a_{tu}^j), \tag{107}$$

such that the sum of (83) and (84) is

$$\begin{aligned} & n^{-1} \sum_{i=1}^{n} \left(F_i - EF_1 + \sum_{j=1}^{D} T_i^j \Psi_{\bullet i}^j (\hat{q}_{Ii}^{\bullet j} - \mu_{Ii}^{\bullet j}) \right) \\ = \;\; & n^{-1} \sum_{t=1}^{n} \left(F_i - EF_1 + \sum_{j=1}^{D} (W_{tiu}^j - \mu_{Wiu}^j) \right), \end{aligned} \tag{108}$$

where $\mu_{Wiu}^j = \Psi_{\bullet i}^j \mu_{Iiu}^j$. Since (108) is an i.i.d. sum of random variables, to establish the asymptotic variance, I look at the variances and covariances of the summands in (108).

Lemma 16

$$VF_1 = \kappa^{d_0+D} h^{-d_0-D} \mathcal{V}_{\bullet}^0 + o(h^{-d_0-D}).$$

Proof: Recall from lemma 15 that $EF_1 = o(h^r)$. I hence need to show that

$$EF_1^2 = E\big(K_0^2(X_1^0, \tilde{q}_1) M_1^2 \Lambda_1^2\big) = \kappa^{d_0+D} h^{-d_0-D} \mathcal{V}_{\bullet}^0 + o(h^{-d_0-D}). \tag{109}$$

Note that $K^2 = h^{-2(d_0+D)} k^2 = h^{-2(d_0+D)} \kappa^{d_0+D} k^* = h^{-d_0-D} \kappa^{d_0+D} K^*$, with $k^* = k/\kappa^{d_0+D}$ a second order kernel (i.e. for this kernel $r = 2$). I then need to show that

$$E\big(K_0^*(X_1^0, \tilde{q}_1) M_1^2 \Lambda_1^2\big) = \Delta_{\bullet}(\sigma^2 \Lambda)/\Delta_{\bullet}^2 + o(1). \tag{110}$$

The LHS in (110) equals

$$E\big(K_0^*(X_1^0, \tilde{q}_1) \omega(X_1) \Lambda_1\big), \tag{111}$$

where $\omega(x) = \Big(\sigma^2(x) + \big(a(x) - a(x_0)\big)^2\Big)\Lambda(x)/\Delta_{\bullet}^2$, since $M_1 = \big(Y_1 - a(x_0)\big)/\Delta_{\bullet}$ by definition. By lemmas 12 and 13, (111) is

$$\Delta_{\bullet}(\omega) + O(h^2) = \Delta_{\bullet}(\sigma^2 \Lambda)/\Delta_{\bullet}^2 + o(1),$$

because $a(x) = a(x_0)$ whenever $x^0 = x_0^0, g(z) = g(z_0)$. □

Let E_i denote expectation treating X_i, Y_i as constant.

Lemma 17

$$E(E_1 W^j_{12u} - U^j_1)^2 = O(h^{2r-d_j}), \quad j = 1, \ldots, D; u = 0, 2$$

Proof: I show the result for $u = 2$, which involves one more step than $u = 0$. Let $\nu^j_{it} = a^j_{it} f^j_{it}$. Note that

$$\begin{aligned} W^j_{122} &= T^j_2 \Psi^j_{\bullet 2} G^j_{102}(Y_1 - a^j_{12}) = K^j_{10}\lambda^j_{10}\big(T^j_2 \Psi^j_{\bullet 2}\lambda^j_{12} f^j_{12}(Y_1 - a^j_{12})\big) \\ &= K^j_{10}\lambda^j_{10}\big((T^j_2 \Psi^j_{\bullet 2}\lambda^j_{12} f^j_{12} - \Delta^j_{\bullet f1})Y_1 - (T^j_2 \Psi^j_{\bullet 2}\lambda^j_{12}\nu^j_{12} - \Delta^j_{\bullet \nu 1})\big) \qquad (112) \\ &\quad + K^j_{10}\lambda^j_{10}(\Delta^j_{\bullet f1} Y_1 - \Delta^j_{\bullet \nu 1}). \qquad (113) \end{aligned}$$

Note first that since $a(x) = a^j(x^{-j}, x^j_0)$ for all x for which $g^j(x^j) = g^0(x^j_0)$, $\Delta^j_{\bullet \nu 1} = a^j_{10}\Delta^j_{\bullet f1}$ and hence (113) is U^j_1. But since

$$\begin{aligned} &E\big(T^j_2 \Psi^j_{\bullet 2}\lambda^j(x^{-j}, X^j_2) f^j(x^{-j}, X^j_2) - \Delta^j_{\bullet f}(x^{-j})\big) = O(h^r), \\ &E\big(T^j_2 \Psi^j_{\bullet 2}\lambda^j(x^{-j}, X^j_2)\nu^j(x^{-j}, X^j_2) - \Delta^j_{\bullet \nu}(x^{-j})\big) = O(h^r), \end{aligned}$$

uniformly in x^{-j} by lemma 14, the squared expectation of (112) is

$$E\Big(K^j_{10}\lambda^j_{10} E_1\big((T^j_2 \Psi^j_{\bullet 2}\lambda^j_{12} f^j_{12} - \Delta^j_{\bullet f1})Y_1 - (T^j_2 \Psi^j_{\bullet 2}\lambda^j_{12}\nu^j_{12} - \Delta^j_{\bullet \nu 1})\big)\Big)^2 = O(h^{2r-d_j}),$$

by lemma 2. □

Lemma 18 *For any $j, j^* = 1, \ldots, D$, $u, u^* = 0, 2$,*

$$\mathrm{Cov}(E_1 W^j_{12u}, E_1 W^{j^*}_{12u^*}) = \begin{cases} O(1), & j \neq j^*, \\ h^{-d_j}\kappa^{d_j}\mathcal{V}^j_\bullet + o(n\rho^2_n), & j = j^*. \end{cases} \qquad (114)$$

Proof: By lemmas 17 and 1, $E(W^j_{12u}) = EU^j_1 + O(h^r) = O(h^r)$. Therefore, $\mathrm{Cov}(E_1 W^j_{12u}, E_1 W^{j^*}_{12u^*}) = E(E_1 W^j_{12u} E_1 W^{j^*}_{13u^*}) + O(h^{2r})$. But

$$\begin{aligned} &\big|E(E_1 W^j_{12u} E_1 W^{j^*}_{13u^*}) - E(U^j_1 U^{j^*}_1)\big| \\ &\quad \leq \Big|E\big((E_1 W^j_{12u} - U^j_1)(E_1 W^{j^*}_{12u^*} - U^{j^*}_1)\big)\Big| + \Big|E\big(U^j_1(E_1 W^{j^*}_{12u^*} - U^{j^*}_1)\big)\Big| + \Big|E\big((E_1 W^j_{12u} - U^j_1)U^{j^*}_1\big)\Big| \\ &\quad = O(h^{2r-d_j}) + O(h^{r-d_j}) = O(h^{r-d_j}) = o(1), \end{aligned}$$

by assumption G, which follows from the Schwarz inequality, lemma 17 and the fact that $E(U_1^j)^2 = O(h^{-d_j})$ by lemma 2. Finally, for $j^* \neq j$, $E(U_1^j U_1^{j^*}) = O(1)$ by lemma 2 and

$$E(U_1^j)^2 = E\Big((K_{10}^j \lambda_{10}^j \Delta_{\bullet f1}^j)^2 (Y_1 - a_{10}^j)^2\Big) = E\Big((K_{10}^j \lambda_{10}^j \Delta_{\bullet f1}^j)^2 \big(\sigma_1^2 + (a_1 - a_{10}^j)^2\big)\Big) = h^{-d_j} \kappa^{d_j} \mathcal{V}_\bullet^j + O(h^{2-d_j}),$$

by lemma 1. But $h^{2-d_j} = o(h^{-d_m}) = o(n\rho_n^2)$ by assumption G. □

Lemma 19 *For any $j = 1, \ldots, D$, $u = 0, 2$,*

$$\mathrm{Cov}(F_1, E_1 W_{12u}^j) = o(n\rho_n^2).$$

Proof: First, since the means of the quantities whose covariance is taken are $O(h^r)$ by lemma 15,

