Kernel estimation in regression on vector and function spaces

Sid Kankanala Marcia Schafgans
Yale University London School of Economics

Victoria Zinde-Walsh*
McGill University and CIREQ

October 2, 2023

Abstract

We investigate the application of kernel-weighted nonparametric regression to scenarios where the distribution of the regressors may not possess absolute continuity with respect to the Lebesgue measure. Our theoretical framework encompasses a wide range of distributions, including those characterized by low-dimensional measures, distributions featuring mass points, and Hausdorff measures on self-similar fractals. We establish a novel central limit theorem and demonstrate that convergence rates are influenced non-trivially by the underlying distribution. In the case of absolutely continuous measures, our approach weakens the usual regularity conditions. Furthermore, we extend our analysis to encompass kernel regression with multiple functional regressors.

1 Introduction

This paper is centered on the nonparametric regression model

$$Y = m(X) + u, \quad E(u|X) = 0, \tag{1}$$

where the regressor X can be a vector in R^q ($X = (X^1, ..., X^q)$), or alternatively X may belong to a function space, such as a Banach space of continuous functions on a compact set, or a more general metric space. In fact, X does not necessarily have to be a vector or a function; it could be an interval, a set, a graph, or a network, as long as a metric (or even a semi-metric) can be defined

*Support from the Natural Sciences and Engineering Research Council of Canada (NSERC) grant 253139 is gratefully acknowledged.
The authors want to thank participants at the Bristol Econometric Study Group meetings.

for it. Y represents a scalar dependent variable, u denotes an unobserved error, and the conditional mean function $m(X)$ satisfies some smoothness assumptions.

The random object X is supported on some domain in a vector or metric, semi-metric space denoted as Ξ. Depending on the nature of Ξ, we can identify three distinct situations:

(*i*) If $\Xi = R^q$, we encounter the standard widely studied nonparametric regression model. Nadaraya (1964) and Watson (1964) introduced the kernel estimator for this case, which was further extended by Stone (1977) to local linear and local polynomial estimators.

(*ii*) For a general univariate metric space $\Xi = \Xi^{[1]}$, the regression problem is known as functional regression. Here, $m(X)$ can represent, for example, the action of a functional on Banach space of continuous functions. Estimation and inference techniques for functional regression have been developed and discussed in a growing body of literature (see Ferraty et al., 2006); local linear as well as local constant estimator have been used (e.g., Ferraty and Nagy, 2022).

It is important to note that while R^q can be treated as a metric space, regression analysis for R^q cannot necessarily be reduced to a special case of functional regression. This distinction arises because the vector X may have non-commensurate components, such as monetary and demographic variables which require a multivariate kernel estimator, e.g., with a product kernel. Hence, case (i) cannot be reduced to case (ii).

(*iii*) $X = (X^1, ..., X^q)$ may have q non-commensurable components, with each X^l defined on a Banach or metric space Ξ_l. In such cases, X could be considered in a product space $\Xi = \Xi^{[q]} \equiv \Xi_1^{[1]} \times \cdots \times \Xi_q^{[1]}$ with a corresponding product measure.

On the Euclidean space R^q, the conventional local weights are defined via a vector $W(x) = h^{-1}(x - X)$ with components $W^j(x) = \left(h^j\right)^{-1}(x^j - X^j)$. The functional regression version of kernel estimator for x in an infinite-dimensional Banach space $\Xi = \Xi^{[1]}$ with norm $\|.\|$ utilizes local weights based on $W(x) = h^{-1}\|x - X\|$. In the case of a metric space for the objects of interest, the Banach norm can be replaced by a **suitable distance measure;** without loss of generality we keep the $\|.\|$ notation to represent the distance measure. For example, Severn et al. (2021) employed the NW estimator for nonparametric regression of networks, where Ξ represents the space of Laplacian matrices.

On $\Xi = \Xi^{[q]}$ consider $W(x) = h^{-1}\|x - X\|$ with $W^j(x) = \left(h^j\right)^{-1}\left\|x^j - X^j\right\|_j$, where $\|.\|_j$ is the metric in $\Xi^{[j]}$. The multivariate kernel function is $K(W(x)) = K(W^1(x), ..., W^q(x))$. It is common in R^q to utilize the product kernel:

$$K(W(x)) = \prod_{j=1}^{q} k\left(W^j(x)\right) \tag{2}$$

with the kernel functions $k(.)$ and bandwidth vector h chosen in a way that provides desirable properties for the kernel estimator. In a metric space $\Xi = \Xi^{[1]}$ kernel functions $k(.)$ are defined for a univariate nonnegative argument. Here we also simplify by using a product kernel to extend the functional regression to the case $\Xi = \Xi^{[q]}$; while different kernel functions could apply to the different components $W^j(x)$, we do not make this distinction.

