Feasible Multivariate Nonparametric Regression Estimation Using Weak Separability

Joris Pinkse*
The University of British Columbia

This Version: August 1999

Abstract

One of the main practical problems of nonparametric regression estimation is the *curse of dimensionality*. The curse of dimensionality arises because nonparametric regression estimates are dependent variable averages local to the point at which the regression function is to be estimated. The number of observations 'local' to the point of estimation decreases exponentially with the number of dimensions. The consequence is that the variance of unconstrained nonparametric regression estimators of multivariate regression functions is often so great that the unconstrained nonparametric regression estimates are of no practical use.

In this paper I propose a new estimation method of *weakly separable* multivariate nonparametric regression functions. Weak separability is a weaker condition than required by other dimension–reduction techniques, although similar asymptotic variance reductions obtain. Indeed, weak separability is weaker than generalized additivity (see Härdle and Linton, 1996 and Horowitz, 1998). The proposed estimator is relatively easy to compute. Theoretical results in this paper include (i) a uniform law of large numbers for *marginal integration* estimators, (ii) a uniform law of large numbers for *marginal summation* estimators, (iii) a uniform law of large numbers for my new nonparametric regression estimator for weakly separable regression functions, (iv) both a uniform strong and weak law of large numbers for *U–statistics*, and (v) three central limit theorems for my nonparametric regression estimator for weakly separable regression functions.

*This paper is based on research supported by a UBC Humanities and Social Sciences grant. I thank Don Andrews, Chuck Blackorby, Richard Blundell, Craig Brett, Erwin Diewert, David Green, Nancy Heckman, Joel Horowitz, Oliver Linton, Peter Robinson, Margaret Slade, Thanasis Stengos and seminar participants at the University of British Columbia (statistics and economics), the London School of Economics and Political Science, University College London, the University of Bristol, Yale University and the University of Groningen for useful suggestions.

1 Introduction

One of the main practical problems of nonparametric regression estimation is the *curse of dimensionality*. The curse of dimensionality arises because nonparametric regression estimators are dependent variable averages local to the point at which the regression function is to be estimated. The number of observations 'local' to the point of estimation decreases exponentially with the number of dimensions. The consequence is that the variance of unconstrained nonparametric regression estimators of multivariate regression functions is often so great that the unconstrained nonparametric regression estimates are of no practical use.

In this paper I propose a new estimation method of *weakly separable* multivariate nonparametric regression functions. Weak separability is a weaker condition than required by other dimension–reduction techniques, although similar asymptotic variance reductions obtain. Indeed, weak separability is weaker than generalized additivity (see Härdle and Linton, 1996 and Horowitz, 1998). In section 2 I give an example related to returns of education which highlights the differences between additive and weak separability. The proposed estimator is relatively easy to compute. Theoretical results in this paper include (i) a uniform law of large numbers for *marginal integration* estimators, (ii) a uniform law of large numbers for *marginal summation* estimators, (iii) a uniform law of large numbers for my new nonparametric regression estimator for weakly separable regression functions, (iv) both a uniform strong and weak law of large numbers for *U–statistics*, and (v) three central limit theorems for my nonparametric regression estimator for weakly separable regression functions..

Some of these results are applicable outside the direct context of this paper. The marginal summation estimator is an alternative to the marginal integration estimator; it is easier to compute and requires less computer time particularly when the number of dimensions is large. The marginal summation estimator can equally be used to facilitate the computation of estimators under generalized additivity (Härdle and Linton, 1996, or Horowitz, 1998) or indeed additivity (Linton and Nielsen, 1995). The uniform strong and weak laws of large numbers of U–statistics have applicability far beyond the context of this paper. Indeed, many commonly encountered statistics are U–statistics including average derivative estimators (Powell et al., 1989) and various nonparametric test statistics.

Many authors have addressed the curse of dimensionality by imposing a structure on the regression function which allows for more efficient estimation. Robinson's (1988) partial linear model

additively separates the regression function into a linear parametric part and a low–dimensional nonparametric part. The regression coefficients can be estimated $\sqrt{n}$–consistently and the nonparametric regression function is estimated at a faster rate of convergence than if the entire regression function were estimated by multivariate nonparametric regression. Others have assumed the regression function to be *additively separable*, i.e. to be the summation over nonparametric regression functions (usually univariate), where the regressors that appear in any one of these functions does not appear in any of the others.

In the context of *series estimation*, imposing additive separability is straightforward since both additive parts can be expanded separately. The general results of Andrews (1991) can then be applied to ensure asymptotic normality of the series estimator.

For kernel estimators, the *backfitting* method of Friedman and Stützle (1981), see also Hastie and Tibshirani (1990), is one example. Alternatively, one can use a two–step procedure in which the first step consists of computing the usual multivariate nonparametric regression estimator where in the second step the estimator of each term is determined from the multivariate estimator by integrating over all regressors which do not enter as arguments. The gain is that the (large sample) variance of this estimator is smaller than that of the multivariate nonparametric regression estimator. This idea was put forward by Linton and Nielsen (1995) and it and variants have been studied in depth. Nielsen and Linton (1997) have studied the relationship between the backfitting algorithm and the marginal integration estimator under additive separability. They found that the asymptotic properties of the backfitting method are generally better than those obtainable by marginal integration.

One variant is the *generalized additive model* in which the unknown regression function is a (*link*) function of a summation over univariate regression functions. For known link functions, Linton and Härdle (1996) have shown that similar results obtain and Horowitz (1998) has obtained similar results for when the link function is unknown.

Rilstone (1996) proposed an estimator for a nonparametric regression function where one of its arguments is itself an estimable conditional mean. In the first step the conditional mean is estimated, and the estimates are used as regressors in the second step estimator of the nonparametric regression function of interest. This procedure is again asymptotically more efficient than full multivariate nonparametric regression estimation because the nonparametric regression function is separated into two functions with fewer arguments. Another, less related but no less interesting, example of

generated regressors in a nonparametric context is Ahn (1997).

Section 2 introduces the concept of weak separability. Section 3 discusses identification conditions. The estimation method is outlined in section 4. Section 5 contains the main results of the paper, in section 6 I discuss the choice of input parameters, section 7 discusses tests for weak separability of a nonparametric regression function and in section 8 I discuss computational issues. Section 9 outlines some avenues yet to be explored and section 10 contains some modest Monte Carlo simulations. Section 11 concludes. All proofs are in the appendix.

2 Weak Separability

Let $\{(X_i, Z_i, Y_i)\}$ be an independent and identically distributed (i.i.d.) sequence of random vectors for which $a(x,z) = E(Y_1|X_1 = x, Z_1 = z)$. I assume that X_i and Z_i have continuous distributions, though this assumption could potentially be relaxed (see for instance Delgado and Mora, 1995). Let d_ϖ denote the dimension of any variable ϖ, and suppose that a is weakly separable.

Definition 1 *The function a is weakly separable (x,z) if two functions m, g exist such that for all values of (x,z),*

$$a(x,z) = m\left\{x, g(z)\right\}, \tag{1}$$

where $d_g = 1, d_z \geq 2$ and m is monotonic in g.

The above definition of weak separability is the simplest form. More general forms are discussed in section 9.

Weak separability is an assumption that has been frequently used in the context of demand systems and production functions. In demand theory, weak separability is imposed on the utility function. If a in (1) were a utility function, then demand for z_i only depends on total expenditure on goods in the z–vector and prices of goods in the z–vector. Similarly, if a production function is assumed weakly separable then the input demand function for z_i depends only on total expenditure on inputs in the z–vector and prices of goods in the z–vector. In both instances, the gain is a considerable reduction in the dimensionality of the demand or input demand functions, the objects of estimation.[1]

[1]Note that it is often possible to estimate a production function directly, if one is willing to assume away any endogeneity concerns relating to the choice of inputs.

Here, the focus is on the estimation of the weakly separable function a itself instead of on functions derived thereof. Nevertheless, the consequences of the weak separability assumption are similar, albeit that in the general case no direct conclusions relating to economic theory can be drawn. The most important limitation of weak separability is that the ratio $\frac{\partial a}{\partial z_i}/\frac{\partial a}{\partial z_j} = \frac{\partial g}{\partial z_i}/\frac{\partial g}{\partial z_j}$ cannot depend on x for any i, j.

Weak separability nests the generalized additive model with unknown link function of Horowitz (1998), and is hence also more general than any of the specifications encompassed by the Horowitz estimator, including those mentioned in the introduction. The generalized additive specification has

$$a(x,z) = m_L\left\{\sum_{i=1}^{d_x} g_{x_i}(x_i) + \sum_{i=1}^{d_z} g_{z_i}(z_i)\right\},$$

where m_L and the g_{x_i}'s and g_{z_i}'s are unknown functions with scalar argument. Generalized additivity imposes that for any (i,j), $\frac{\partial a}{\partial x_i}/\frac{\partial a}{\partial x_j}, \frac{\partial a}{\partial x_i}/\frac{\partial a}{\partial z_j}, \frac{\partial a}{\partial z_i}/\frac{\partial a}{\partial z_j}$ only depend on $(x_i, x_j), (x_i, z_j)$ and (z_i, z_j) respectively. Generalized additivity is hence a stronger assumption than weak separability, which can be seen if one chooses $g(z) = \sum_{i=1}^{d_z} g_{z_i}(z_i)$.[2]

One example which illustrates the difference between the (generalized) additivity and weak separability assumptions relates to returns to education.[3] In the most narrow model (Mincer, 1974, chapter 2) the difference in expected log earnings between two individuals with the same level of experience depends only on the differences in schooling and other characteristics, not on the experience level itself. Hence, the model can be described by an additively separable specification with experience level in one term and all other characteristics in the other.[4] If differences in expected log earnings between different groups do depend on the experience level, then a weakly separable specification is more appropriate. Expected log earnings of two groups (high and low schooling) diverge as a function of the experience level if the amount of time spent on "post–schooling" is positively

Some useful further references in these areas are Blackorby, Davidson and Schworm (1991), Blackorby and Schworm (1988), Blackorby, Schworm and Fisher (1986), Blundell (1988), Diewert and Wales (1987,1988,1992) and Woodland (1978). An application of weak separability in the context of monetary aggregation is Barnett (1980).

[2]Note that Horowitz (1998) also allows for $a(x,z) = m_L\left\{\sum_{i=1}^{d_x} g_{x_i}(x_i) + \sum_{i=1}^{d_z} g_{z_i}(z_i), g_w(w_i)\right\}$ for scalar w_i. Horowitz concentrates on estimation of the g_{z_i}'s and g_{x_i}'s and does not use the weakly separable structure further. Further use would indeed require an extension of the model since the g_{z_i}'s, g_{x_i}'s and g_w are scalar–valued.

[3]I thank David Green for this example.

[4]Note however that most results for additively separable functions require each nonparametric function to have scalar argument.

correlated with the amount of time spent on schooling (Mincer, 1974, p.31). They converge if the correlation is negative. An additive specification thus assumes the correlation to be zero.

The definition of weak separability (Definition 1) can be generalized in many ways. Weak separability can be nested or $m\{x, g(z)\}$ can be replaced with $m\{g_x(x), g_z(z), \ldots\}$. Some generalizations are discussed in section 9.1. The main results of this paper apply to Definition 1.

3 Identification

For any separable function a there are many functions m and g which satisfy (1). For instance, if $\varkappa$ is any monotonic function, then $a(x,z) = m\{x, g(z)\} = m^*\{x, g^*(z)\}$ with $g^* = \varkappa^{-1}(g)$ and $m^*(x,g) = m\{x, \varkappa(g)\}$, and hence (m, g) and (m^*, g^*) can not be separately identified. It is assumed here that m and g are not of separate interest, and hence one can impose any identification condition on m and g.[5] An identification condition guarantees that any weakly separable function a can be reproduced by one and only one combination of m and g that satisfies the identification condition.

One commonly used identification condition is $g(z) = m\{0, g(z)\}$. This identification condition does not allow for increased efficiency since it does not involve any averaging. Instead, my identification condition allows for g to be estimated by marginal integration, thereby ensuring that a is estimated more efficiently than if it were estimated by an unconstrained multivariate nonparametric regression estimator $\hat{a}$.

Let λ be some practitioner–chosen nonnegative function for which $0 < \int_{\mathcal{X}} \lambda(x)dx < \infty$,where $\mathcal{X}$ is the support of the density f_X of X_1.

Theorem 1 *Once the practitioner has chosen λ, setting $g(z) = \int_{\mathcal{X}} a(x,z)\lambda(x)dx$ uniquely identifies (g, m).*

All proofs are in an appendix. A discussion on the choice of λ follows in section 6.

4 Estimation Method

My estimation method consists of three steps. In the first step, an unconstrained nonparametric regression estimator $\hat{a}$ of a is computed. The second step consists of finding an estimator $\hat{g}$ of g,

[5] If m and g are of separate interest, then the application should provide appropriate identification conditions.

which converges at a faster rate than $\hat{a}$ converges to a. In the third step $\hat{g}$ is used to regress Y_i on $\{X_i, \hat{g}(Z_i)\}$ nonparametrically, giving an estimate of m. Since $d_x + 1$, the dimension of $\{X_i, \hat{g}(Z_i)\}$ is less than $d_x + d_z$, the resulting estimator of m also generally converges at a faster rate than does $\hat{a}$.