$$\mathrm{Cov}(F_1, E_1 W_{12u}^j) = E(F_1 W_{12u}^j) + O(h^{2r}). \tag{115}$$

With U_1^j as defined in (107), the first term on the RHS in (115) is

$$E\Big(F_1 (E_1 W_{12u}^j - U_1^j)\Big) + E\big(F_1 U_1^j\big). \tag{116}$$

Lemmas 17 and 16 together with the Schwarz inequality imply that the first term in (116) is $O(h^{r-(d_0+D+d_j)/2}) = o(h^{-d_m}) = o(n\rho_n^2)$. The absolute value of the second term in (116) is

$$\Big|E\Big(K_0(X_1^0, \tilde{q}_1)\Lambda_1 M_1 K_{10}^j \lambda_{10}^j \Delta_{\bullet f}^j (Y_1 - a_{10}^j)\Big)\Big| \\ \leq Ch^{-1} E\Big|K_0(X_1^0, \tilde{q}_1^{-j}, X_1^j)\Lambda_1 M_1 \lambda_{10}^j \Delta_{\bullet f}^j (Y_1 - a_{10}^j)\Big| = O(h^{-1}) = o(n\rho_n^2),$$

for some $C < \infty$ by lemma 2 since $K_0(\tilde{q}^j)$ is bounded by Ch^{-1} for some fixed $C < \infty$ by assumption F. □

Lemma 20 *Let μ_{Ii0}^j, μ_{Iii}^j be as defined earlier in this appendix. Then, for $u = 0, i$,*

$$\rho_n^{-1} n^{-1} \sum_{i=1}^n \left(\big(F_i - EF_1\big) + \sum_{j=1}^D T_i^j \Psi_{\bullet i}^j (\hat{q}_{Iiu}^j - \mu_{Iiu}^j) \right) \xrightarrow{\mathcal{L}} N(0, \mathcal{V}_\bullet). \tag{117}$$

Proof: I show the result for $u = i$; showing the results for $u = 0, i$ simultaneously only complicates notation. The LHS in (117) is by (107) equal to

$$\rho_n^{-1} n^{-2} \sum_{t=1}^n \sum_{i=1}^n \left((F_t - EF_1) + \sum_{j=1}^D (W_{tii}^j - E_i W_{tii}^j) \right), \tag{118}$$

which is a V–statistic with asymmetric kernel. In line with lemma 3, denote the expression in large brackets in (118) by ζ_{it} and set $c_n = \rho_n$. Clearly, $E_i\zeta_{it} = 0$ a.s. and hence $E\zeta_{12} = 0$. By lemma 2,

$$\begin{aligned} E|F_1| &= O(1), \\ E|W^j_{111}| &= E|T^j_1 \Psi^j_{\bullet 1} K^j_{10} \lambda^j_{10} \lambda_1 f_1(Y_1 - a_1)| \\ &\leq C_1 h^{-2} E\big|K^0_{10} K^j_{10}(|Y_1| + 1)\big| = O(h^{-2}), \end{aligned}$$

for some $C_1 > 0$ because $|K'| = h^{-2}|k'|$ and $|k'|$ is bounded. Hence $E|\zeta_{11}| = O(h^{-2}) = o(n^{1/2}h^{-d_m}) = o(n\rho_n)$. Further, by (109) and again by lemma 2,

$$\begin{aligned} EF_1^2 &= O(h^{-d_0-D}), \\ E(W^j_{122})^2 &= E\big(T^j_2 \Psi^j_{\bullet 2} K^j_{10} \lambda^j_{10} \lambda^j_{12}(f^j_{12} Y_1 - \nu^j_{12})\big)^2 \\ &\leq C_2 E(T^j_2)^2 E\big(K^j_{10} \lambda^j_{10}(|Y_1| + 1)\big)^2 = O(h^{-d_0-D-3}) O(h^{-d_j}) = O(h^{-d_0-D-3-d_j}), \end{aligned}$$

for some $C_2 < \infty$. Hence $E\zeta^2_{12} = O(h^{-d_0-D-3-d_j}) = O(h^{-2d_m-3}) = o(nh^{-d_m}) = o(n^2\rho_n^2)$ by assumption G. Hence lemma 3 implies that

$$\rho_n^{-1} n^{-2} \sum_{t=1}^{n} \sum_{i=1}^{n} \zeta_{it} = \rho_n^{-1} n^{-1} \sum_{t=1}^{n} E_t \zeta_{it} + o(1).$$

All that remains to be done is to establish the limiting variance. Now, $E\big(E_1(\zeta_{21})\big)^2$ is

$$\begin{aligned} E(\zeta_{21}\zeta_{31}) &= VF_1 && (119) \\ &+ \sum_{j=1}^{D} \mathrm{Cov}(F_1, E_1 W^j_{122}) && (120) \\ &+ \sum_{j=1}^{D} V E_1 W^j_{122} && (121) \\ &+ \sum_{j=1}^{D} \sum_{j^* \neq j}^{D} \mathrm{Cov}(E_1 W^j_{122}, E_1 W^{j^*}_{122}). && (122) \end{aligned}$$

By lemma 16, (119) is $h^{-d_0-D}\kappa^{d_0+D}\mathcal{V}^0_\bullet + o(n\rho_n^2)$. Lemma 19 implies that (120) is $o(n\rho_n^2)$. Finally, lemma 18 establishes that (122) is $o(n\rho_n^2)$ and that (121) is $\sum_{j=1}^{D} h^{-d_j}\kappa^{d_j}\mathcal{V}^j_\bullet + o(n\rho_n^2)$. □

A.6 Negligble Terms

Lemma 21 establishes that (85) is asymptotically negligble, lemma 22 does so for (86) and lemma 23 for (87).

Lemma 21 *Let* $\mu^j_{IIi0} = E_i\hat{q}^j_{IIi0}$, $\mu^j_{IIii} = E_i\hat{q}^j_{IIii}$. *Then for* $u = 0, i$,

$$n^{-1}\sum_{i=1}^{n} T^j_i \Psi^j_{\bullet i}(\hat{q}^j_{IIiu} - \mu^j_{IIiu}) = o_p(\rho_n). \tag{123}$$

Proof: I show the result for $u = 0$ where the result for $u = i$ follows similarly. Let $\Upsilon_t = \lambda^j_{t0}(f^j_{t0}Y_t - \nu^j_{t0})$ and choose $\zeta_{it} = T^j_i\Psi^j_{\bullet i}\big(K^j_{ti}\lambda^j_{ti}\Upsilon_t - E_i(K^j_{ti}\lambda^j_{ti}\Upsilon_t)\big)$ in lemma 3, such that the LHS in (123) is $n^{-2}\sum_{i=1}^n\sum_{t=1}^n \zeta_{it}$. Clearly, $E\zeta_{12} = 0$. The procedure to establish that $E|\zeta_{11}| = o(n\rho_n)$ and $E\zeta^2_{12} = o(n^2\rho^2_n)$ is essentially the same as in lemma 20 and is not repeated here. I now show that $E(\zeta_{21}\zeta_{31}) = o(n\rho^2_n)$, which is sufficient for (123). Note that

$$E(\zeta_{21}\zeta_{31}) = E(E_1\zeta_{21})^2 \leq E\big(\Upsilon_1 E_1(T^j_2\Psi^j_{\bullet 2}K^j_{12}\lambda^j_{12})\big)^2 = E\big(T^j_2T^j_3Q(X^j_2, X^j_3)\big), \tag{124}$$

with (setting $\Upsilon^*(x) = E(\Upsilon_1|X_1 = x)$),

$$\begin{aligned}
Q(x^j, \tilde{x}^j) &= E\big(K_1(x^j)K_1(\tilde{x}^j)\Upsilon^2_1\lambda^j(X^{-j}_1, x^j)\lambda^j(X^{-j}_1, \tilde{x}^j)\big) \\
&= \int K(\bar{x}^j - x^j)K(\bar{x}^j - \tilde{x}^j)\Upsilon^*(\bar{x})\lambda^j(\bar{x}^{-j}, x^j)\lambda^j(\bar{x}^{-j}, \tilde{x}^j)d\bar{x} \\
&= h^{-d_j}\int k(u)k\left(\frac{\tilde{x}^j - x^j}{h} + u\right)\Upsilon^{*j}(\bar{x}^{-j}, \tilde{x}^j + hu)\lambda^j(\bar{x}^{-j}, x^j)\lambda^j(\bar{x}^{-j}, \tilde{x}^j)d\bar{x}^{-j}du,
\end{aligned}$$