The probability measure in univariate metric space $\Xi = \Xi^{[1]}$ is characterized by the small ball probability: for ball $B(x,h) = \{X : \|x - X\| \leq h\}$ in Ξ the probability measure $P_X(B(x,h))$ is defined.

Kankanala and Zinde-Walsh (2023) introduced small cube probability for a cube[1] $C(x,h) = \left\{X \in R^q : \left|X^l - x^l\right| \leq h^l, l = 1, \cdots, q\right\}$ in R^q. With distribution function of X given byF_X the corresponding probability measure is given by

$$F_X(x - h, x + h) = P_X(C(x,h)). \tag{3}$$

We use small cube probability to extend the regression to a product of general metric spaces. Suppose $X \in \Xi = \Xi^{[q]} \equiv \Xi_1^{[1]} \times \cdots \times \Xi_q^{[1]}$, where the metrics on $\Xi_l^{[1]}$, $\left\|X^l - x^l\right\|_l$, could differ for each of the q components of function spaces. We consider a small cube around $x = \left(x^1, ..., x^q\right)$ where for each component the small ball $B^l\left(x^l, h^l\right) = \left\{X^l : \left\|x^l - X^l\right\|_l \leq h^l\right\}$ is defined for a vector $h = \left(h^1, ..., h^q\right)$, and the cube is

$$C(x,h) = \left\{X : \left\|X^l - x^l\right\|_l \leq h^l\right\} = \left\{X : X^l \in B^l\left(x^l, h^l\right)\right\}. \tag{4}$$

The corresponding small cube probability is denoted $P_X(C(x,h))$.

The existing literature as well as the results presented in this paper regarding the pointwise limit properties of the estimator can be categorized based on assumptions made about the space Ξ, the small cube probability $P_X(C(x,h))$ and the kernel function $K(\cdot)$. A kernel is said to be of type I if it is nonzero everywhere on its support (e.g. Uniform), while a kernel is type II if it can be zero on the boundary of its support (e.g. Epanechnikov). Below, A-E outlines the results in the literature for the kernel estimators (Nadaraya-Watson and local linear) and the contributions of this paper.

A For $\Xi = R^q$, assuming a sufficiently smooth distribution F_X and considering various kernels, well-established results found in standard textbooks demonstrate the consistency and asymptotic normality of the estimator using appropriate bandwidth sequences (e.g., Racine and Li, 2007).

B This paper establishes the asymptotic normality of the kernel estimator for $\Xi = R^q$ with a more general distribution F_X and various kernels, including Type I and Type II product kernels. For a Type I kernel, a sufficient restriction for asymptotic normality is that $nP_X(C(x,h) \to \infty$.

[1]We use the shorthand "cube" for the set $\left\{X \in R^q : \left|X^l - x^l\right| \leq h^l, l = 1, \cdots, q\right\}$ even though the sides given by $\left\{h^i\right\}$ may not be equal.

However, for commonly used Type II kernels, an additional condition is required :

$$\frac{P_X\left(C\left(x,h\right)\right)}{P_X\left(C\left(x,\varepsilon h\right)\right)} < C_F < \infty \tag{5}$$

for some $\varepsilon \in (0,1)$ (see Assumption 4(b) below). This condition is less restrictive than a similar analog proposed in Ferraty and Vieu (2006, Lemma 4.4). We verify that these conditions are sufficient to establish the limit theory. Furthermore, the rate of convergence is shown to depend non-trivially on the underlying distribution of the regressors. Kankanala and Zinde-Walsh (2023) study the closely related problem of a kernel-weighted specification test for the regression function under a general distribution of the regressors.

C The pointwise asymptotic normality for a Banach or metric space Ξ is established by Ferraty et al. (2007) and Geenens (2015) for kernels of Type I. Under a condition similar to (5), the result holds for kernels of Type II as well. Ferraty and Nagy (2022) provide results for a LL estimator in a Hilbert space Ξ.