All nonparametric regression estimators used are nonparametric (Nadaraya–Watson, see Nadaraya, 1964, and Watson, 1964) kernel regression estimators. The unconstrained Nadaraya–Watson estimator of a is

$$\hat{a}(x,z) = \frac{\frac{1}{nh_g^{d_b}} \sum_{i=1}^{n} k_{h_g}(x - X_i) k_{h_g}(z - Z_i) Y_i}{\frac{1}{nh_g^{d_b}} \sum_{i=1}^{n} k_{h_g}(x - X_i) k_{h_g}(z - Z_i)}, \tag{2}$$

with $d_b = d_x + d_z$, h_g the practitioner–chosen *bandwidth*, $k_{h_g}(u) = k(u/h_g)$, with k the *kernel*. I use the symbol k as a generic symbol for kernel, its exact form being determined by the dimension of its argument. So the functions k used on $(x - X_i)$ and $(z - Z_i)$ in (2) are different unless the dimensions of X_i and Z_i are the same.

It is the choice of functional form rather than the choice of estimation method which allows the dimension reduction result. The Nadaraya–Watson kernel regression estimator is but one choice. In section 9 I discuss potential alternatives. The trade–offs between *local* nonparametric methods like kernel regression estimation and local polynomial estimation and *global* nonparametric methods like series estimation and artificial neural networks are well–known. My reason for opting for kernel regression estimation instead of local polynomial estimation is simplicity of proofs and arguments. I expect that similar results can be obtained for local polynomial estimators.

The natural estimator of g is the *marginal integration* estimator

$$\hat{g}(z) = \int_{\mathcal{X}} \hat{a}(x,z) \lambda(x) dx \tag{3}$$

for a judiciously chosen function λ, and some unconstrained multivariate nonparametric regression estimator $\hat{g}$. Note that $\hat{g}$ has exactly the same form as the Linton and Nielsen (1995) estimator. There are two differences: $\hat{g}$ is only the first stage of my estimation procedure and g is a function of one scalar variable in Linton and Nielsen (1995), and of at least two variables in this paper.

There are many ways λ can be chosen and the choice can affect the asymptotic variance matrix (see section 6). One particular choice is $\lambda(x) = f_X(x) I(x \in \mathcal{B}_X)$, with I the indicator function, f_ϖ the density of ϖ_1 for any random variable ϖ_1, and $\mathcal{B}_X$ some compact subset of $\mathcal{X}$, on which I will

impose conditions in Assumption A. Since f_X is not observed, one could instead use the *marginal summation* estimator

$$\hat{g}_\Sigma(z) = n^{-1} \sum_{i=1}^{n} \hat{a}_{-i}(X_i, z) I(X_i \in \mathcal{B}_X),$$

where the subscript $-i$ denotes that observation i is not used in the determination of $\hat{a}_{-i}$ (*leave one out*). $\hat{g}_\Sigma$ can be preferable to $\hat{g}$ for reasons of computational ease (see section 8). Although the results for the marginal summation estimator in my paper are for $\lambda = f_X$, other choices can be incorporated; see section 9.

Once g is estimated using (3), its estimates can be used to obtain a more efficient estimator of a, namely

$$\hat{a}_S(x, z) = \frac{\frac{1}{nh_m^{d_x+1}} \sum_{i=1}^{n} k_{h_m}(x - X_i) k_{h_m} \{\hat{g}(z) - \hat{g}_{-i}(Z_i)\} Y_i}{\frac{1}{nh_m^{d_x+1}} \sum_{i=1}^{n} k_{h_m}(x - X_i) k_{h_m} \{\hat{g}(z) - \hat{g}_{-i}(Z_i)\}},$$

where h_m is again a bandwidth and $\hat{g}_{-i}$ is $\hat{g}$ when observation i is omitted in its estimation.

5 Main Results

The main results are divided into two separate subsections. In section 5.1, I study the properties of the marginal integration estimator $\hat{g}$ and the marginal summation estimator $\hat{g}_\Sigma$. In particular, I prove that $\hat{g}$ and $\hat{g}_\Sigma$ converge uniformly to g (Theorems 2 and 4). In doing so, I also prove (Theorem 3) both a uniform strong law of large numbers and a uniform weak law of large numbers for *U–statistics*.

In section 5.2, I establish the limiting distribution of $\hat{a}_S$. Depending on the dimensions of X_1 and Z_1, the results are in Theorems 5, 6 and 7. In Theorem 8 I show that $\hat{a}_S$ converges uniformly to a.

5.1 Estimating g

I first state the assumptions and then discuss them all at once.

Assumption A *There is a Cartesian product of intervals* $\mathcal{B} = \mathcal{B}_X \times \mathcal{B}_Z \subset \mathbb{R}^{d_b}$ *for which* $\inf_{b \in \mathcal{B}} f_{XZ}(b) > 0$ *and for which* $\int_{\mathcal{X}} \lambda(x)dx = \int_{\mathcal{B}_X} \lambda(x)dx > 0$.

Assumption B *For some* $p > 2, \sup_{x \in B_X, z \in B_Z} E\left(|Y_1|^p \,|X_1 = x, Z_1 = z\right) < \infty$ *and* $\sigma^2_{XZ}(x,z) = E\left(Y_1^2 | X_1 = x, Z_1 = z\right)$ *is continuous at all* $(x,z) \in \mathcal{B}$.

Assumption C *Both* $f_{XZ}(x,z)$ *and* $a(x,z)$ *are at least* $r \geq 2$ *times boundedly differentiable on the interior of* $\mathcal{B}$.

Assumption D *The kernels used in the nonparametric kernel regression estimator are products of even univariate* r*–th order kernels* $k_{(1)}$ *with bounded support, i.e. they are bounded functions* k *such that* $k(u) = \prod_{i=1}^{d_u} k_{(1)}(u_i)$, $\int k_{(1)}(t)dt = 1$, $\int k_{(1)}(t)t^\ell dt = 0$, $\ell = 1, \ldots, r-1$, $\int \left|k_{(1)}(t)t^r\right| dt < \infty$.

Assumption E *For some* $\epsilon > 1 - 2/p$, $n^{1-\epsilon} h_g^{d_b} \to \infty$ *and* $h_g \to 0$, *as* $n \to \infty$.

Since $\hat{a}$ contains a denominator term, stronger results obtain when $\hat{a}$ is integrated only over a bounded set. Assumption A says that λ should be chosen positive only over a set on which the joint density of (X_i, Z_i) is known to be bounded away from zero. *I do not assume anywhere that any density has bounded support.*

Assumption B contains a mild moment condition and a continuity condition on the conditional variance. Assumption C is a smoothness condition, where smoothness is measured in terms of the number of existing (bounded) derivatives.

Assumptions D and E do not impose conditions on the data. Instead, they restrict the set of kernels and bandwidths the practitioner can use. Kernels of order $r = 2$ are standard. Higher order kernels are a theoretical tool and are useful to increase the rate at which the bias disappears with an increase in the sample size. For small and moderate samples the increase in the variance is such that second order kernels often work better in practice. Higher order kernels take negative values. An example of a fourth order kernel is $k_{(1)}(u) = \left\{(3 - u^2)/(4\pi)\right\} \exp(-u^2/2)$, but kernels of any order, including infinite order, can be constructed.

Finally, Assumption E contains a weak restriction on the way the bandwidth choice should change with the sample size in the limit. Like Assumption F further on, it is merely a technical construct and provides no guidance on how to choose bandwidths for a sample of finite size. See section 6 for a discussion of the choice of input parameters.

The first result is a uniform convergence result for $\hat{g}$.

Theorem 2 *Under Assumptions A, B, C, D and E, then for $\mathcal{Z}$ a strict subset of $\mathcal{B}_Z$,*

$$\sup_{z\in\mathcal{Z}}|\hat{g}(z)-g(z)| = o_p(n^{-1/2}h_g^{-d_z/2}\log n + n^{-1}h_g^{-d_b}\log n) + O_p(h_g^r). \tag{4}$$

In particular, when $h_g \sim n^{-1/\{r+d_z+\max(r,d_x)\}}$, as $n\to\infty$, then

$$\sup_{z\in\mathcal{Z}}|\hat{g}(z)-g(z)| = o_p\left\{n^{-r/\{r+d_z+\max(r,d_x)\}}\log n\right\}.$$

Theorem 2 establishes that when r can be chosen greater than d_x, i.e. when a is more than d_x times differentiable, then the uniform convergence rate of $\hat{g}$ is the same as the best attainable for d_z–variate nonparametric kernel regression estimators using r–th order kernels.

For the marginal summation estimator $\hat{g}_\Sigma$ the proof is a little more complicated. Instead of proving directly that $\hat{g}_\Sigma$ converges to g at a particular rate I establish that $\hat{g}_\Sigma - \hat{g}$ converges no slower than $\hat{g}-g$, where $\hat{g}, g$ are defined in terms of $\lambda = f_X$. Since $\hat{g}_\Sigma$ involves a double–sum, *U–statistic* theory applies (Hoeffding, 1948). To my knowledge there are no uniform laws of large numbers for U–statistics which apply in the current scenario. Since a uniform law of large numbers for U–statistics is of interest in its own right I present it in the text.

Let $\{\xi_i\}$ be an independent and identically distributed sequence of random variables and consider the U-statistic $\tilde{S}_n(t) = \sum_{s=1}^n\sum_{i=1}^{s-1}\tilde{U}_{nsi}(t)$ with $\tilde{U}_{nsi} = \tilde{U}_n(t,\xi_s,\xi_i)$ with $\tilde{U}_n$ a function symmetric in its last two arguments.

Theorem 3 *If t indexes a function class $\mathcal{F} = \left\{\tilde{U}_n(t,\cdot,\cdot), t\in\mathcal{T}\right\}$ with polynomial discrimination (see Pollard, 1984, Definition II.13) and if in addition for some $p_U > 0$,* $\limsup_{n\to\infty}\sup_{t\in\mathcal{T}} E\left|E\left\{\tilde{U}_{n12}(t)|\xi_1\right\}\right|^{p_U} < \infty$, *and if for all n, $\sup_{t\in\mathcal{T}} V\left\{\tilde{U}_{n12}(t)\right\} \le \sigma_{Un}^2$ and $\sup_{t\in\mathcal{T}} V\left[E\left\{\tilde{U}_{n12}(t)|\xi_1\right\}\right] \le \sigma_{UCn}^2$ such that for some $\iota > 0$, $n^{1/2-1/p_U}\sigma_{UCn}(\log n)^{\iota/2}\to\infty$ as $n\to\infty$, then*

$$n^{-2}\sup_{t\in\mathcal{T}}\left|\tilde{S}_n(t) - E\left\{\tilde{S}_n(t)\right\}\right| = o_p\left(n^{-1/2}\sigma_{UCn}\log n + n^{-1}\sigma_{Un}\log n\right).$$

If in addition $n^{1/2-2/p_U}\sigma_{UCn}(\log n)^{\iota/2}\to\infty$ as $n\to\infty$ then

$$n^{-2}\sup_{t\in\mathcal{T}}\left|\tilde{S}_n(t) - E\left\{\tilde{S}_n(t)\right\}\right| = o_{a.s.}\left(n^{-1/2}\sigma_{UCn}\log n + n^{-1}\sigma_{Un}\log n\right)$$

The only difference between the uniform strong law and uniform weak law in Theorem 3 is that the uniform strong law imposes a stronger moment condition. If σ_{UCn}^2 and σ_{Un}^2 do not depend on

the sample size, then $p_U > 2$ suffices for the uniform weak law and $p_U > 4$ for the uniform strong law. With Theorem 3, Theorem 4 is relatively easy to prove.

Theorem 4 *Under Assumptions A, B, C, D and E, then*

$$\sup_{z \in \mathcal{B}_Z} |\hat{g}_\Sigma(z) - \tilde{g}_{f_X}(z)| = o_p(n^{-1} h_g^{-d_b/2} \log n) + O_p(h_g^r), \tag{5}$$

where $\tilde{g}_{f_X}$ is the $\tilde{g}$ estimator with choice $\lambda(x) = f_X(x) I(x \in \mathcal{B}_X)$.

Note that the $O_p(h_g^r)$ term in (5) also occurs in (4). Note also that since $h_g \to 0$ as $n \to \infty$ by Assumption E, $n^{-1} h_g^{-d_b/2} \log n$ is of smaller order than $n^{-1} h_g^{-d_b} \log n$. Hence the marginal summation and marginal integration estimator (with choice $\lambda = f_X$) are asymptotically equivalent.

5.2 Properties of $\hat{a}_S$

Two further assumptions are required to obtain results for $\hat{a}_S$. The first assumption, Assumption F, is like Assumption E a technical restriction on the rate at which the bandwidths should decrease with an increase in the sample size. Hence. there is little to be learnt from Assumption F in terms of the optimal choice of bandwidth in a sample of finite size. The discussion below is hence limited to demonstrating that, given sufficient smoothness, bandwidth sequences satisfying Assumption F indeed exist.

Let $\Phi > 1$ denote the number of bounded derivatives of k and set $\Gamma = I_z / (I_z + I_x)$, where $I_z = I(d_z \geq d_x + 1)$ and $I_x = I(d_z \leq d_x + 1)$ such that $\Gamma = 1, 1/2, 0$ according to whether $d_z > d_x + 1$, $d_z = d_x + 1$, and $d_z < d_x + 1$, respectively. Let Π_1, Π_2 denote some finite positive constants.