where the last equality follows with the substitution of $u = (\bar{x}^j - \tilde{x}^j)/h$. Let $\bar{Q}(x) = E(\Lambda_1M_1|X_1 = x)f(x)$. Since λ is bounded and $\sup_{x^j}\int \Upsilon^{*j}(x^{-j}, x^j)dx^{-j} < \infty$ by assumption D, the RHS in (124) is bounded by some constant times

$$\begin{aligned}
&h^{-d_j}\sup_u\left|E\left(T^j_2T^j_3k\left(\frac{X^j_2 - X^j_3}{h} + u\right)\right)\right| \\
&= h^{-d_j}\sup_u\left|E\left(K_0(X^0_2, \tilde{q}^{-j}_2)K_0(X^0_3, \tilde{q}^{-j}_3)K'_0(\tilde{q}^j_2)K'_0(\tilde{q}^j_3)\Lambda_2\Lambda_3M_2M_3k\left(\frac{X^j_2 - X^j_3}{h} + u\right)\right)\right| \\
&= h^{-d_j}\sup_u\Bigg|\int K_0\big(x^0, \tilde{q}^{-j}(z^{-j})\big)K_0\big(\tilde{x}^0, \tilde{q}^{-j}(\tilde{z}^{-j})\big)K'_0\big(\tilde{q}^j(x^j)\big)K'_0\big(\tilde{q}^j(\tilde{x}^j)\big) \\
&\quad\times k\left(\frac{x^j - \tilde{x}^j}{h} + u\right)\bar{Q}(x)\bar{Q}(\tilde{x})dxd\tilde{x}\Bigg| \\
&= \sup_u\Bigg|\int K_0\big(x^0, \tilde{q}^{-j}(z^{-j})\big)K_0\big(\tilde{x}^0, \tilde{q}^{-j}(\tilde{z}^{-j})\big)K'_0\big(\tilde{q}^j(x^j)\big)K'_0\Big(\tilde{q}^j\big(\tilde{x}^j + h(u+v)\big)\Big) \\
&\times k(v)\bar{Q}(x)\bar{Q}^j\big(\tilde{x}^{-j}, x^j + h(u+v)\big)dxd\tilde{x}^{-j}dv\Bigg|, \\
&\leq \sup_u\int\Bigg|K_0\big(\tilde{x}^0, \tilde{q}^j(x^j)\big)K'_0\Big(\tilde{q}^j\big(\tilde{x}^j + h(u+v)\big)\Big)k(v) \\
&\qquad\times\sup_{u,v}\left|\int K_0\big(x^0, \tilde{q}^{-j}(z^{-j})\big)K'_0\big(\tilde{q}^j(x^j)\big)\bar{Q}(x)\bar{Q}^j\big(\tilde{x}^{-j}, x^j + h(u+v)\big)dx\right|\Bigg|d\tilde{x}^{-j}dv
\end{aligned} \tag{125}$$

by substitution of $v = (x^j - \tilde{x}^j)/h + u$. The inner supremand is $O(1)$, uniformly in u, v. Thus, (125) is

$$O(1)\sup_u \int \left| K_0\big(\tilde{x}^0, \tilde{q}^j(x^j)\big) K_0'\Big(\tilde{q}^j\big(\tilde{x}^j + h(u+v)\big)\Big) k(v) \right| d\tilde{x}^{-j} dv = O(h^{-1}) = o(h^{-d_m}) = o(n\rho_n^2),$$

by lemma 2. □

Lemma 22

$$n^{-1}\sum_{i=1}^n T_i^j(\mu_i^{*j} - q_i^j) - E\big(T_1^j(\mu_1^{*j} - q_1^j)\big) = o_p(\rho_n), \quad j = 1, \ldots, D.$$

Proof: In the proof of lemma 15, I showed that $\sup_{x^j \in \bar{\mathcal{S}}^j} |\mu^{*j}(x^j) - q^j(x^j)| = O(h^r)$. By the Schwarz inequality and lemma 2,

$$E\big(T_1^j(\mu_1^{*j} - q_1^j)\big)^2 \leq E\big(K_0(X_1^0, \tilde{q}_1^{-j}) K_0'(\tilde{q}_1^j) M_1\big)^2 \sup_x \Big(\Lambda(x)\big(\mu^{*j}(x^j) - q^j(x^j)\big)\Big)^2$$
$$= O(h^{-d_0 - D - 3}) O(h^{2r}) = O(h^{2r - d_m - 3}) = o(\rho_n). \quad \square$$

Lemma 23

$$n^{-1}\sum_{i=1}^n T_i^j \Psi_{\bullet i}^j \pi_i^j (\hat{\delta}_i^{\bullet j} - \mu_{\delta i}^{\bullet j}) = o_p(\rho_n), \quad j = 1, \ldots, D. \tag{126}$$

Proof: Choose $\zeta_{it} = T_i^j \Psi_{\bullet i}^j \pi_i^j (G_{t0i}^j - E_i G_{t0i}^j + G_{ti0}^j - E_i G_{ti0}^j)$ in lemma 3, such that the LHS in (126) is $n^{-2}\sum_{t=1}^n \sum_{i=1}^n \zeta_{it}$. Using the same steps as in the proof of lemma 20, it can be shown that $E|\zeta_{11}| = o_p(n\rho_n)$ and $E\zeta_{12}^2 = o_p(n^2\rho_n^2)$. Here I show that $E(\zeta_{21}\zeta_{31}) = o(\rho_n)$, which is sufficient since $E_1\zeta_{12} = 0$ a.s.. Note that

$$E(\zeta_{21}\zeta_{31}) \leq 2E\Big(E_1\big(T_2^j \pi_2^j (G_{102}^j - E_2 G_{102}^j)\big)/\delta_2^j\Big)^2 + 2E\Big(E_1\big(T_2^j \pi_2^j (G_{120}^j - E_2 G_{120}^j)/\delta_2^j\big)\Big)^2. \tag{127}$$

I show that the first RHS term in (127) is $o(n\rho_n^2)$, where the second term follows similarly. Thus,

$$E\Big(E_1\big(T_2^j \pi_2^j \Psi_{\bullet 2}^j (G_{102}^j - E_2 G_{102}^j)\big)\Big)^2 = VE_1(T_2^j \Psi_{\bullet 2}^j \pi_2^j G_{102}^j) \leq E\big(E_1(T_2^j \Psi_{\bullet 2}^j \pi_2^j G_{102}^j)\big)^2$$
$$= E\Big(K_{10}^j \lambda_{10}^j E_1(T_2^j \Psi_{\bullet 2}^j \pi_2^j \lambda_{12}^j f_{12}^j)\Big)^2 = o(h^{2r - d_j}),$$

since the inner expectation is $o(h^r)$ a.s. by lemma 14. □

A.7 Proof of Theorem 1

Consider the expansion of $\hat{a}_{S\bullet}(x_0) - a(x_0)$ in (83)–(88). Lemmas 21–23 establish that (85)–(87) are $o_p(\rho_n)$. In lemma 15 I showed that the asymptotic bias is $O(h^r)$, which is $o(\rho_n)$ if $\ell = 0$ and $O(\rho_n)$, otherwise by the definition of ℓ. By lemma 20, ρ_n^{-1} times the sum of (83) and (84) has a limiting $N(0, \mathcal{V}_\bullet)$–distribution. □

A.8 Lemmas for Theorem 2

Lemma 24 is an auxiliary result, which is used in the remaining lemmas.

Lemma 24 *For any sequence of functions $\{\xi_i\}$ for which $\sup_x |\xi_i(x)|$ has moments greater than one and for any $j = 1, \ldots, D$,*

$$\sup_x n^{-1} \sum_{i=1}^n \left| K_0(X_i^0)\big(K_0(\hat{q}_i^{-j})K_0'(\hat{q}_i^j) - K_0(\tilde{q}_i^{-j})K_0'(\tilde{q}_i^j)\big)\xi_i(x)\Lambda_i \right| = O_p(\rho_n h^{-2}\log n) = o_p(1).$$

Proof: The proof is close to that of lemma 8 and the steps below are essentially a shorter repetition of lemma 8. To understand the steps below it is helpful to read lemma 8 first. I will show that

$$\sup_x n^{-1} \sum_{i=1}^n \left| K_0(X_i^0)\big(K_0(\hat{q}_i^{-j})K_0'(\hat{q}_i^j) - K_0(q_i^{-j})K_0'(q_i^j)\big)\xi_i(x)\Lambda_i \right| = O_p(\rho_n h^{-2}\log n), \tag{128}$$

$$\sup_x n^{-1} \sum_{i=1}^n \left| K_0(X_i^0)\big(K_0(q_i^{-j})K_0'(q_i^j) - K_0(\tilde{q}_i^{-j})K_0'(\tilde{q}_i^j)\big)\xi_i(x)\Lambda_i \right| = O_p(\rho_n h^{-2}\log n), \tag{129}$$

First, (129) follows from (74). Now (128). Let $\varsigma, \iota_\varsigma$ be defined as in lemma 8 and let ς^* be ς except that $\varsigma_j^* = \varsigma_j - 1$. Applying the mean value theorem, like in (71), the LHS in (128) is

$$\sup_x \sum_{2\le|\varsigma|\le\Phi} \sum_{i=1}^n \left| K_0(X_i^0)\xi(x)\Lambda_i \prod_{j^*=1}^D \left(\frac{(\hat{q}_i^{j^*} - q_i^{j^*})^{\varsigma_{j^*}^*}}{\varsigma_{j^*}^*!} \right) \prod_{j^*=1}^D \left(\big(K_0^{(\varsigma_{j^*})}(q_i^{j^*})\big)^{1-\iota_{\varsigma j^*}} \big(K_0^{(\Phi)}(\cdot)\big)^{\iota_{\varsigma j^*}} \right) \right|, \tag{130}$$

where $(\cdot)$ denotes some quantity between q_i^j and $\hat{q}_i^j$. Following the same steps as in lemma 8 after (71) and using the same notation, the LHS in (130) is

$$O_p\left(\psi_n^{|\varsigma|-1} h^{-|\varsigma|-|\iota_\varsigma|} \sup_x n^{-1} \sum_{i=1}^n \left| K_0(X_i^0)\xi_i(x)\Lambda_i \prod_{j^*=1}^D \big(\bar{K}_{(\varsigma_{j^*})0}(q_i^{j^*})\big)^{1-\iota_{\varsigma j^*}} \right| \right) = O_p(\psi_n^{|\varsigma|-1} h^{-|\varsigma|-|\iota_\varsigma|}). \tag{131}$$

Note that since $\Phi > 3$ and $|\iota_\varsigma| = 1$ only if $|\varsigma| = \Phi$,

$$\psi_n^{|\varsigma|-1} h^{-|\varsigma|-|\iota_\varsigma|} = O\big(\psi_n h^{-2} + (\psi_n h^{-2})^{\Phi-1} h^{\Phi-3}\big).$$

But $\psi_n h^{-2} = \rho_n h^{-2}\log n = O(n^{-1/2}h^{d_m/2-2}\log n) = o(1)$ by assumption G. □

Let $P_t^j(x) = K_0(X_t^0, \tilde{q}_t^{-j})K_0'(\tilde{q}_t^j)\Lambda_t\lambda^j(x^{-j}, X_t^j)$. Then lemma 24 implies that

$$\sup_x |\hat{P}_t^j(x) - P_t^j(x)| = o_p(1), \quad j = 1, \ldots, D. \tag{132}$$

Let for $\omega = f, \nu$,

$$R_\omega^j(x) = E\big(T_1^j \Psi_{\bullet 1}^j (\lambda^j \cdot \omega^j)(x^{-j}, X_1^j)\big). \tag{133}$$

with $\hat{P}^j$ as defined in (28).

Lemma 25 *Let ω be one of f, ν. Then*

$$\sup_x \left| n^{-1} \sum_{t=1}^n \hat{R}_{\omega t}^j(x) - R_\omega^j(x) \right| = o_p(1). \tag{134}$$

Proof: Recall that $M_t = (Y_t - a_0)/\Delta_\bullet$. I establish (134) in two steps:

$$\sup_x \left| n^{-1} \sum_{t=1}^n \Big(\hat{P}_t^j(x) \hat{\Psi}_{\bullet t}^j \frac{Y_t - \hat{a}(x_0)}{\hat{\Delta}_\bullet} \hat{\omega}^j(x^{-j}, X_t^j) - P_t^j(x) \Psi_{\bullet t}^j M_t \omega^j(x^{-j}, X_t^j) \Big) \right| = o_p(1), \tag{135}$$

$$\sup_x \left| n^{-1} \sum_{t=1}^n P_t^j(x) \Psi_{\bullet t}^j M_t \omega^j(x^{-j}, X_t^j) - E\big(T_1^j \Psi_{\bullet 1}^j (\lambda^j \cdot \omega^j)(x^{-j}, X_1^j)\big) \right| = o_p(1). \tag{136}$$

First (135). Since

$$\frac{1}{\hat{\Delta}_\bullet} - \frac{1}{\Delta_\bullet} = \frac{\Delta_\bullet - \hat{\Delta}_\bullet}{\Delta_\bullet} \frac{1}{\hat{\Delta}_\bullet}, \quad \frac{Y_t - \hat{a}_{S\bullet}(x_0)}{\Delta_\bullet} - M_t = \frac{a(x_0) - \hat{a}_{S\bullet}(x_0)}{\Delta_\bullet},$$

the LHS in (135) is

$$\sup_x \left| n^{-1} \sum_{t=1}^n \Big(\hat{P}_t^j(x) \hat{\Psi}_{\bullet t}^j \frac{Y_t - \hat{a}(x_0)}{\hat{\Delta}_\bullet} \hat{\omega}^j(x^{-j}, X_t^j) - P_t^j(x) \Psi_{\bullet t}^j M_t \omega^j(x^{-j}, X_t^j) \Big) \right| \tag{137}$$

$$\leq \left| \frac{\Delta_\bullet - \hat{\Delta}_\bullet}{\Delta_\bullet} \right| \sup_x \left| n^{-1} \sum_{t=1}^n \Big(\hat{P}_t^j(x) \hat{\Psi}_{\bullet t}^j \frac{Y_t - \hat{a}(x_0)}{\hat{\Delta}_\bullet} \hat{\omega}^j(x^{-j}, X_t^j) - P_t^j(x) \Psi_{\bullet t}^j M_t \omega^j(x^{-j}, X_t^j) \Big) \right| \tag{138}$$

$$+ \sup_x \left| n^{-1} \sum_{t=1}^n M_t \Big(\hat{P}_t^j(x) \hat{\Psi}_{\bullet t}^j \hat{\omega}^j(x^{-j}, X_t^j) - P_t^j(x) \Psi_{\bullet t}^j \omega^j(x^{-j}, X_t^j) \Big) \right| \tag{139}$$

$$+ \left| \frac{a(x_0) - \hat{a}(x_0)}{\Delta_\bullet} \right| \sup_x \left| n^{-1} \sum_{t=1}^n \Big(\hat{P}_t^j(x) \hat{\Psi}_{\bullet t}^j \hat{\omega}^j(x^{-j}, X_t^j) - P_t^j(x) \Psi_{\bullet t}^j \omega^j(x^{-j}, X_t^j) \Big) \right| \tag{140}$$