Assumption F *The bandwidth sequences satisfy*

$$\begin{cases} \left(h_m h_g^{-1}\right)^{2\Gamma-1} \to \Pi_1 \Gamma (1-\Gamma), & n^{-1} h_g^{-d_z} h_m^{2(\Phi+1)/(1-\Phi)} (\log n)^2 \to 0, \\ \left(h_g^{d_z} h_m^{-d_x-1}\right)^{2\Gamma-1} \to \Pi_2 \Gamma (1-\Gamma), & n^{-1} h_g^{-d_z-d_x} h_m^{(\Phi+1)/(1-\Phi)} \log n \to 0, \\ h_g^r h_m^{\{2+(1-\Gamma)r\}/(\Gamma-2)} \to 0, & n^{-1} h_g^{d_z(\Gamma-2)} h_m^{\{(d_x+1)(1-\Gamma)-4\}} (\log n)^4 \to 0, \\ h_g^r h_m^{(\Phi+1)/(1-\Phi)} \to 0, & n^{-1} h_g^{\{d_z(\Gamma-4)-4d_x\}/3} h_m^{\{(d_x+1)(1-\Gamma)-4\}/3} (\log n)^{4/3} \to 0, \end{cases} \tag{6}$$

all as $n \to \infty$.

The last (three) condition(s) in Assumption F are the most restrictive. Sometimes higher order kernels are required. Regardless of the values of d_z and d_x, (sufficiently large) values of Φ, r can

be found such that Assumption F holds. For the case $\Gamma = 1$, all conditions are satisfied for $h_g = n^{-1/(2r+d_z)}, h_m = n^{-1/(2r+d_x+1)}$, provided r and Φ are chosen sufficiently large. Indeed, for fixed Φ sufficient conditions on r for the left column conditions in (6) to hold are $r \geq \sqrt{d_z} + I(d_x < 3)$ for the third left column condition and $r \geq \sqrt{R_\Phi d_z/2} + I(d_x < 2R_\Phi - 1)\{2R_\Phi - d_x - 1\}/2$, for the fourth left column condition, where $R_\Phi = (\Phi+1)/(\Phi-1)$. The first two left column conditions are satisfied for all $r \geq 2$ and all Φ. Sufficient conditions for the right column are $r \geq \sqrt{R_\Phi d_z/2} + I(d_x < 2R_\Phi - 1)\{R_\Phi - (d_x+1)/2\}$, $r \geq (d_x+1)/2 + \sqrt{R_\Phi d_z}/2 + (R_\Phi - 1)/2$, $r \geq \sqrt{d_z} + I(d_x < 3)$, and $r \geq \sqrt{d_x(d_x+d_z+1)/3} + I(5d_x > 3d_z+3)(5d_x - 3d_z - 3)/6$. Specifically, for $d_z = 3, d_x = 1$, one can choose $\Phi = 5$ $(R_\Phi = 3/2)$ and $r = 3$.

The second condition imposes a restriction on the function g. Let (x, z),the point at which a is to be estimated, be an interior point of $\mathcal{B}$.

Assumption G *g is monotonically increasing in its first argument in a neighborhood of z.*

Assumption G is more fundamental than Assumption F. Since m is monotonically increasing in g, it asks that a is monotonically increasing in at least one element of z. Without Assumption G, $g(z)$ could be on the boundary of the support of $g(Z_1)$, and boundary behavior of nonparametric kernel regression estimators is poor.

I am now in a position to posit three theorems which establish asymptotic normality and an optimal convergence rate for $\hat{a}_S$ under varying conditions on d_x and d_z. Let $\kappa_2 = \int k_{(1)}^2(u)du$.

Theorem 5 *Under Assumptions A–F, if $d_z < d_x + 1$, then (i) when $h_m \sim n^{-1/(2r+d_x+1)}$,*

$$n^{r/(2r+d_x+1)}\{\hat{a}_S(x,z) - a(x,z)\} = O_p(1),$$

and (ii) $\exists \varepsilon > 0$ such that for $h_m \sim n^{-1/(2r+d_x+1)-\varepsilon}$,

$$n^{r/(2r+d_x+1)-\varepsilon(d_x+1)/2}\{\hat{a}_S(x,z) - a(x,z)\} \xrightarrow{\mathcal{L}} N\left[0, \frac{\sigma_{XG}^2\{x, g(z)\}\kappa_2^{d_x+1}}{f_{XG}(x, g(z))}\right],$$

with $\sigma_{XG}^2\{x, g(z)\} = V\{Y_1 | X_1 = x, g(Z_1) = g(z)\}$.

Theorem 6 *Under Assumptions A–F, if $d_z > d_x + 1$,then (i) when $h_g \sim n^{-1/(2r+d_z)}$,*

$$n^{r/(2r+d_z)}\{\hat{a}_S(x,z) - a(x,z)\} = O_p(1),$$

and (ii) $\exists \varepsilon > 0$ *such that for* $h_g \sim n^{-1/(2r+d_z)-\varepsilon}$,

$$n^{r/(2r+d_z)-\varepsilon d_z/2}\left[\hat{a}_S(x,z) - a(x,z)\right] \xrightarrow{\mathcal{L}} N\left[0, \left[\frac{\partial m}{\partial g}\{x, g(z)\}\right]^2 \sigma_J^2(z)\kappa_2^{d_z}\right], \tag{7}$$

with $\sigma_J^2(z) = \int_{\mathcal{X}} \sigma_{XZ}^2(x,z)\lambda^2(x) f_{XZ}^{-1}(x,z)dx$.

Theorem 7 *Under Assumptions A–F, if* $d_z = d_x + 1$,*then (i) when* $h_g, h_m \sim n^{-1/(2r+d_z)}$,

$$n^{r/(2r+d_z)}\left\{\hat{a}_S(x,z) - a(x,z)\right\} = O_p(1),$$

and (ii) $\exists \varepsilon > 0$ *such that for* $h_g = h_m \sim n^{-1/(2r+d_z)-\varepsilon}$,

$$n^{r/(2r+d_z)-\varepsilon d_z/2}\left[\hat{a}_S(x,z) - a(x,z)\right] \xrightarrow{\mathcal{L}} N\left(0, \mathcal{V}\right),$$

where $\mathcal{V}$ *is the sum of the variance matrices in the previous two theorems.*

There are a number of remarks to be made here. First, from Theorems 5,6 and 7, it follows that the optimal convergence rate of $\hat{a}_S$ is $n^{-r/\{2r+\max(d_z,d_x+1)\}}$ when $h_g \sim n^{-1/(2r+d_z)}$ and $h_m \sim n^{-1/(2r+d_x+1)}$. Theorem 5 implies that under the conditions in the theorem, there is no (asymptotic) penalty for not knowing g. Under the conditions of Theorem 6, the convergence rate of $\hat{a}_S$ is identical to that of $\hat{g}$, which is the same as that of a nonparametric kernel regression estimator of a d_z–variate regression function. Theorems 5–7 say that the convergence rate of $\hat{a}_S$ is the slowest of the estimator of m (with g known) and $\hat{g}$.

Like other nonparametric kernel regression estimators, the convergence rate can be made arbitrarily close to $n^{-1/2}$ by choosing r large, provided a is very smooth. Theorems 5,6 and 7 are similar to the theorem in Rilstone (1996) in the sense that nonparametric generated regressors are used.[6]

Second, like in nonparametric kernel regression estimation, asymptotic normality only obtains under *undersmoothing*, i.e. when the bandwidth goes to zero at a faster rate than the rate at which the asymptotic mean square error is minimized. When the bandwidth goes to zero fast, the squared

[6]Rilstone's (1996) proof is incorrect and stronger assumptions are needed to obtain the stated results. I will outline the problem in my notation. While Rilstone shows that $\tilde{m}\{x, \hat{g}(z)\} - \tilde{m}\{x, g(z)\}$ has a limiting normal distribution where $\tilde{m}$ is the nonparametric kernel regression estimator of Y_i on $\{X_i, g(Z_i)\}$ and $\hat{g}$ is his estimator (quite different from mine) of his function g, Rilstone ignores the fact that the $g(Z_i)$'s are themselves estimated, also. This is the most difficult part of the proof. For another, correct, proof of a nonparametric generated regressor result, see Ahn (1997).

bias decreases faster than the variance with an increase in the sample size and hence does not impact the asymptotic distribution. In small or moderate samples, the bias can still be large and unless the bandwidth is chosen very small, bias reduction techniques like the *bootstrap* or *jackknife* may be appropriate.

Finally, the variance matrices in Theorems 5,6 and 7 can be estimated using nonparametric regression and density estimation.

Aside from the pointwise asymptotic normality results of Theorems 5, 6 and 7, it is also possible to establish a uniform convergence result for the estimator under weak separability. The rate of convergence is slower, particularly when d_x is small relative to d_z. Let $\Psi_{gn} = n^{-1/2}h_g^{-d_z/2}\log n + n^{-1}h_g^{-d_b}\log n + h_g^r$ denote the uniform convergence rate of $\hat{g}$.

Theorem 8 *Under Assumptions A, B, C, D and E,*

$$\sup_{b\in\mathcal{B}}\left|\hat{a}_S(b) - a(b)\right| = O_p\left(\Psi_{gn}h_m^{-1} + \Psi_{gn}^2 h_m^{-3} + n^{-1/2}h_m^{-(d_x+1)/2}\log n + h_m^r\right).$$

In particular, when $h_g \sim n^{-1/(2r+d_z)}, h_m \sim n^{-1/(2r+d_x+1)}$,

$$\sup_{b\in\mathcal{B}}\left|\hat{a}_S(b) - a(b)\right| = O_p\left(n^{-\nu_a}\log^2 n\right),$$

where

$$\nu_a = \frac{\min\left\{2r^2 - (1-d_x)r - d_z, 4r^2 - (4-d_x)r - 3d_z, 4r^2 + 2d_z r\right\}}{(2r+d_z)(2r+d_x+1)}.$$

A sufficient condition for $\nu_a > 0$, i.e. for convergence, is $r \geq \sqrt{3d_z}/2 + I(d_x < 4)$. For $d_x = 1, d_z = 2, r \geq 3$, $\nu_a = (r-1)/\{2(r+1)\}$ and the (nonuniform) convergence rate of $\hat{a}_S$ for the same bandwidth choice is $n^{-r/\{2(r+1)\}}$. So, for uniform convergence, a kernel of one order higher is needed to get the same approximate convergence rate for $\hat{a}_S$ when $d_x = 1, d_z = 2, r \geq 3$.

6 Choice of Input Parameters

In the proposed estimation method, the practitioner chooses four input parameters, i.e. a kernel, two bandwidths and the function λ. The choice of kernel shape is generally less important than the choice of bandwidth.

The choice of the bandwidth h_m is likely determined by similar concerns as the choice of the bandwidth in an ordinary nonparametric kernel regression problem. A second generation bandwidth choice algorithm which has been found to have good properties is Sheather and Jones (1991).

The other bandwidth h_g should primarily be chosen such as to maximize the accuracy of $\hat{g}$. Work on bandwidth choice for marginal integration estimators can be found in a number of sources including Horowitz (1998).

Now the choice of λ. Linton and Nielsen (1995) have studied the optimal choice of λ within the context of their estimator. Optimality in Linton and Nielsen (1995) is defined in terms of the density–weighted integrated mean square error. I allow the optimal choice of λ to depend on the point (x, z) at which a is to be estimated.

Allowing λ to depend on the point of estimation (x, z) leads to an asymptotic variance which is less than or equal to the asymptotic variance when λ is chosen the same for the whole range. Unfortunately, it also increases the computational burden if a is to be estimated at multiple points since the optimal λ needs to be determined for each individual point and for each λ all $\hat{g}(Z_i)$'s need to be recomputed. See section 8 for a discussion. I have failed to find an explicit solution for the function λ which, like in Linton and Nielsen (1995), minimizes some global measure of dispersion.

In Theorem 5, the choice of λ does affect the asymptotic variance. Although $\sigma^2_{XG}\{x, g(z)\}$ is not affected by the choice of λ, the choice of λ does affect $f_{XG}\{x, g(z)\}$. Indeed, if $g(z) = \chi_\lambda\{g_0(z)\}$, with χ a monotonically increasing differentiable transformation and g_0 one fixed choice for g, then $f_{XG}\{x, g(z)\} = f_{XG_0}\{x, g_0(z)\}/\chi'_\lambda\{g_0(z)\}$. Note however, that the only reason the asymptotic variance is affected is that choosing λ amounts to transforming one of the explanatory variables in the nonparametric kernel regression of Y_i on $\{X_i, g(Z_i)\}$; it is effectively a bandwidth choice. For Theorem 6, the situation is more interesting.

Theorem 9 *The limiting variance in Theorem 6 is minimized for*

$$\lambda(t) = \frac{\frac{\partial a}{\partial z_1}(t, z) f_{XZ}(t, z)}{\sigma^2_{XZ}(t, z)}.$$

The optimal choice of λ is intuitive. For g to be estimated accurately, more weight should be put at points at which the unconstrained estimator of a is the most accurate, i.e. at those points around which there are relatively many data points (large f_{XZ}) and small error variance σ^2_{XZ}.