By lemma 9, (138) is of lesser order than (137). Expressions (139) and (140) are also $o_p(1)$, which I now show. Let M_t^* be one of $M_t, 1$, Then

$$\sup_x \Big| n^{-1} \sum_{t=1}^n M_t^* \big(\hat{P}_t^j(x) \hat{\Psi}_{\bullet t}^j \hat{\omega}^j(x^{-j}, X_t^j) - P_t^j(x) \Psi_{\bullet t}^j \omega^j(x^{-j}, X_t^j) \big) \Big| \tag{141}$$

$$\leq \sup_x \Big| n^{-1} \sum_{t=1}^n M_t^* \big(\hat{P}_t^j(x) - P_t^j(x) \big) \big(\hat{\omega}^j(x^{-j}, X_t^j) \hat{\Psi}_{\bullet t}^j - \omega^j(x^{-j}, X_t^j) \Psi_{\bullet t}^j \big) \Big| \tag{142}$$

$$+ \sup_x n^{-1} \sum_{t=1}^n \Big| M_t^* \big(\hat{P}_t^j(x) - P_t^j(x) \big) \Psi_{\bullet t}^j \omega^j(x^{-j}, X_t^j) \Big| \tag{143}$$

$$+ \sup_x \Big| n^{-1} \sum_{t=1}^n M_t^* P_t^j(x) \big(\hat{\omega}^j(x^{-j}, X_t^j) \hat{\Psi}_{\bullet t}^j - \omega^j(x^{-j}, X_t^j) \Psi_{\bullet t}^j \big) \Big|. \tag{144}$$

(142) is bounded by

$$\sup_x n^{-1} \sum_{t=1}^n \Big| M_t^* \big(\hat{P}_t^j(x) - P_t^j(x) \big) \Big| \sup_x \max_{t=1,\dots,n} \big| \hat{\omega}^j(x^{-j}, X_t^j) - \omega^j(x^{-j}, X_t^j) \big|, \tag{145}$$

which is $o_p(n^{-1/2} h^{-d/2} \log n)$, because the first factor in (145) is $o_p(1)$ by (132) and because the second factor is $o_p(n^{-1/2} h^{-d/2} \log n)$ since
$\sup_x \big(|\hat{f}(x) - f(x)| + |\hat{\nu}(x) - \nu(x)| \big) = o_p(n^{-1/2} h^{-d/2} \log n)$ by lemma 1 of Horowitz (1999).
Similarly, (143) is $o_p(1)$, again by (132). Further, (144) is bounded by

$$\sup_x n^{-1} \sum_{t=1}^n |M_t^* P_t^j| \sup_x \max_{t=1,\dots,n} \big| \hat{\omega}^j(x^{-j}, X_t^j) \hat{\Psi}_{\bullet t}^j - \omega^j(x^{-j}, X_t^j) \Psi_{\bullet t}^j \big|$$

$$= O_p(h^{-1}) o_p(n^{-1/2} h^{-d/2} \log n) = o_p(n^{-1/2} h^{-d/2-1} \log n) = o_p(1),$$

by assumption G. Hence (147) is $o_p(1)$, and therefore so is (141). So (135) holds.
Now (136). By lemma 1 of Horowitz (1999), (136) is

$$O_p\Big(\sup_x n^{-1/2} \sqrt{E\big(P_1^j(x) \Psi_{\bullet 1}^j M_1 \omega^j(x^{-j}, X_1^j) \big)^2} \log n \Big).$$

But

$$\sup_x E\big(P_1^j(x) \Psi_{\bullet 1}^j M_1 \omega^j(x^{-j}, X_1^j) \big)^2 = \sup_x E\big(T_1^j(x) \Psi_{\bullet 1}^j (\lambda^j \cdot \omega^j)(x^{-j}, X_1^j) \big)^2$$

$$= O(h^{-d_0 - D - 2}) = O(h^{-d_m - 2}),$$

by lemma 2. Finally, $n^{-1} h^{-d_m - 2} \log^2 n = o(1)$ by assumption G. □

A.9 Proof of Theorem 2

I first show that $\hat{\mathcal{V}}_\bullet^0 = \mathcal{V}_\bullet^0 + o_p(1)$ and further down that $\hat{\mathcal{V}}_\bullet^j = \mathcal{V}_\bullet^j + o_p(h^{d_j - d_m})$, $j = 1, \dots, D$. Consider (31). Let $\vartheta(x) = E(Y_1^2 | X_1 = x)$. By lemma 9,

$$
\begin{aligned}
n^{-1} \sum_{i=1}^n K_0(X_i^0, \hat{q}_i) \Lambda_i^2 &= \Delta_\bullet(\Lambda) + o_p(1), \\
n^{-1} \sum_{i=1}^n K_0(X_i^0, \hat{q}_i) \Lambda_i^2 Y_i &= \Delta_\bullet(\Lambda a) + o_p(1), \\
n^{-1} \sum_{i=1}^n K_0(X_i^0, \hat{q}_i) \Lambda_i^2 Y_i^2 &= \Delta_\bullet(\Lambda \vartheta) + o_p(1), \\
\hat{\Delta}_\bullet(1) &= \Delta_\bullet(1) + o_p(1).
\end{aligned}
$$

Further, by theorem 1, $\hat{a}(x_0) - a(x_0) = o_p(1)$. Therefore, by Slutsky,

$$
\begin{aligned}
\hat{\mathcal{V}}_\bullet^0 &= \frac{n^{-1} \sum_{i=1}^n K_0(X_i^0, \hat{q}_i) \Lambda_i^2 \big(Y_i - \hat{a}(x_0)\big)^2}{\hat{\Delta}_\bullet^2(1)} \\
&= \frac{\Delta_\bullet(\Lambda^2 \vartheta) - 2\Delta_\bullet(\Lambda a) a(x_0) + \Delta_\bullet(\Lambda) a^2(x_0)}{\Delta_\bullet^2(1)} + o_p(1) = \frac{\Delta_\bullet(\Lambda^2 \sigma^2)}{\Delta_\bullet^2(1)} + o_p(1) = \mathcal{V}_\bullet^0 + o_p(1),
\end{aligned}
$$

because $\Delta_\bullet(\omega a) = a(x_0) \Delta_\bullet(\omega)$ for any function ω since $a(x) = a(x_0)$ for any x for which $g(x) = g(x_0)$.

I now show that $\hat{\mathcal{V}}_\bullet^j = \mathcal{V}_\bullet^j + o_p(h^{d_j - d_m})$. Let $\zeta_{ti}^j = K_0(X_i) \lambda^j(X_i^{-j}, x_0^j) R_f(X_i^j) \big(Y_i - a^j(X_i^{-j}, X_t^j)\big)$. From (133) it then follows that

$$
K_0(X_i) \lambda^j(X_i^{-j}, x_0^j) \big(Y_i R_f^j(X_i) - R_\nu^j(X_i)\big) = E_i \zeta_{1i}^j.
$$

Therefore, by (32),

$$
\begin{aligned}
& h^{d_m-d_j}\big|\hat{\mathcal{V}}_\bullet^j - \mathcal{V}_\bullet^j\big| \\
& \quad \le \kappa^{-d_j} h^{d_m} n^{-1} \sum_{i=1}^n K_0^2(X_i^j)\big(\lambda^j(X_i^{-j}, x_0^j)\big)^2 \\
& \qquad \times \left(n^{-1} \sum_{t=1}^n \Big(\big(\hat{R}_{ft}^j(X_i) - R_f^j(X_i)\big) Y_i - \big(\hat{R}_{\nu t}^j(X_i) - R_\nu^j(X_i)\big) \Big) \right)^2 && (146) \\
& \quad + \kappa^{-d_j} h^{d_m} n^{-1} \sum_{i=1}^n K_0^2(X_i^j)\big(\lambda^j(X_i^{-j}, x_0^j)\big)^2 \\
& \qquad \times \left| n^{-1} \sum_{t=1}^n \Big(\big(\hat{R}_{ft}^j(X_i) - R_f^j(X_i)\big) Y_i - \big(\hat{R}_{\nu t}^j(X_i) - R_\nu^j(X_i)\big) \Big) \right| \\
& \qquad \times \left| n^{-1} \sum_{t=1}^n \big(R_f^j(X_i) Y_i - R_\nu^j(X_i) \big) \right| && (147) \\
& \quad + \kappa^{-d_j} h^{d_m} \Big(n^{-1} \sum_{i=1}^n (E_i \zeta_{1i}^j)^2 - E(E_2 \zeta_{12}^j)^2 \Big) && (148) \\
& \quad + h^{d_m} \kappa^{-d_j} \Big(E(E_2 \zeta_{12}^j)^2 - h^{-d_j} \kappa^{d_j} \mathcal{V}_\bullet^j \Big) && (149)
\end{aligned}
$$

By lemmas 25 and 2, (146) and (147) are both $o_p(1)$. Squaring (148) and taking expectations implies that the following is a sufficient condition for (148) to be $o_p(1)$.

$$h^{2d_m} E(E_2 \zeta_{12}^j)^4 = o(n). \tag{150}$$

But by lemma 2, the LHS in (150) is $O(h^{2d_m - 3(d_0+D)-4}) = O(h^{-d_m-4}) = o(n)$ by assumption G. Finally, (149) is $o(1)$ by lemma 18.

The result for when λ is replaced with $\tilde{\lambda}/\hat{f}$ follows similarly. □

A.10 Identification

Lemma 26

$$\hat{g}^1(x^1) - g^1(x^1) = O_p\big(n^{-1/2} h^{-d_1}\big). \tag{151}$$

Proof: To simplify notation, let $\hat{\omega} = \hat{\underline{g}}_{(1)}, \omega = \underline{g}_{(1)}$, and let $\hat{\xi} = \hat{\underline{g}}^1(x^1), \xi = \underline{g}^1(x^1)$. Then (151) is equivalent to $\hat{\omega}^-(\hat{\xi}) - \omega^{-1}(\xi) = O_p\big(n^{-1/2} h^{-d_1}\big)$. Note that
$\hat{\xi} - \xi = \hat{\underline{g}}^1(x^1) - \underline{g}^1(x^1) = O_p\big(n^{-1/2} h^{-d_1}\big)$, which follows, with minor adjustments for the fact that

no uniform convergence is required here, from lemma 6. Then by the continuous differentiability of ω and by the assumption that ω' is bounded away from zero, the mean value theorem implies that

$$|\omega^{-1}(\hat{\xi}) - \omega^{-1}(\xi)| = \left|\frac{\hat{\xi} - \xi}{\omega'(\cdot)}\right| = O_p\big(n^{-1/2}h^{-d_1}\big). \tag{152}$$

It now remains to be shown that $\hat{\omega}^-(\hat{\xi}) - \omega^{-1}(\hat{\xi})$ converges at the same rate. Let $\aleph_0 = \max\big(\omega(0), \hat{\omega}(0)\big)$ and $\aleph_1 = \min\big(\omega(1), \hat{\omega}(1)\big)$. Then

$$|\hat{\omega}^-(\hat{\xi}) - \omega^{-1}(\hat{\xi})| = |\hat{\omega}^-(\hat{\xi}) - \omega^{-1}(\hat{\xi})| I\big(\hat{\xi} \in [\aleph_0, \aleph_1]\big) \tag{153}$$

$$+ |\hat{\omega}^-(\hat{\xi}) - \omega^{-1}(\hat{\xi})| I\big(\hat{\xi} \in [\omega(0), \aleph_0)\big) \tag{154}$$

$$+ |\hat{\omega}^-(\hat{\xi}) - \omega^{-1}(\hat{\xi})| I\big(\hat{\xi} \in (\aleph_1, \omega(1)]\big). \tag{155}$$

First (153). By construction of $\hat{\omega}^-$, some s^* exists for which $\hat{\omega}(s^*) = \xi$ and such that $\hat{\omega}^-\big(\hat{\omega}(s^*)\big) = s^*$. But then

$$\hat{\omega}^-(\hat{\xi}) - \omega^{-1}(\hat{\xi}) = s^* - \omega^{-1}\big(\hat{\omega}(s^*)\big) = \omega^-\big(\omega(s^*)\big) - \omega^{-1}\big(\hat{\omega}(s^*)\big).$$

Proceed as in (152). Since (154) and (155) are very similar, I only deal with (154). So $\omega(0) \leq \hat{\xi} < \hat{\omega}(0)$. Then

$$|\hat{\omega}^-(\hat{\xi}) - \omega^{-1}(\hat{\xi})| = |\omega^{-1}(\hat{\xi})| \leq \omega^{-1}\big(\hat{\omega}(0)\big) = \omega^{-1}\big(\hat{\omega}(0)\big) - \omega^{-1}\big(\omega(0)\big).$$

Proceed again as in (152). □

A.11 Iteration

Since there are only two groups, 0 and 1, in the derivation below, the superscript 2 will mean 'square'. The derivation uses many shortcuts, which are always repetitions of similar results derived in detail in the proof of theorem 1.

Let $\hat{g}(x^1) = \int \hat{a}(x)\tilde{\lambda}^2(x^0)dx^0$ such that $\hat{\chi}^1(x^1) = \hat{g}(x_0^1) - \hat{g}(x^1)$. Let moreover $\hat{g}_0 = \hat{g}(x_0^1)$ and $\hat{g}_i = \hat{g}(X_i^1)$ and similarly for $\hat{\hat{g}}, g$. It then follows from appendix A.7 that

$$\hat{a}_{S\chi}(x_0) - a(x_0) \approx \hat{a}_I(x_0) - a(x_0) + n^{-1}\sum_{i=1}^{n} K'(g_0 - g_i)\frac{Y_i - a_0}{f_{X^0 g}(x_0^0, g_0)}(\hat{g}_0 - g_0), \tag{156}$$

where $\hat{a}_I$ is again the Nadaraya–Watson estimator with regressors X_i^0, g_i and $\approx$ means that the remaining terms are irrelevant. The $f_{X^0 g}$–component in the denominator in (156) is the joint

density of X_i^0, g_i and replaces Δ since $\Lambda = 1$ everywhere. Similarly,

$$\hat{\hat{a}}_{S\chi}(x_0) - a(x_0) \approx \hat{a}_I(x_0) - a(x_0) + n^{-1}\sum_{i=1}^{n} K'(g_0 - g_i)\frac{Y_i - a_0}{f_{X^0 g}(x_0^0, g_0)}(\hat{\hat{g}}_0 - g_0).$$

$\hat{a}_{S\chi}$ and $\hat{\hat{a}}_{S\chi}$ are hence asymptotically equally efficient if $\hat{\hat{g}}_0 - g_0 \approx \pm(\hat{g}_0 - g_0)$. When $\hat{\hat{g}}_0 - g_0 \approx -(\hat{g}_0 - g_0)$, the result of section 4.2 applies.

Now, following steps similar to those in the proof of theorem 1,

$$\begin{aligned}
\hat{\hat{g}}_0 - g_0 &= \int \hat{a}_S(x^0, x_0^1)\tilde{\lambda}^2(x^0)dx^0 - g_0 \\
&\approx n^{-1}\sum_{i=1}^{n}\int K(x^0 - X_i^0, \hat{g}_0 - \hat{g}_i)\frac{Y_i - a(x^0, x_0^1)}{f_{X^0 g}(x^0, g_0)}\tilde{\lambda}^2(x^0)dx^0 \\
&\approx n^{-1}\sum_{i=1}^{n} K(\hat{g}_0 - \hat{g}_i)\frac{Y_i - a(X_i^0, x_0^1)}{f_{X^0 g}(X_i^0, g_0)}\tilde{\lambda}_i^2 \\
&\approx n^{-1}\sum_{i=1}^{n} K'(g_0 - g_i)\frac{Y_i - a(X_i^0, x_0^1)}{f_{X^0 g}(X_i^0, g_0)}\tilde{\lambda}_i^2(\hat{g}_0 - g_0), \qquad (157)
\end{aligned}$$

where the last step involves the mean value theorem. The zero–order term is omitted because it is asymptotically negligble, and the $\hat{g}_i - g_i$ bit in the first order term disappears because it is of a smaller order of magnitude when averaged across i.