The additional $\frac{\partial a}{\partial z_1}$ factor is caused by the $\partial m/\partial g$ factor in (7). Note that $\frac{\partial m}{\partial g} = \frac{\partial a}{\partial z_1} / \frac{\partial g}{\partial z_1}$; $\frac{\partial a}{\partial z_1}$ does not depend on λ but $\frac{\partial g}{\partial z_1}(z) = \int \frac{\partial a}{\partial z_1}(t,z)\,\lambda(t)\,dt$ does depend on λ. Note that $\frac{\partial a}{\partial z_1}$ could be equally replaced with $\frac{\partial a}{\partial z_i}$ for any i for which $\frac{\partial a}{\partial z_i}$ is everywhere positive. The only effect is a scale change in λ (z is fixed).

The optimal choice $\lambda(x) = \frac{\partial a}{\partial z_1}(x,z)\, f_{XZ}(x,z)\, \sigma_{XZ}^{-2}(x,z)$ depends on unknown quantities. λ can be estimated by first estimating $\frac{\partial a}{\partial z_1}, f_{XZ}, \sigma_{XZ}^2$ and then setting $\hat{\lambda}(x) = \frac{\partial \hat{a}}{\partial z_1}(x,z)\, \hat{f}_{XZ}(x,z)\, \hat{\sigma}_{XZ}^{-2}(x,z)$.

There are two problems with using an estimated weight function. The first is the computational problem mentioned earlier. The second concern is that the small sample performance of the estimator with estimated weight function may in fact be poorer than for a prudently chosen weight function, which does not depend on the data, in the same way that the feasible generalized least squares estimator can in practice be worse than the ordinary least squares estimator in the context of a linear regression model.

7 Testing for Weak Separability

Since $\hat{a}_S$ only estimates a consistently when a is weakly separable, it is important to establish whether a is indeed weakly separable, unless there is prior information that weak separability is indeed a reasonable assumption. There are many ways in which this can be done.

One possibility is to use the property of weakly separable functions that $(\partial a/\partial z_i)/(\partial a/\partial z_j)$ does not depend on x for any i, j. One could then test the hypothesis $\int V \left\{ \frac{\partial a}{\partial z_2}(X_1,z) / \frac{\partial a}{\partial z_1}(X_1,z) \right\} \lambda_\tau(z)\,dz = 0$ for some nonnegative weight function λ_τ (unrelated to λ) which ensures that the integral exists. I have not pursued this possibility because nonparametric derivative estimation is generally less accurate than nonparametric kernel regression estimation and the denominator of the integrand can be small.

Instead, it is preferable to use the test for additive separability of Gozalo and Linton (1997). Since the uniform convergence rate of $\hat{a}_S$ is faster than the pointwise convergence rate of $\hat{a}$, their results should carry over to the case of weak separability.

An alternative possibility is to use the test of independence of Pinkse (1999). If a is weakly separable and the errors are homoskedastic, then $Y_1 - \hat{a}_S(X_1, Z_1)$ is asymptotically independent of X_1, Z_1.

8 Computation

The proposed estimator $\hat{a}_S$ is computationally considerably more demanding than the unconstrained estimator $\hat{a}$. The way in which the computational burden increases with an increase in the sample size and the dimensionality of X_i differs depending on whether the marginal integration or the marginal summation estimator is used.

First the case of marginal integration. For the computation of $\hat{g}$ at each Z_i, one needs to marginally integrate $\hat{a}$ in d_x directions. This can be accomplished by most *quadrature*–based routines. To compute $\hat{a}$ takes $O(n)$ operations (assuming no *binning* or *fast Fourier transforms* are used). The time it takes to marginally integrate a function increases exponentially in the number of integration dimensions, so to compute $\hat{g}$ at a single point takes $O\left(n\Pi_3^{d_x}\right)$ operations, with $\Pi_3 > 1$ some positive constant. Suppose that a is to be estimated at n_a points (x, z) with different z. Then, $\hat{g}$ needs to be computed at $n + n_a$ different points. The last estimation step involves a number of steps which is of lower order. Hence, the number of operations required is $O\left\{(n + n_a)\, n\Pi_3^{d_x}\right\}$, compared to $O(nn_a)$ for the unconstrained estimator. Particularly when d_x is large, the computational burden can be substantial, although in my experience it takes less than an hour on a 200 MHz Pentium running NextStep for a single data set when $d_x = 4$ and $n = 500$.

When λ is unknown and needs to be estimated, or indeed is chosen differently for each z, the whole procedure needs to be repeated for each choice of z, and hence the total number of operations is $O\left\{n_a n^2 \Pi_3^{d_x}\right\}$, compared to $O(nn_a)$ for $\hat{a}$. z–dependent λ could still be feasible for individual data sets when n, n_a and d_x are small, but for the simulation study in section 10 it is too demanding.

In the case of marginal summation, it takes $O\left(n^2\right)$ operations to compute $\hat{g}$ at a single point. The total number of operations is hence $O\left\{(n + n_a)\, n^2\right\}$. So marginal summation is preferable when n is relatively small and d_x is relatively large. Marginal summation is in this paper limited to the choice $\lambda = f_X$, but see section 9 for possible extensions.

9 Extensions

There are many ways in which the ideas put forward in this paper can be extended. Section 9.1 discusses extensions to the model (1). In section 9.2 I discuss a way in which different choices of λ can be implemented in the context of the marginal summation estimator and section 9.3 looks at

alternative nonparametric regression estimation techniques that could be used.

9.1 Convenient Forms of Weak Separability

It is possible to extend the model (1) to say,

$$a(\upsilon) = \Upsilon_0 \left[\Upsilon_1 \left\{ \Upsilon_{11}(\upsilon_{11}), \Upsilon_{12}(\upsilon_{12}), \ldots \right\}, \Upsilon_2 \left\{ \cdot \right\}, \ldots \right]. \tag{8}$$

One could allow for an arbitrarily high level of nesting, but despite the asymptotic results that obtain, small sample performance is likely to deteriorate with both d_υ and the level of nesting. Υ_{1j} can be identified by $\Upsilon_{1j}(\upsilon_{1j}) = \int a(\upsilon)\lambda(\upsilon_{-1j})d\upsilon_{-1j}$ where υ_{-1j} denotes all elements of υ which are not in υ_{1j}. Intermediate functions can be estimated by repeated use of the estimation methods described in Sections 4 and 5 using say $\int \hat{a}(\upsilon)\lambda(\upsilon_{-1})d\upsilon_{-1}$as the dependent variable and $\hat{\Upsilon}_{1j}(\upsilon_{1j})$ for various values of j as explanatory variables.

Although the results in this paper are for (1), some preliminary explorations have shown that extensions like (8) follow relatively easily, albeit under stronger conditions. In principle, then, if I knew that the regression function a had a particularly convenient (i.e. nested) weakly separable form, I could estimate a with an estimator which had the same convergence rate as that of a bivariate nonparametric kernel regression estimator, regardless of the number of arguments. Under generalized additive separability the convergence rate of the Linton and Härdle (1996) and Horowitz (1998) estimators compare to that of a univariate nonparametric kernel regression estimator.

Even if a has the convenient form mentioned above, the above–described estimator which uses it is not likely to be very good in even fairly large samples in view of the compounded approximation errors.

9.2 Marginal Summation

The marginal summation estimator $\hat{g}_\Sigma$ assumes $\lambda = f_X$. Other choices of λ could be implemented by replacing $\hat{g}_\Sigma$ with a weighted equivalent, say

$$\hat{g}_{\Sigma w}(z) = n^{-1} \sum_{i=1}^{n} \hat{a}(X_i, z) I(X_i \in \mathcal{B}_X) \lambda_w(X_i),$$

which uses $\lambda = f_X \lambda_w$ as the weight function.

9.3 Other Nonparametric Estimation Techniques

It is possible to obtain results similar to those discussed here for other nonparametric estimation methods. I discuss three here: local polynomial estimation, K–nearest neighbor estimation and series estimation.

The potential benefits of local polynomial estimation over kernel regression estimation are well–documented (see Fan and Gijbels, 1996, for an overview). They include improved estimation of peaks and troughs and, in some circumstances, at boundaries.

Similar results to those obtained in this paper can be obtained for local polynomial estimators at the expense of some complications in and lengthening of the proofs. There is no essential difference in the current context between asymptotics for kernel regression estimators and local polynomial estimators.

K–nearest neighbor estimators (Fix and Hodges, 1951; see Stone, 1977, for some fundamental results) could be used in both steps. Because of discontinuities the use of generated regressors is considerably more complex and it is unclear whether similar results obtain. For marginal integration K–nearest regression estimation could in fact be simpler since there is no denominator term to contend with. I have experimented a little with both but failed to get comparable results to that for the kernel regression estimation method employed here.

One can impose weak separability on a series expansion of a regression function. Except for the fact that the number of terms is allowed to increase with the sample size this is similar to parametric specifications used in the past.

It is also possible to use series estimation only in the determination of $\hat{g}$. One particularly convenient way this can be accomplished is by letting $a(x,z) = \zeta_S(x)\sum_{j=1}^{\infty}\alpha_j e_j(z) + \sum_{i=2}^{\infty}\sum_{j=1}^{\infty}\alpha_{ij}^* e_{Xi}(x)e_j(z)$. The functions ζ, e_j, e_{Xi}, $i = 2,\ldots\infty, j = 1,\ldots\infty$ are chosen by the practitioner so that they form a basis for the function class a belongs to. ζ_S can be chosen such that $\int \zeta_S(x)e_{Xi}(x)dx = 0$ for all i. Then set $\lambda = \zeta_S$ and employ the normalization $\int \lambda^2(x)dx = 1$ to obtain $g(z) = \sum_{j=1}^{\infty}\alpha_j e_j(z)$.

From Andrews (1991), it follows that g can be estimated more efficiently than a. The estimate of g thus obtained can then be used as a generated regressor in the last step of my estimation procedure.

10 Simulations

The simulation study in this section has three goals. To find out in which type of models $\hat{a}_S$ performs well relative to $\hat{a}$, how the sample size affects the relative performs and how performance is affected by a change in dimensionality. My study therefore includes several combinations of model structure, sample size and dimensions (d_x and d_z).

Except where indicated, 100 data sets are drawn for each specification, sample size, dimension combination and the regression functions are estimated at 50 randomly selected points $\left(\tilde{X}_{ji}, \tilde{Z}_{ji}\right)$ (each of which is independently drawn from the same distribution as the (X_{ji}, Z_{ji}) pair used in the data). While 100 is a relatively small number, what is of interest here is a comparison of means or distributions, which takes far fewer replications then say, determining the rejection rate of a test statistic under the null hypothesis, i.e. a *tail probability*.

A pseudo–RMSE figure is then computed for each data set, using

$$RMSE_{\hat{a}j} = \sqrt{\frac{1}{50}\sum_{i=1}^{50}\left\{\hat{a}\left(\tilde{X}_{ji}, \tilde{Z}_{ji}\right) - a\left(\tilde{X}_{ji}, \tilde{Z}_{ji}\right)\right\}^2}.$$

A lower RMSE number is better. The RMSE results are ordered (by increasing magnitude) and plotted against the RMSE results of other specifications/estimation methods. So in each case the lowest RMSE number of one specification is matched with the lowest RMSE number of another, the next lowest with the next lowest and so on. The results are in the figures which I discuss further below.

In all cases the regressors have a joint normal distribution with moderate correlations. The errors are drawn from a normal distribution and independently from the regressors; I do not study the effects of heteroskedasticity here. No truncation was used, which violates the compactness condition of Assumption B.

In all scenarios a standard normal kernel was used, all regressors were normalized by dividing through by their standard deviations and bandwidths were chosen by setting $h = n^{-1/(4+d)}$, where d denotes the dimension of the entire regressor vector used in that particular nonparametric regression. While not satisfying Assumption F for all choices of d_x, d_z, typically lower order kernels give more accurate resuts in small to moderate samples. Unless otherwise specified, λ is positive and constant on $[-2, 2]$ and zero elsewhere. All simulations were carried out in the programming language C on

a 200 Mhz Pentium running NeXTSteP 3.3.

The specifications used were the following.

Model 1. $m(x,g) = \sum_{i=1}^{d_x} x_i + g$, $g(z) = \sum_{i=1}^{d_z} z_i$.

Model 2. $m(x,g) = x_1 g + x_2 g^2$, $g(z) = z_1(z_2+1)$.

Model 3. $m(x,g) = x_1 g + x_2 g^2$, $g(z) = z_1 + z_2$.

Model 4. $m(x,g) = \sqrt{g\sqrt{x_1^2+1} + x_2\sqrt{g}}$, $g(z) = \sqrt{\sqrt{z_1^2+1} + \sqrt{z_2^2+1}}$,

Model 5. $m(x,g) = \log\left[\{(x_1+x_2)-g\}^2 + 1\right]$, $g(z) = z_1 + z_2$.

All specifications studied are weakly separable. The different specifications were chosen to determine which functional forms lead to the best performance. The functional forms chosen are not necessarily models one would ever encounter in reality.

Figures 1 and 2 show how changes in sample size affect the performance of $\hat{a}$ and $\hat{a}_S$ in model 1 using marginal integration. The figures are the same except that in figure 2, $d_x = 2$ instead of $d_x = 1$. On the horizontal axis are the RMSE numbers for $\hat{a}$ and $\hat{a}_S$ for $n = 100$. On the vertical axis are the same numbers for $n = 200$ and $n = 500$. Figures 1 and 2 show how performance varies with n; they contain no information about the relative performance of $\hat{a}_S$ vis–a–vis $\hat{a}$. In figure 1, the performance improvement of $\hat{a}$ and $\hat{a}_S$ for a sample of size 200 over one of size 100 appears roughly the same. For $n = 500$ the improvement for $\hat{a}_S$ appears greater than that for $\hat{a}$, suggesting that the asymptotically superior performance of $\hat{a}_S$ starts becoming noticeable at a sample size between 200 and 500.

For the data set of figure 2, $\hat{a}_S$ has the convergence rate of a nonparametric regression estimator with three regressors instead of two and $\hat{a}$ is now a four–dimensional nonparametric regression estimator, as opposed to three–dimensional as in figure 1. The improvement in performance is now noticeable at $n = 200$. These results suggest that the smaller is $\max\{d_x+1, d_z\}/(d_x+d_z)$, the earlier are the gains of $\hat{a}_S$ realized.

Figures 3, 4 and 5 compare the performance of $\hat{a}_S$ using marginal summation (MS) and marginal integration (MI) for model 1 for sample sizes 100, 200 and 500 and $d_x = 1, d_z = 2$. With MI, $\lambda(x) = 1$ and with MS $\lambda(x) = f_X(x)$. The optimal choice of λ is here $\lambda(x) = f_{XZ}(x,z)$, more in line with MS, and my expectation is that MS will do better. In each case, the horizontal axis shows

the RMSE of $\hat{a}$ and the vertical axis the RMSE of both variants of $\hat{a}_S$. To facilitate a comparison, the three graphs also contain shaded bars; the closer the bars are together, the more often the RMSE of $\hat{a}$ takes values in that region.

In all three figures, both variants of $\hat{a}_S$ have RMSE graphs which are below the 45–degree line. Hence, both MI and MS outperform $\hat{a}$, even at $n = 100$. MI and MS perform similarly at $n = 100$ but at $n = 200$ and $n = 500$ MS appears to do a little better, although the difference is negligible compared to the difference between either MS or MI and $\hat{a}$. These conclusions are also borne out by figure 6, the equivalent of figure 2 for the marginal summation estimator.

Figures 7 and 8 contain a comparison of $\hat{a}_S$ and $\hat{a}$ for the five models. In each case the sample size is 100 and $d_x = d_z = 2$. In all cases MS is used. The graphs show that $\hat{a}_S$ does better for models 1 and 4, slightly worse for model 5 and worse for models 2 and 3. These results suggest that the greater the degree of nonlinearity, the greater the sample size at which $\hat{a}_S$ starts becoming preferable over $\hat{a}$. Further experiments (not graphed) have shown that even at $n = 500$, $\hat{a}$ performs better than $\hat{a}_S$.

Part of the problem is the choice of λ. In figure 9, I have plotted the results where the experiments with models 2 and 3 are carried out again but λ is chosen equal to $f_X \partial a/\partial z_1$.[7] This choice of λ ignores the dependence of the optimal choice on f_{XZ} but nonetheless demonstrates the dependence of performance on the choice of λ. For model 2, the perfomance of $\hat{a}$ and $\hat{a}_S$ are very similar. For model 3 $\hat{a}_S$ still does worse than $\hat{a}$ but performance appears somewhat better. Not graphed is a comparison of the performance of $\hat{a}_S$ and $\hat{a}$ in model 2 with $\lambda = f_X \partial a/\partial z_1$ and $n = 200$ (instead of $n = 100$); $\hat{a}_S$ does a little better than $\hat{a}$.

Finally, a comparison across $d_x + d_z$. Figures 10 and 11 are identical except that figure 10 has $n = 100$ and figure 11 has $n = 200$. In both cases MS is used and both figures apply to model 1. As expected, accuracy decreases with an increase in the number of dimensions. Clearly, 16–dimensional nonparametric regression using a sample of size 100 is not advisable, whether weak separability is imposed or not. Asymptotically, $\hat{a}_S$ with $d_x = d_z = 8$ should do a little worse than $\hat{a}$ with $d_x = d_z = 4$, but asymptotics appear to have little bearing on figure 10. The conclusion re figure 10 must be that $\hat{a}_S$ does better than $\hat{a}$ but not as much as asymptotics has one believe. Figure 11 suggests that asymptotics have still not taken full effect at $n = 200$ but that a dimension reduction

[7] The graphs are a little jerkier because the number of replications was smaller here.

appears attainable in samples of a few hundred observations.

11 Conclusions

I have proposed a nonparametric kernel regression estimator under weak separability. The estimator uses marginal integration to obtain dimension reduction like estimation methods using the stronger concept of (generalized) additive separability.

Weak separability is a considerably weaker condition than additive separability because it allows for more interaction between regressors. The downside is that because of the very nature of the weak separability condition the dimensionality can at best be reduced to that of a two–dimensional problem instead of a one–dimensional problem as is the case under additive separability.

Simulation results suggest that the proposed estimator generally does better than the unconstrained estimator, even in small samples, provided that the weight function λ is chosen appropriately. However, the degree of dimension reduction promised by asymptotics is not generally realized in small samples.

12 References

- Ahn, H. (1997), "Semiparametric estimation of a single–index model with nonparametrically generated regressors," Econometric Theory 12, 3–31.
- Andrews, D.W.K. (1987), "Consistency in nonlinear models: a generic uniform law of large numbers," Econometrica 55, 1465–1471.
- Andrews, D.W.K. (1991), "Asymptotic normality of series estimators for nonparametric and semiparametric regression models," Econometrica 59, 307–345.
- Andrews, D.W.K. (1992), "Generic uniform convergence," Econometric Theory 8, 241–257.
- Barnett, W.A. (1980), "Economic monetary aggregation: an application of index number and aggregation theory," Journal of Econometrics 14, 11–48.
- Blackorby, C., R. Davidson and W. Schworm (1991), "Implicit separability: characterisation and implications for consumer demands," Journal of Economic Theory 55, 364–399.
- Blackorby, C. and W. Schworm (1988), "The existence of input and output aggregates in aggregate production functions," Econometrica 56, 613–643.
- Blackorby, C., Schworm, W. and T. Fisher (1986), "Testing for the existence of input aggregates in an economy production function," UBC discussion paper 86–26.
- Blundell, R. (1988), "Consumer behaviour: theory and empirical evidence – a survey," Economic Journal 98, 16–65.
- Blundell, R. and J.–M. Robin (1997), "Latent separability: grouping goods without weak separability,", forthcoming in Econometrica.
- Burkholder, D.L. (1973), "Distribution function inequalities for martingales," Annals of Probability 1, 19–42.
- Delgado, M. and J. Mora (1995), "Nonparametric and semiparametric estimation with discrete regressors," Econometrica 63, 1477–1482.

- Diewert, W.E. and T.J. Wales (1987), "Flexible functional forms and global curvature conditions," Econometrica 55, 43–68.
- Diewert, W.E. and T.J. Wales (1988), "Quadratic spline models for producer's supply and demand functions," International Economic Review 33, 705–722.
- Diewert, W.E. and T.J. Wales (1992), "Normalized quadratic systems of consumer demand functions.", Journal of Business and Economic Statistics 6, 303–312.
- Diewert, W.E. and T.J. Wales (1995), "Flexible functional forms and tests of homogeneous separability," Journal of Econometrics 67, 259–302.
- 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.
- Fix, E. and J.L. Hodges (1951), "Discriminatory analysis, nonparametric estimation: consistency properties," Report Number 4, Project No. 21–49–004, U.S. Air Force School of Aviation Medicine (Randolph Field, Texas).
- Friedman, J.H. and W. Stützle (1981), "Projection pursuit regression,", Journal of the American Statistical Association 76, 817–823.
- Gozalo, P.L. (1993), 'A Consistent Model Specification Test for Nonparametric Estimation of Regression Function Model," Econometric Theory 9, 451–477.
- Gozalo, P.L. and O.B. Linton (1997), "Testing Additivity in Generalized Nonparametric Regression Models," Yale University working paper.
- Hall, P. (1984), "Central limit theorem for integrated square error of multivariate nonparametric density estimators," Journal of Multivariate Analysis 14, 1–16.
- Hall, P. and C.C. Heyde (1980), "Martingale limit theory and its application," Academic Press (New York).

- 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. (1998), "Nonparametric estimation of a generalized additive model with an unknown link function," University of Iowa mimeo.
- Konakov, V.D. (1977), "On a global measure of deviation for an estimate of the regression line,", Theory of Probability and its Applications 22–4, 858–868.
- Liero, H. (1992), "Asymptotic normality of a weighted integrated squared error of kernel regression estimates with data–dependent bandwidth," Journal of Statistical Planning and Inference 30, 307–325.
- 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.
- 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. (1991), "Uniform convergence in probability and stochastic equicontinuity," Econometrica 59, 1161–1167.
- Nielsen, J.P. and O.B. Linton (1997), "An optimization interpretation of integration and backfitting estimators for separable nonparametric methods," Yale University mimeo.
- Pinkse, J. (1999), "Nonparametric misspecification testing," UBC working paper.
- Pollard, D. (1984), "Convergence of stochastic processes", Springer Verlag, New York.

- Pötscher, B.M. and I.R. Prucha (1989), "A uniform law of large numbers for dependent and heterogeneous data processes," Econometrica 57, 675–683.
- 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.
- Rosenblatt, M. (1975), "A quadratic measure of deviation of two–dimensional density estimates and a test of independence," Annals of Statistics 3, 1–14.
- 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. (1977), "Consistent nonparametric regression," Annals of Statistics 5, 595–645.
- Tjøstheim, D. and B. Auestad (1994), "Nonparametric identification of nonlinear time series: projections," Journal of the American Statistical Association 89, 1398–1409.
- Watson, G.S. (1964), "Smooth regression analysis," Sankhyā A26, 359–372.
- Woodland, A.D. (1978), "On testing weak separability," Journal of Econometrics 8, 383–398.

A Proofs

Proof of Theorem 1

I need to show that (i) any weakly separable function a can be written as $m\{x, g(z)\}$ with the identification condition imposed and that (ii) the identification condition uniquely identifies m, g.

From the definition of weak separability it follows that unknown functions m^*, g^* exist such that for all x, z, $a(x,z) = m^*\{x, g^*(z)\}$ with m^* monotonic in g^*. Thus,

$$g(z) = \int_{\mathcal{X}} a(x,z)\lambda(x)dx = \int_{\mathcal{X}} m^*\{x, g^*(z)\}\lambda(x)dx = \vartheta\{g^*(z)\},$$

for some function ϑ which is monotonic because of the monotonicity of m^* with respect to g^*. Hence g is a monotonic transformation of g^*. Thus, $a(x,z) = m\{x, g(z)\}$ with $m(x,g) = m^*\{x, \vartheta^{-1}(g)\}$. Hence both (i) and (ii) hold. □

Lemma 1 *If for some* $p_\zeta > 0$, $\limsup_{n\to\infty} \max_{1\le i\le n} E|A_{ni}|^{p_A} \le C_A$, *then* $P(\max_{1\le i\le n}|A_{ni}| > M_n) \le nC_A M_n^{-p_A}$.

Proof: Note that $P(\max_{1\le i\le n}|A_{ni}| > M_n) \le \sum_{i=1}^n P(|\zeta_{ni}| > M_n)$. Apply the Markov inequality. □

Lemma 2 *Let* $\mathcal{G} = \{\zeta_{nt} : t \in \mathcal{T}\}$ *be a class of functions with polynomial discrimination. Let* $\{\xi_i\}$ *be a sequence of i.i.d. random variables and set* $\sigma^2_{\zeta n} = \sup_{t\in\mathcal{T}} V\{\zeta_{nt}(\xi_1)\} < \infty$.. *If a* $p_\zeta > 0$ *exists such that* $\limsup_{n\to\infty} E\{\sup_{t\in\mathcal{T}}|\zeta_{nt}(\xi_1)|^{p_\zeta}\} < \infty$ *and for some fixed* $1/2 < \iota < 1$, $n^{1/2-1/p_\zeta}\sigma_{\zeta n}(\log n)^{\iota/2} \to \infty$ *as* $n \to \infty$ *then* $\sup_{t\in\mathcal{T}}\left|n^{-1}\sum_{i=1}^n \zeta_{nt}(\xi_i) - E\{\zeta_{nt}(\xi_1)\}\right| = o_p\left\{n^{-1/2}\sigma_{\zeta n}(\log n)^{\iota}\right\}$ *and if* $n^{1/2-2/p_\zeta}\sigma_{\zeta n}(\log n)^{\iota/2} \to \infty$ *as* $n \to \infty$ *then* $\sup_{t\in\mathcal{T}}\left|n^{-1}\sum_{i=1}^n \zeta_{nt}(\xi_i) - E\{\zeta_{nt}(\xi_1)\}\right| = o_{a.s.}\left\{n^{-1/2}\sigma_{\zeta n}(\log n)^{\iota}\right\}$.