By the weak law of large numbers (157) is

$$\approx E\left(K'(g_0 - g_1)\frac{Y_1 - m(X_1^0, g_0)}{f_{X^0 g}(X_1^0, g_0)}\tilde{\lambda}_1^2\right)(\hat{g}_0 - g_0).$$

But by standard kernel derivative estimation procedures,

$$\begin{aligned}
E\left(K'(g_0 - g_1)\frac{Y_1 - m(X_1^0, g_0)}{f_{X^0 g}(X_1^0, g_0)}\tilde{\lambda}_1^2\right) &\approx \int K'(g_0 - g)\big(m(x^0, g) - m(x^0, g^0)\big)\frac{f(x^0, g)}{f(x^0, g_0)}\tilde{\lambda}^2(x^0)dx^0 dg \\
&\approx -\int \frac{\partial m}{\partial g}(x^0, g_0)\tilde{\lambda}^2(x^0)dx^0 \\
&= -\frac{\int \frac{\partial a}{\partial x^{11}}(x^0, x_0^1)\tilde{\lambda}^2(x^0)dx^0}{\frac{\partial g}{\partial x^{11}}(x_0^1)} = -1.
\end{aligned}$$

Linear, $\sigma = 1$

		MSE		MSE99		MDAE	
d	n	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
3	100	297	339	214	290	282	323
	200	201	256	141	212	221	264
4	100	6648	547	304	489	350	450
	200	430	449	202	395	271	385
9	100	717	987	654	914	553	668
	200	581	984	518	910	480	668

Probit–Like, $\sigma = 1$

MSE		MSE99		MDAE	
$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
⧫	303	098	257	224	394
6909	227	078	185	210	238
⧫	513	110	456	290	424
9901	417	101	364	293	358
⧫	986	145	913	324	667
⧫	983	134	909	333	667

Product of Logs, $\sigma = 1$

MSE		MSE99		MDAE	
$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
2064	313	255	266	234	298
⧫	239	260	196	217	242
658	511	013	454	078	421
397	415	006	362	052	356
6370	986	022	913	090	666
⧫	983	010	909	061	667

Linear, $\sigma = 2$

		MSE		MSE99		MDAE	
d	n	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
3	100	1219	1242	622	1052	460	598
	200	678	931	415	761	363	484
4	100	⧫	2079	789	1851	533	857
	200	1773	1693	545	1481	422	727
9	100	2075	3946	1852	3653	896	1335
	200	1688	3932	1478	3638	769	1335

Probit–Like, $\sigma = 2$

MSE		MSE99		MDAE	
$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
⧫	1205	281	1019	314	583
⧫	902	177	734	260	470
⧫	2045	272	1818	303	843
⧫	1661	175	1450	272	713
⧫	3945	644	3652	416	1336
⧫	3931	459	3637	346	1335

Product of Logs, $\sigma = 2$

MSE		MSE99		MDAE	
$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
⧫	1216	424	1029	336	586
⧫	915	349	745	286	473
⧫	2043	160	1816	196	842
⧫	1659	058	1448	128	712
⧫	3945	552	3652	325	1333
⧫	3931	336	3637	216	1335

Probit, $\sigma = 1$

		MSE		MSE99		MDAE	
d	n	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
3	100	570	029	064	023	204	071
	200	563	020	058	002	195	006
4	100	1060	047	024	040	100	071
	200	2457	037	016	030	080	064
9	100	1549	071	038	063	114	022
	200	094	071	031	062	105	022

Flat, $\sigma = 1$

MSE		MSE99		MDAE	
$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
7220	301	037	254	117	291
775	225	020	183	085	234
658	511	013	454	078	421
397	415	006	362	052	356
6370	986	022	913	090	667
⧫	983	010	909	061	667

Arctan–Power, $\sigma = 1$

MSE		MSE99		MDAE	
$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
2056	348	244	295	205	318
⧫	273	254	222	185	260
3472	512	066	455	163	422
2086	416	060	363	153	357
7966	986	092	913	154	667
⧫	983	082	909	136	667

Probit, $\sigma = 2$

		MSE		MSE99		MDAE	
d	n	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
3	100	040	047	035	041	119	120
	200	028	034	023	028	097	097
4	100	⧫	078	027	070	112	158
	200	⧫	061	020	054	096	132
9	100	⧫	127	039	119	126	172
	200	277	126	032	118	118	173

Flat, $\sigma = 2$

MSE		MSE99		MDAE	
$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
⧫	1202	212	1017	254	583
⧫	900	109	732	182	468
⧫	2043	160	1816	196	842
⧫	1659	058	1448	128	712
⧫	3945	552	3652	325	1333
⧫	3931	336	3637	216	1335

Arctan–Power, $\sigma = 2$

MSE		MSE99		MDAE	
$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$	$\hat{a}_{S\pi}$	$\hat{a}$
⧫	1249	436	1058	307	597
⧫	948	367	772	245	483
⧫	2044	215	1817	242	843
⧫	1660	118	1449	187	713
⧫	3945	604	3652	364	1334
⧫	3931	404	3637	262	1334

Table 1: Simulation results. Entries were multiplied by 1,000. Models as described in the text. ⧫=entry at least 10,000. MSE=mean square error, MSE99=mean square error of 99% of replications, MDAE=median absolute error. Please refer to text for a precise description.