Proof: This fixes a minor oversight in Horowitz, Lemma 1, at the expense of the loss of a.s. convergence in lieu of convergence in probability. The problem is in the choice of $\delta_n = \sigma_f/n$ and $\alpha_n = n^{-1/2}\delta_n^{-1}\log n = \sigma_f^{-1}n^{1/2}\log n$, which violates Pollard's (1984, Theorem II.37)[8] condition that α_n be non–increasing. Instead, for any $\iota > 1/2$ choose $\delta_n = n^{-1/2}(\log n)^{\iota}$ after multiplying all $f(Z_i)$'s (Horowitz notation) by $\sigma_f^{-1}n^{-1/2}(\log n)^{\iota}$ instead of by n^{-1} as in Horowitz. For the convergence in probability result, replace the treatment of T_{n2} in Horowitz with Lemma 1 with

[8]Other results on uniform convergence, whose conditions are often easier to verify, are Andrews (1987,1992), Newey (1991) and Pötscher and Prucha (1989).

$A_{ni} = \sup_{f \in \mathcal{J}_n} |f(Z_i)|$ and $M_n = n^{1/2}\sigma_f (\log n)^{-\iota}$ (Horowitz's notation) or $A_{ni} = \sup_{t \in \mathcal{T}} |\zeta_{nt}(\xi_i)|$ and $M_n = n^{1/2}\sigma_{\zeta n}(\log n)^{-\iota}$ (my notation). For a.s. convergence, (similar to Horowitz) use the fact that $\sum_{n=1}^{\infty} P(\max_{1 \le i \le n} |A_{ni}| > M_n) < \infty$. Apply Borel–Cantelli.□

Proof of Theorem 2

In the proof I omit arguments wherever possible without causing confusion. Let $\hat{N}_a$ and $\hat{D}_a$ be the numerator and denominator of $\hat{a}$ respectively. Then at all $b \in \mathcal{B}$,

$$\hat{a} - a = \frac{(\hat{N}_a - N_a)D_a - N_a(\hat{D}_a - D_a)}{D_a^2} \sum_{i=0}^{\infty} \left(\frac{D_a - \hat{D}_a}{D_a} \right)^i. \tag{9}$$

The expansion is allowed because of the well–established uniform convergence of kernel density estimators (denominator) and the fact that the joint density is assumed bounded away from zero in Assumption A. From Pollard (1984), Theorem 2.37 and my adaptation (Lemma 2 observing that $\sigma_{\zeta n} = O\left(h_g^{d_b/2}\right)$ and $p_\zeta = p$) of Horowitz (1998), Lemma 1 and a standard kernel bias expansion, it follows that

$$\sup_{b \in \mathcal{B}} |\hat{N}_a - N_a| + \sup_{b \in \mathcal{B}} |\hat{D}_a - D_a| = o_p(n^{-1/2} h_g^{-d_b/2} \log n) + O_p(h_g^r).$$

Thus,

$$\sup_{b \in \mathcal{B}} \left| \hat{a} - a - \frac{\hat{N}_a D_a - N_a \hat{D}_a}{D_a^2} \right| = o_p(n^{-1} h_g^{-d_b} \log n) + O_p(h_g^{2r}).$$

But,

$$\int \frac{\hat{N}_a(x,z)D_a(x,z) - N_a(x,z)\hat{D}_a(x,z)}{D_a^2(x,z)} \lambda(x)dx = n^{-1}h_g^{-d_z} \sum_{i=1}^{n} k_{h_g}(z - Z_i) J_{ni}(z) = \tilde{g}(z), \tag{10}$$

where $J_{ni}(z) = h_g^{-d_x} \int k_{h_g}(x - X_i) \{Y_i - a(x,z)\} \lambda(x) / f_{XZ}(x,z)dx$. Finally, apply Lemma 2 again to the right hand side in (10). □

Proof of Theorem 3

Let $S_n(t) = \sum_{s=1}^{n} \sum_{i=1}^{s-1} U_{si}(t)$ with $U_{si}(t) = \tilde{U}_{si}(t) - E\left\{\tilde{U}_{si}(t) | \xi_i\right\} - E\left\{\tilde{U}_{si}(t) | \xi_s\right\} + E\left\{\tilde{U}_{si}(t)\right\}$. I first show that $\sup_{t \in \mathcal{T}} \left| \tilde{S}_n(t) - S_n(t) \right| = o_p\left(n^{3/2}\sigma_{UCn} \log n\right)$ and that under the stronger moment condition, $\sup_{t \in \mathcal{T}} \left| \tilde{S}_n(t) - S_n(t) \right| = o_{\text{a.s.}}\left(n^{3/2}\sigma_{UCn} \log n\right)$. Note that $\tilde{S}_n(t) - S_n(t) = (n-1)\sum_{s=1}^{n} \left[E\left\{\tilde{U}_{si}(t) | \xi_s\right\} - E\left\{\tilde{U}_{si}(t)\right\} \right]$ (with i implicitly different from s). Apply Lemma 2.

I now show that $\sup_{t\in\mathcal{T}}|S_n(t)| = o_{\text{a.s.}}\{n\sigma_{Un}\log n\}$. Set $\varepsilon_n = n\sigma_{Un}\log n$. Since $\mathcal{F}$ is of polynomial discrimination I need to show that for any fixed t, $P(|S_n(t)| > 2\varepsilon_n)$ decreases faster with n than any power of n. Choose arbitrary t. I omit the dependence on t in my notation from hereon. Let $T_s = \sum_{i=1}^{s-1} U_{si}$ and set $I_s = I\left(\sum_{i=1}^{s-1} T_i^2 < \delta_n\right)$ with $\delta_n = n^2\sigma_{Un}^2(\log n)^{\iota}$. Thus, $P(|S_n| > 2\varepsilon_n) \leq P(|\sum_{s=1}^n T_s I_s| > \varepsilon_n) + P\{|\sum_{s=1}^n T_s(1-I_s)| > \varepsilon_n\}$.

I first deal with the second majorant term. By the Markov and Schwarz inequalities

$$\begin{aligned}\varepsilon_n P\left\{\left|\sum_{s=1}^n T_s(1-I_s)\right| > \varepsilon_n\right\} &\leq E\left|\sum_{s=1}^n T_s(1-I_s)\right| \leq \sum_{s=1}^n E|T_s(1-I_s)| \\ &\leq \sum_{s=1}^n \sqrt{E(T_s^2)P\left(\sum_{i=1}^{s-1} T_i^2 > \delta_n\right)} \leq \sqrt{nP\left(\sum_{s=1}^n T_s^2 > \delta_n\right)\sum_{s=1}^n E(T_s^2)}.\end{aligned}$$

Since $\sum_{s=1}^n E\left(T_s^2\right)$ and ε_n increase no faster than a power of n, it suffices to show that $P\left(\sum_{s=1}^n T_s^2 > \delta_n\right)$ decreases faster than any power of n. Note that

$$\begin{aligned}P\left(\sum_{s=1}^n T_s^2 > \delta_n\right) &= P\left\{\sum_{s=1}^n\left(\sum_{i=1}^{s-1} U_{si}\right)^2 > \delta_n\right\} \leq nP\left(\left|n^{-1}\sum_{i=1}^{s-1} U_{si}\right| > \sqrt{\tfrac{\delta_n}{n^3}}\right) \\ &= nP\left\{\left|n^{-1}\sum_{i=1}^{s-1} U_{si}\right| > n^{-1/2}\sigma_{Un}(\log n)^{\iota}\right\}.\end{aligned}$$

Use the fact that the right hand side probability decreases faster than any power of n by Lemma 2 in conjunction with Theorem II.37 of Pollard (1984).

Now consider $P(|\sum_{s=1}^n T_s I_s| > \varepsilon_n)$. Set $\check{S}_n = \sum_{s=1}^n T_s I_s$. Now,
$P\left(\left|\check{S}_n\right| > \varepsilon_n\right) = P\left(\check{S}_n > \varepsilon_n\right) + P\left(-\check{S}_n > \varepsilon_n\right) \leq e^{-\varepsilon_n^2/(2\delta_n)}\left[E\left\{e^{\varepsilon_n\check{S}_n/(2\delta_n)}\right\} + E\left\{e^{-\varepsilon_n\check{S}_n/(2\delta_n)}\right\}\right]$.
But, by the Burkholder inequality (Burkholder, 1973, Hall and Heyde, 1980 and Davidson, 1994),

$$\begin{aligned}E\left\{e^{\varepsilon_n\check{S}_n/(2\delta_n)}\right\} + E\left\{e^{-\varepsilon_n\check{S}_n/(2\delta_n)}\right\} &\leq 2\sum_{j=0}^{\infty}\frac{\{\varepsilon_n/(2\delta_n)\}^{2j}E\left(\check{S}_n^{2j}\right)}{(2j)!} \leq 2\sum_{j=0}^{\infty}\frac{C_j}{(2j)!/j!}\frac{\left\{\varepsilon_n^2/\left(4\delta_n^2\right)\right\}^j E\left(\check{Q}_n^j\right)}{j!} \\ &\leq \mathcal{K}E\left\{e^{\varepsilon_n^2\check{Q}_n/\left(4\delta_n^2\right)}\right\},\end{aligned}$$

with $\{C_j\}$ the Burkholder constants, $\mathcal{K} = 2\sup_j C_j(j!)/(2j)!$ and $\check{Q}_n = \sum_{s=1}^n T_s^2 I_s \leq \delta_n$ by construction. Hence $P\left(\left|\check{S}_n\right| > \varepsilon_n\right) \leq \mathcal{K}e^{-\varepsilon_n^2/(2\delta_n)}e^{\varepsilon_n^2/(4\delta_n)} = \mathcal{K}e^{-\varepsilon_n^2/(4\delta_n)} = \mathcal{K}e^{-\varepsilon_n^2/(4\delta_n)} = \mathcal{K}n^{-(\log n)^{1-\iota}/4} \to 0$ faster than any power of n. □

Proof of Theorem 4

Instead of the integral on the left hand side in (10) I need to look at (making the implicit

assumption that X_s's outside $\mathcal{B}_X$ are omitted in the sum),

$$n^{-2}\sum_{s=1}^{n}\sum_{i\neq s}^{n}h_g^{-d_b}k_{h_g}\left(z-Z_i\right)\left[k_{h_g}\left(X_s-X_i\right)\left\{\frac{Y_i-a\left(X_s,z\right)}{f_{XZ}\left(X_s,z\right)}\right\}-\int_{\mathcal{B}_X}k_{h_g}\left(x-X_i\right)\left\{\frac{Y_i-a\left(x,z\right)}{f_{XZ}\left(x,z\right)}\right\}f_X(x)dx\right]. \tag{11}$$

Expression (11) is almost in the required form to apply Theorem 3. The only problem is that the U–statistic in (11) is not symmetric in ξ_i,ξ_s with $\xi_i=(X_i,Z_i)$. Denote the summand in (11) by $\check{U}_{nsi}\left(z\right)$ and set $\bar{U}_{nsi}\left(z\right)=\check{U}_{nsi}\left(z\right)+\check{U}_{nis}\left(z\right)$, $\tilde{U}_{nsi}\left(z\right)=\bar{U}_{nsi}\left(z\right)-E\left\{\bar{U}_{nsi}\left(z\right)|\xi_i\right\}-E\left\{\bar{U}_{nsi}\left(z\right)|\xi_s\right\}+E\left\{\bar{U}_{nsi}\left(z\right)\right\}$. Then (11) is $n^{-2}\sum_{s=1}^{n}\sum_{i=1}^{s-1}\tilde{U}_{nsi}\left(z\right)+n^{-1}\sum_{s=1}^{n}\left[E\left\{\bar{U}_{nsi}\left(z\right)|\xi_s\right\}-E\left\{\bar{U}_{nsi}\left(z\right)\right\}\right]$. The second right hand side term is $O_p\left(h_g^r\right)$. Apply Theorem 3 to the first right hand side term, noting that $\sigma_{Un}=h_g^{-d_b/2}$ and $\sigma_{UCn}=0$. □

Lemma 3 *For* $t=2,\ldots,\Phi-1$,

$$n^{-1}h_m^{-d_x-1-t}\sum_{i=1}^{n}k_{h_m}\left(x-X_i\right)k_{h_m}^{(t)}\left\{g\left(z\right)-g\left(Z_i\right)\right\}\frac{Y_i-a(x,z)}{f_{XG}(x,g(z))}\left\{\hat{g}(z)-g(z)\right\}-\left\{\hat{g}(Z_i)-g(Z_i)\right\}^t$$
$$=o_p\left\{h_m^{-t}\left(n^{-1}h_g^{-d_b}\log n+n^{-1.2}h_g^{-d_z}\log n\right)^t\right\}+O_p\left(h_m^{-t}h_g^{rt}\right).$$

Proof: Use Theorem 2 to get uniform convergence results on the $\hat{g}$'s. Take absolute values of the remainder of the summand and take expectations, making the substitution $s=(t-x)/h_m$, $s_g=\left\{t_g-g\left(z\right)\right\}/h_m$, where t,t_g are the integration variables. □

Lemma 4 $n^{-1}h_m^{-d_x-1-\Phi}\sum_{i=1}^{n}k_{h_m}\left(x-X_i\right)k_{h_m}^{(\Phi)}\left\{\cdot\right\}\frac{Y_i-a(x,z)}{f_{XG}(x,g(z))}\left[\left\{\hat{g}(z)-g(z)\right\}-\left\{\hat{g}(Z_i)-g(Z_i)\right\}\right]^{\Phi}$
$=o_p\left\{h_m^{-\Phi-1}\left(n^{-1}h_g^{-d_b}\log n+n^{-1/2}h_g^{-d_z}\log n\right)^{\Phi}\right\}+O_p\left(h_m^{-1-\Phi}h_g^{r\Phi}\right)$.
Proof: Like Lemma 3 with $t=\Phi$, but using the fact that $k_{h_m}^{(\Phi)}$ is bounded instead of integrating in that direction. □

Lemma 5 *Let* $\Phi_i=\Phi(\Omega_i),\Psi_{ij}=\Psi(\Omega_i,\Omega_j)$ *be such that* $E(\Phi_1)=0$. *Then if* $E\left\{\left(\Phi_1\Psi_{12}\right)^2\right\}<\infty$,

$$E\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\{\Phi_i\Psi_{ij}-E\left(\Phi_i\Psi_{ij}|\Omega_j\right)\right\}\right]^2\leq\frac{n(n-1)}{4}E\left\{\left(\Xi_{12}+\Xi_{21}\right)^2\right\}, \tag{12}$$

with $\Xi_{12}=\Phi_1\Psi_{12}-E\left(\Phi_1\Psi_{12}|\Omega_2\right)-E\left(\Phi_1\Psi_{12}|\Omega_1\right)+E\left(\Phi_1\Psi_{12}\right)$.

Proof: Observe that the double sum in (12) is an asymmetric U–statistic (See Hoeffding, 1948, and Serfling, 1980). The stated result follows immediately from projecting onto the basic observations (Serfling, 1980).□

Lemma 6 *Let*

$$\psi_n(s,t) = [m\{s,g(t)\} - m\{x,g(z)\}]\,[\{\tilde{g}(z) - g(z)\} - \{\tilde{g}(t) - g(t)\}]\,.$$

Then

$$n^{-1}\sum_{i=1}^n h_m^{-d_x-2} k_{h_m}(x - X_i) k'_{h_m}\{g(z) - g(Z_i)\}\tfrac{Y_i - m\{x,g(z)\}}{f_{XG}\{x,g(z)\}}\,[\{\tilde{g}_{-i}(z) - g(z)\} - \{\tilde{g}_{-i}(Z_i) - g(Z_i)\}] \quad (13)$$
$$= E^*\left\{\frac{\frac{\partial\psi_n}{\partial t_1}(X_1,Z_1)}{\frac{\partial g}{\partial z_1}(Z_1)}|X_1 = x, g(Z_1) = g(z)\right\} + O_p\left\{h_m^r + n^{-1/2}h_m^{-(d_x+3)/2}\left(n^{-1/2}h_g^{-d_z/2} + h_g^r\right)\right\},$$

where $E^*\{\omega(X_1,Z_1)\} = \int \omega(x,z) f_{XZ}(x,z)dxdz$ *even if* ω *itself is a random function.*

Proof: Define $\tilde{\psi}_n(x,g) = E\{\psi_n(X_1,Z_1)|X_1 = x, g(Z_1) = g\}$. Denote the summand in expression (13) by $\varsigma_n(X_i,Z_i,Y_i)$. From Lemma 5 it follows that $n^{-1}\sum_{i=1}^n \varsigma_n(X_i,Z_i,Y_i) = E^*\{\varsigma_n(X_1,Z_1,Y_1)\} + O_p\left\{n^{-1/2}h_m^{-(d_x+3)/2}\left(n^{-1/2}h_g^{-d_z/2} + h_g^r\right)\right\}$. But using a standard kernel bias expansion one obtains that

$$E^*\{\varsigma_n(X_1,Z_1,Y_1)\} = \frac{\partial\tilde{\psi}_n}{\partial g}\{x,g(z)\} + O_p(h_m^r). \quad (14)$$

Now, since $\psi_n(t,x) = 0$ for all t for which $g(t) = g(z)$,

$$\frac{\partial\tilde{\psi}_n}{\partial g}\{x,g(z)\} = E^*\left\{\frac{\frac{\partial\psi_n}{\partial t_1}(X_1,Z_1)}{\frac{\partial g}{\partial z_1}(Z_1)}|X_1 = x, g(Z_1) = g(z)\right\}$$

□

Lemma 7 $n^{-1}h_g^{-d_z}\sum_{j=1}^n k_{h_g}(z - Z_j)\left\{J_{nj}(z) - \frac{Y_j - a(X_j,z)}{f_{XZ}(X_j,z)}\lambda(X_j)\right\} = O_p(h_g^r)$.

Proof: Follows from a Taylor series expansion with residual on the first term in curly brackets about the second term. □

Lemma 8 *When* $nh_g^{d_z+2r} \to 0$ *and* $nh_g^{d_z} \to \infty$ *as* $n \to \infty$,

$$\sqrt{nh_g^{d_z}}E^*\left\{\frac{\frac{\partial\psi_n}{\partial t_1}(X_1,Z_1)}{\frac{\partial g}{\partial z_1}(Z_1)}|X_1 = x, g(Z_1) = g(z)\right\} \xrightarrow{\mathcal{L}} N\left[0, \left[\frac{\partial m}{\partial g}\{x,g(z)\}\right]^2 \int k^2(s)ds\sigma_J^2(z)\right].$$

Proof: Since
$E^*\left\{\frac{\frac{\partial\psi_n}{\partial t_1}(X_1,Z_1)}{\frac{\partial g}{\partial z_1}(Z_1)}|X_1 = x, g(Z_1) = g(z)\right\} = \frac{\partial m}{\partial g}\{x,g(z)\}\,[\tilde{g}(z) - E^*\{\tilde{g}(Z_1)|X_1 = x, g(Z_1) = g(z)\}]$, need to consider $\sqrt{nh_g^{d_z}}\,[\tilde{g}(z) - E^*\{\tilde{g}(Z_1)|X_1 = x, g(Z_1) = g(z)\}]$. Its bias is $O(h_g^r)$, again by a standard kernel bias expansion. But by Lemma 7 and the bandwidth condition in the lemma statement, $\tilde{g}(z)$

is asymptotically equivalent with $\tilde{g}^*(z) = n^{-1}h_g^{-d_z}\sum_{j=1}^n k_{h_g}(z-Z_j)\frac{Y_j-a(X_j,z)}{f_{XZ}(X_j,z)}\lambda(X_j)$. Since the variance of $E^*\{\tilde{g}(Z_1)|g(Z_1)=g(z)\}$ is of lower order than the variance of $\tilde{g}^*(z)$, the asymptotic variance of the left hand side in the lemma statement is $\left[\frac{\partial m}{\partial g}\{x,g(z)\}\right]^2$ times the variance of $\sqrt{nh_g^{d_z}}\tilde{g}^*(z)$.□

Lemma 9 *Let $\{W_i\}$ be a sequence of i.i.d. random variables with $E|W_1|^{p_W} < \infty$ for some $p_W > 2$. Then*

$$\sup_{b\in\mathcal{B}}\left|n^{-1}h_m^{-d_x-1}\sum_{i=1}^n k_{h_m}(x-X_i)\left[k_{h_m}\{\hat{g}(z)-\hat{g}(Z_i)\}-k_{h_m}\{g(z)-g(Z_i)\}\right]W_i\right| = o_p\left(h_m^{-1}\Psi_{gn}+h_m^{-3}\Psi_{gn}^2\right),$$

where $\Psi_{gn} = n^{-1/2}h_g^{-d_z/2}\log n + n^{-1}h_g^{-d_b}\log n + h_g^r$, the uniform convergence rate of $\hat{g}$.

Proof: Let $\Delta_{gni}(z) = \{\hat{g}(z)-g(z)\}-\{\hat{g}(Z_i)-g(Z_i)\}$. Bound the left hand side by

$$\sup_{b\in\mathcal{B}} n^{-1}h_m^{-d_x-2}\sum_{i=1}^n\left|k_{h_m}(x-X_i)\,k'_{h_m}\{g(z)-g(Z_i)\}W_i\right|\max_{i\le n}\sup_{z\in\mathcal{B}_Z}|\Delta_{gni}(z)| \tag{15}$$

$$+\sup_{b\in\mathcal{B}} n^{-1}h_m^{-d_x-3}\sum_{i=1}^n\left|k_{h_m}(x-X_i)\,k''_{h_m}\{\cdot\}W_i\right|\max_{i\le n}\sup_{z\in\mathcal{B}_Z}|\Delta_{gni}(z)|^2. \tag{16}$$

Note first that $\max_{i\le n}\sup_{z\in\mathcal{B}_Z}|\Delta_{gni}(z)| = o_p(\Psi_{gn})$. I first deal with (16). Note that k'' is bounded by assumption. But $\sup_{b\in\mathcal{B}} n^{-1}h_m^{-d_x-3}\sum_{i=1}^n|k_{h_m}(x-X_i)W_i| = h_m^{-d_x-3}\sup_{b\in\mathcal{B}}E|k_{h_m}(x-X_1)W_1|+o_p\left(n^{-1/2}h_m^{-3-d_x/2}\right)$ by Lemma 2. But $n^{-1/2}h_m^{-3-d_x/2} = O\left(h_m^{-3}\right)$ by Assumption F and $h_m^{-d_x-3}\sup_{b\in\mathcal{B}}E|k_{h_m}(x-X_1)W_1| = O\left(h_m^{-3}\right)$ by substitution in the integral. Now (16). Note that $\sup_{b\in\mathcal{B}} n^{-1}h_m^{-d_x-2}\sum_{i=1}^n\left|k_{h_m}(x-X_i)\,k'_{h_m}\{g(z)-g(Z_i)\}W_i\right| =$ $h_m^{-d_x-2}\sup_{b\in\mathcal{B}}E\left|k_{h_m}(x-X_1)\,k'_{h_m}\{g(z)-g(Z_1)\}W_1\right| + o_p\left(n^{-1/2}h_m^{-d_x-3/2}\right)$ by Lemma 2. But $n^{-1/2}h_m^{-d_x-3/2} = O\left(h_m^{-1}\right)$ and $h_m^{-d_x-2}\sup_{b\in\mathcal{B}}E\left|k_{h_m}(x-X_1)\,k'_{h_m}\{g(z)-g(Z_1)\}W_1\right| = O\left(h_m^{-1}\right)$. □

Proof of Theorem 5.

Note that analogous to (9),

$$\hat{a}_s - a = \frac{(\hat{N}_{a_s}-N_{a_s})D_{a_s}-N_{a_s}(\hat{D}_{a_s}-D_{a_s})}{D_{a_s}^2}\sum_{i=0}^{\infty}\left(\frac{D_{a_s}-\hat{D}_{a_s}}{D_{a_s}}\right)^i.$$

Since $\hat{D}_{a_S}$ converges uniformly to D_{a_s} in a neighborhood of $\{x,g(z)\}$ by Lemma 9 and Assumption F, all terms except $i=0$ in the expansion can be ignored. Now take a second order Taylor expansion

on the remainder to obtain $T_{0n} + T_{1n} + \sum_{t=2}^{\Phi-1} T_{2nt} + T_{3n}$, with

$$
\begin{aligned}
T_{0n} &= n^{-1}h_m^{-d_x-1}\sum_{i=1}^n k_{h_m}(x-X_i)k_{h_m}\left\{g(z)-g(Z_i)\right\}\tfrac{Y_i-a(x,z)}{f_{XG}\{x,g(z)\}},\\
T_{1n} &= n^{-1}h_m^{-d_x-2}\sum_{i=1}^n k_{h_m}(x-X_i)k'_{h_m}\left\{g(z)-g(Z_i)\right\}\tfrac{Y_i-a(x,z)}{f_{XG}\{x,g(z)\}}\left\{\hat{g}(z)-g(z)-\hat{g}_{-i}(Z_i)+g(Z_i)\right\},\\
T_{2nt} &= n^{-1}h_m^{-d_x-1-t}\sum_{i=1}^n k_{h_m}(x-X_i)k^{(t)}_{h_m}\left\{g(z)-g(Z_i)\right\}\tfrac{Y_i-a(x,z)}{f_{XG}\{x,g(z)\}}\left\{\hat{g}(z)-g(z)-\hat{g}_{-i}(Z_i)+g(Z_i)\right\}^t,\\
T_{3n} &= n^{-1}h_m^{-d_x-1-\Phi}\sum_{i=1}^n k_{h_m}(x-X_i)k^{(\Phi)}_{h_m}\{\cdot\}\tfrac{Y_i-a(x,z)}{f_{XG}\{x,g(z)\}}\left\{\hat{g}(z)-g(z)-\hat{g}_{-i}(Z_i)+g(Z_i)\right\}^\Phi.
\end{aligned}
$$

Standard kernel regression estimation theory implies that T_{0n} has the properties ascribed to $\hat{a}_S$ in the statement of Theorem 5. I hence need to show that T_{1n}, T_{2nt}, T_{3n} for $t = 2, \ldots, \Phi-1$ are of lower order than T_{0n} under the conditions of (ii) and of lower or equal order under (i). The result for T_{2nt} is proved in Lemma 3 using the bandwidth conditions of Assumption F and T_{3n} is dealt with in Lemma 4 again using Assumption F (for T_{3n}, note that it suffices to show that $h_m^{-\Phi-1}\Psi_n^\Phi$ goes to zero faster than Ψ_n where Ψ_n is the convergence rate of $\hat{g}$; the bandwidth conditions guarantee that $h_m^{-(\Phi+1)/(\Phi-1)}\Psi_n \to 0$ as $n \to \infty$).