Figure 1a: Estimator distribution at (0,0,0), linear model

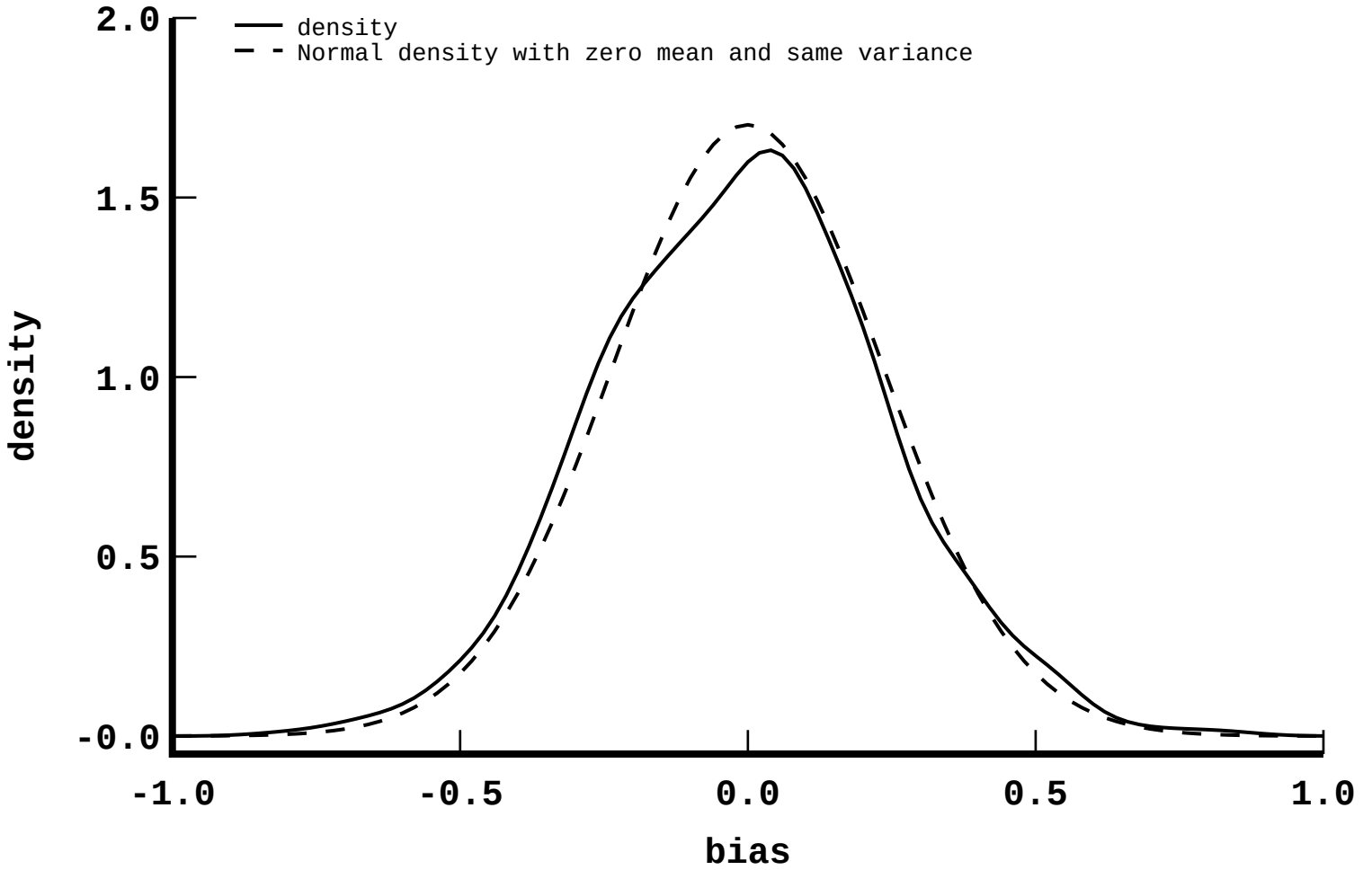

Figure 1b: Estimator distribution at (0,1,0), linear model

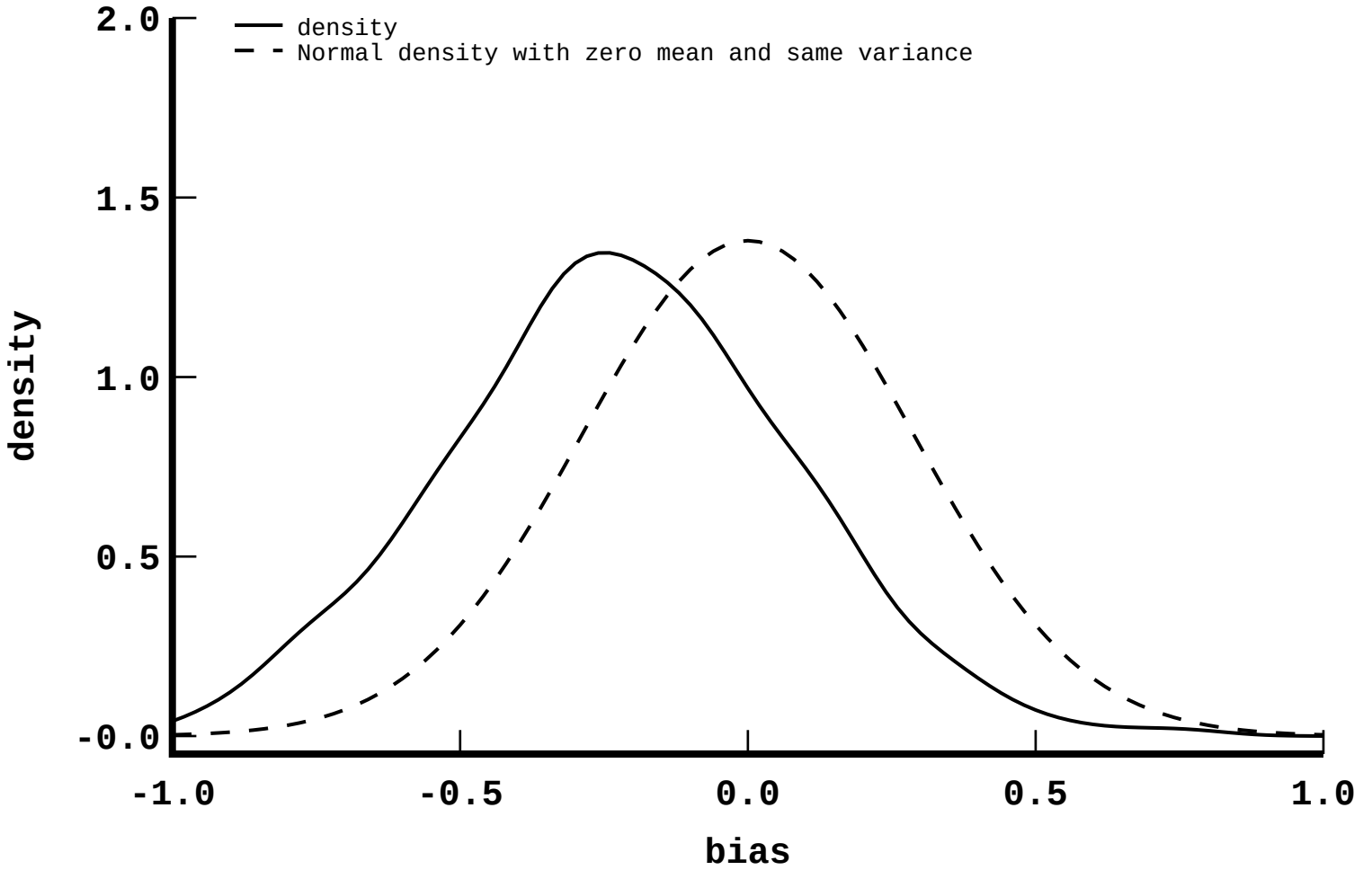

Figure 1c: Estimator distribution at (-1,1,-1), linear model

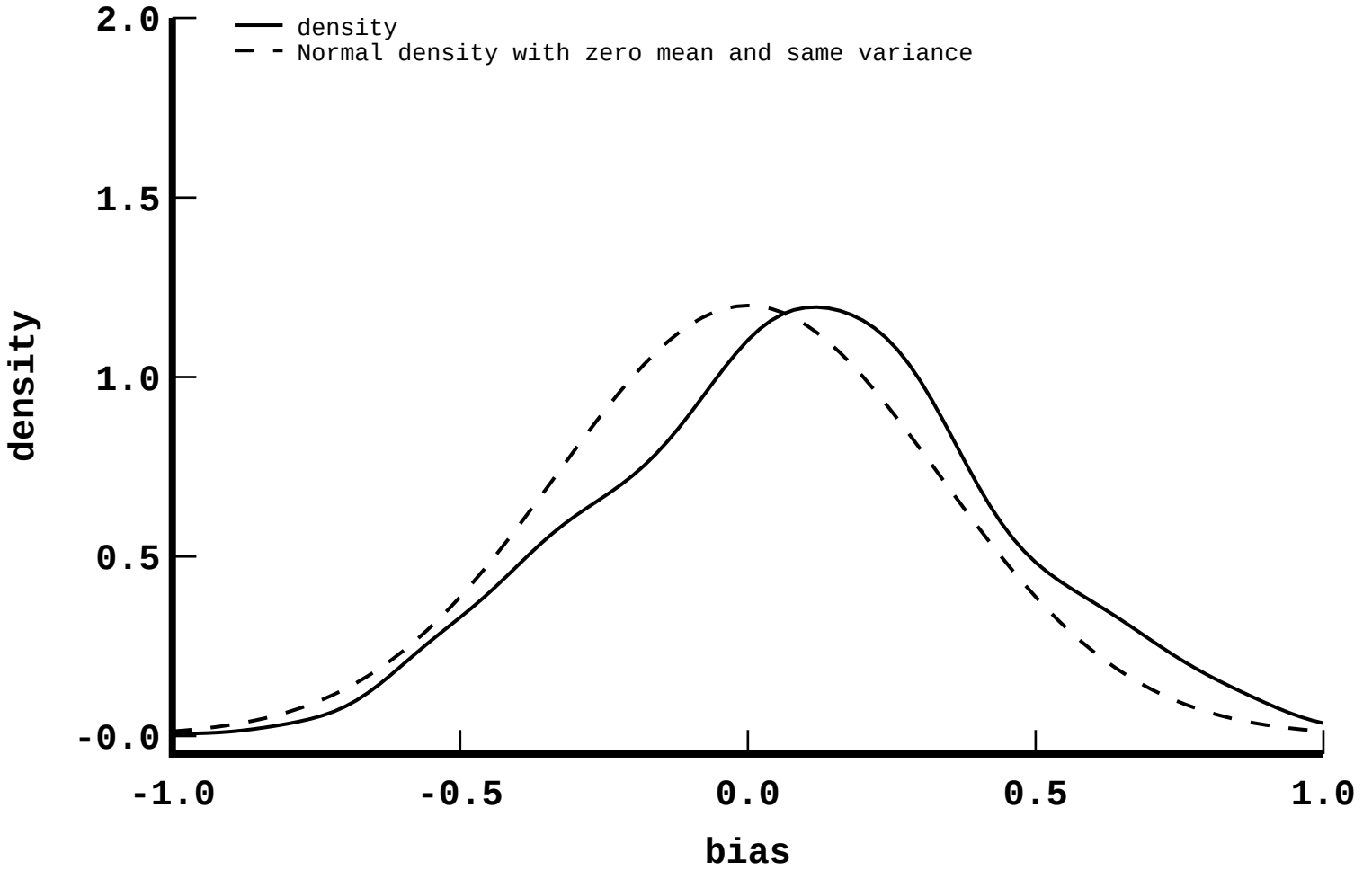

Figure 1d: Estimator distribution at (1,1,1), linear model