Now T_{1n}. Apply Lemmas 6, 7 and 8 to obtain that $T_{1n} = O_p(h_m^r + h_g^r + n^{-1/2}h_g^{-d_z/2})$. The result now follows immediately with Assumption F. □

Proof of Theorem 6

The steps are virtually identical to those of the proof of Theorem 5, albeit that T_{0n} and T_{2n} are now dominated by T_{1n}, which by Lemmas 6, 7 and 8 has the properties described in the theorem statement. □

Proof of Theorem 7

Again, the proof is almost identical to those of Theorems 5 and 6. Part (i) follows trivially from the proofs of Theorems 5 and 6. Now (ii). By Assumption F and Lemmas 6, 7 and 8 both T_{0n} and T_{1n} are asymptotically identical to two partial sums having the limiting normal distributions derived in Theorems 5 and 6. Their sum has hence again a normal distribution with mean zero and variance the sum of the variances plus twice the covariance. But

$$
\begin{aligned}
&n^{-2}h_g^{-d_z}h_m^{-d_x-1}\sum_{i=1}^n\sum_{j=1}^n\\
&\mathrm{Cov}\left\{k_{h_g}(z-Z_j)\tfrac{Y_j-a(X_j,z)}{f_{XZ}(X_j,z)}\lambda(X_j), k_{h_m}(x-X_i)k_{h_m}\left\{g(z)-g(Z_i)\right\}\tfrac{Y_i-a(x,z)}{f_{XG}\{x,g(z)\}}\right\}
\end{aligned}
$$

is $O(n^{-1}h_m^{-1})$ and is hence of lower order than the variance terms. □

Proof of Theorem 8.

Follows from Lemma 9.□

Proof of Theorem 9

Follows immediately from Lagrangean optimization. □

Figure 1: Change in Accuracy â and $\hat{a}_S$ with Sample Size
Marginal Integration, Model 1; dx=1, dz=2

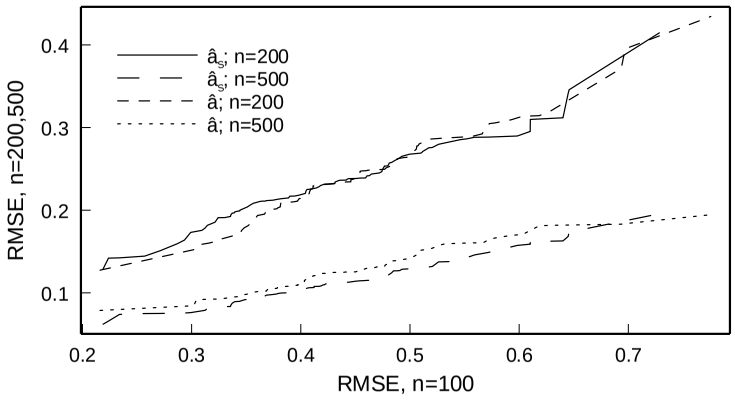

Figure 2: Change in Accuracy â and $\hat{a}_S$ with Sample Size
Marginal Integration, Model 1; dx=dz=2

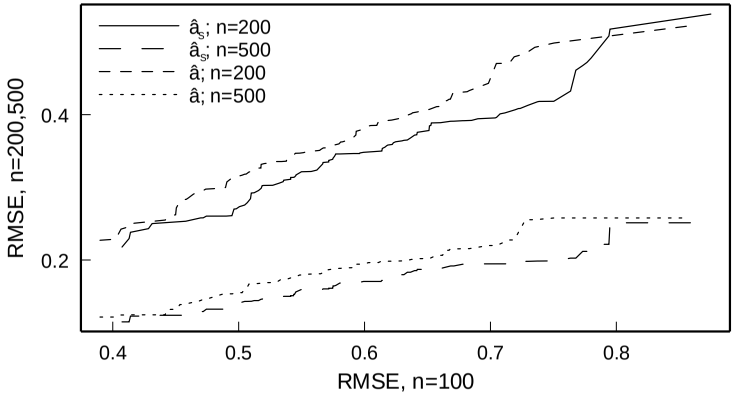

Figure 3: Comparison of $\hat{a}_S$
using Marginal Integration and Marginal Summation
Model 1, n=100, dx=1, dz=2

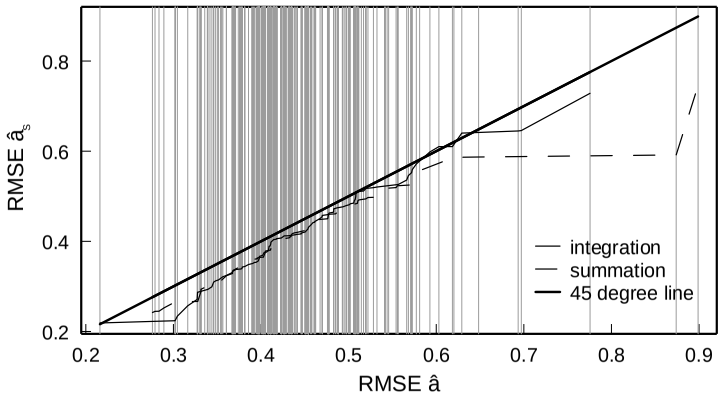

Figure 4: Comparison of $\hat{a}_S$
using Marginal Integration and Marginal Summation
Model 1, n=200, dx=1, dz=2

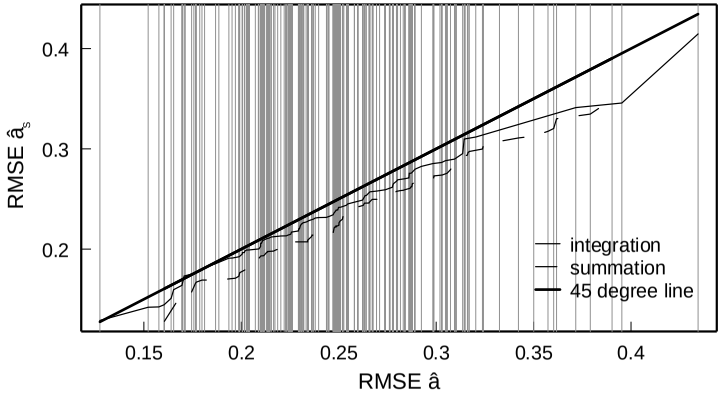

Figure 5: Comparison of $\hat{a}_S$
using Marginal Integration and Marginal Summation
Model 1, n=500, dx=1, dz=2

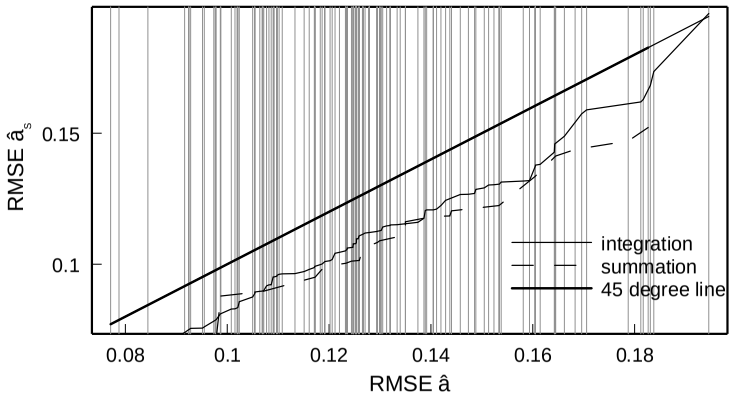

Figure 6: Change in Accuracy â and $\hat{a}_S$ with Sample Size
Marginal Summation, Model 1; dx=dz=2

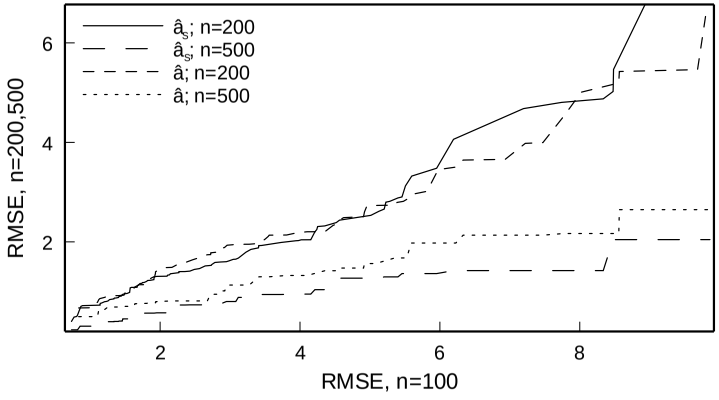

Figure 7: Comparison of â and $\hat{a}_S$
Marginal Summation, n=100, dx=dz=2,
Models 1,4 and 5

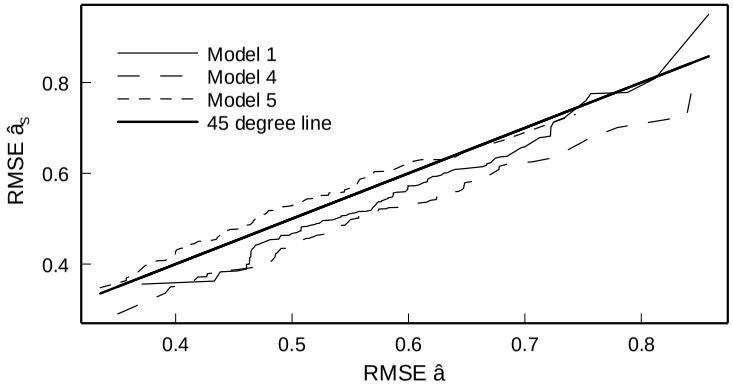

Figure 8: Comparison of â and $\hat{a}_S$
Marginal Summation, n=100, dx=dz=2,
Models 2 and 3

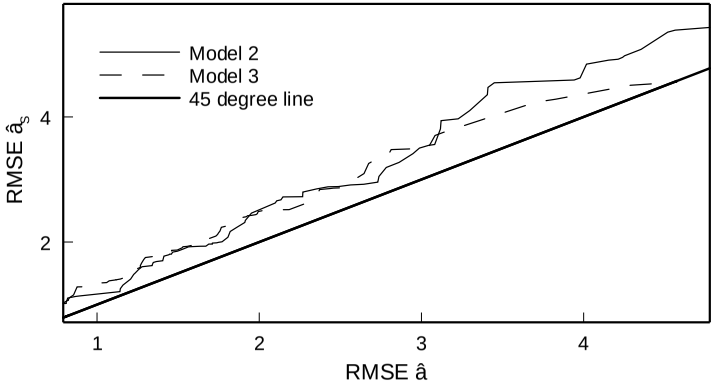

Figure 9: Comparison of $\hat{a}$ and $\hat{a}_S$
Marginal Summation, n=100, dx=dz=2,
Chosen λ; Models 2 and 3.

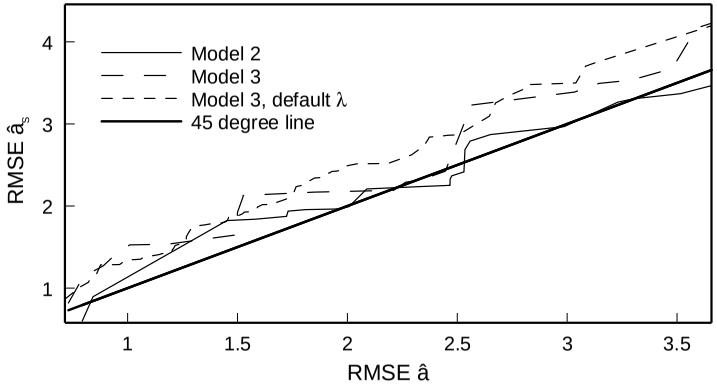

Figure 10: Comparison of $\hat{a}$ and $\hat{a}_S$ across Dimensions
Marginal Summation, Model 1, n=100

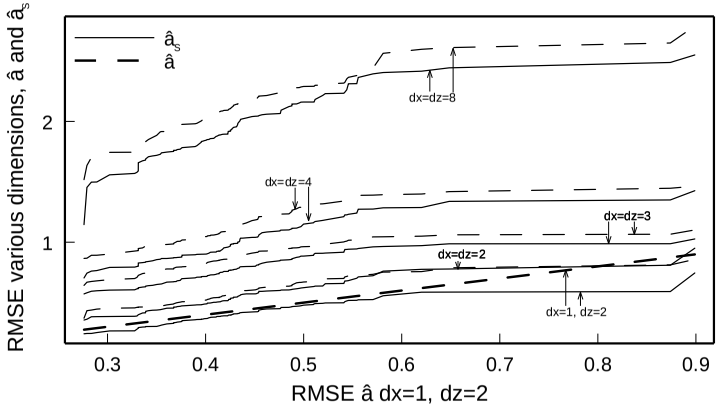

Figure 11: Comparison of $\hat{a}$ and $\hat{a}_S$ across Dimensions
Marginal Summation, Model 1, n=200