D For the case where $\Xi^{[q]}$ represents a finite product of normed or metric spaces, and a corresponding product kernel is used, this paper derives the limit Gaussian distribution.

E We establish moment bounds for local functions that depend on the small cube probability. These bounds generalize the results in the literature (e.g. in Ferraty and Vieu, 2006) and subsequent papers on functional regression to the product space. This result applies to a broad class of local functions, encompassing general kernels as well as product kernels, their powers, and even their partial derivatives.

F We provide simulation evidence on the behavior of kernel estimators under possible singularity of conditioning distribution in R^q. The simulations demonstrate the effect of singularity on point-wise convergence rates for the kernel estimators.

The structure of the paper is as follows. In Section 2, we introduce the general framework, outline the assumptions, and discuss the classes of conditioning distributions in both finite-dimensional and metric spaces where our results hold. Section 3 is dedicated to deriving the limit process for the Nadaraya-Watson (NW) estimator and the local linear estimator in the product space under more general assumptions than considered in the literature. We present the results of our simulations in Section 4 and conclude in Section 5. In Appendix A we provide the general technical result for moments and their bounds on which our proofs rely. Appendices B and C contain the proofs. Supplementary appendices provide ...

2 The set-up

This section provides the (well-known) formulae for the kernel estimators, introduces some useful notation and provides formal assumptions. The assumptions on the model and sampling are the easiest to work with, since the focus of the paper is on the generality of assumptions on the underlying distribution of the conditioning variables (or functions, or other symbolic objects). Those distributional assumptions are very general; the limit theory (in the following section) requires a combination of the degree of generality of the distributional assumptions and the properties of the kernel.

2.1 The kernel

To define the kernel estimators we first formally define some notation and restrict the kernel functions that will be considered. We consider a univariate metric space $\Xi = \Xi^{[1]}$ or a finite set of such spaces, $\left\{\Xi_i^{[1]}\right\}_{i=1}^{q}$ and a product space $\Xi = \Xi^{[q]}$. In a product space we define a vector w as $\left(w^1, ..., w^q\right)^T$ where each component is in the corresponding space, thus for R^q, w is a q-dimensional vector of reals, in $\Xi = \Xi^{[q]}$ each $w^l \in \Xi_l^{[1]}$. The bandwidth vector is $h = \left(h^1, ...h^q\right) \in R^q$ where if all components are the same we write $h\iota$ for the bandwidth vector, where ι denotes a vector of ones. We use the same notation $\|.\|$ for the absolute value (or modulus) of a scalar in R^1, the Euclidean norm for a vector in R^q or norm for a function in $\Xi = \Xi^{[1]}$, with Ξ a Banach space, or semi-norm or metric in a metric space Ξ; when the meaning is not clear from the context we shall specify.

The following assumption is made on the kernel function.

Assumption 1 *(kernel) (a) The kernel function* $K(w) = \prod_{j=1}^{q} k\left(w^j\right)$ *for* $w = \left(w^1, ..., w^q\right)$ *is a product kernel on* $\Xi^{[q]}$;

(b) each $k(z)$ *is non-negative, non-increasing for* $z > 0$;

(c) $k(.)$ *is either symmetric with support on* $[-1, 1]$ *or* $k(.) = k_+(.)$ *is supported on* $[0, 1]$*; it is continuously differentiable in the interior of its support.*

(d) $k(z)$ *is a type I kernel:* $k(z) > 0$ *on its support.*

Assumption 1(a) is made to simplify the exposition, but our technical results in Appendix A provide moment derivations that do not require a product structure for the local functions, such as kernels and their powers. Assumption 1(b,c) are satisfied by the Epanechnikov and quartic kernels, as well as a uniform kernel; the latter additionally satisfies Assumption 1(d). Kernels of type I provide advantages for estimation and inference, but the literature on nonparametric estimators in standard settings with finite-dimensional $X \in R^q$ typically uses kernels such as Epanechnikov that is a type II kernel. We thus investigate the impact of both type I and type II kernel on the properties of the estimator.

2.2 The local constant (NW) estimator

The local constant kernel estimator, solves (with $W(x) = h^{-1}(x - X)$ on R^q or with $W(x) = h^{-1}\|x - X\|$ on a metric space $\Xi^{[q]}$)

$$\arg\min_{a \in R} \sum_{i=1}^{n} (Y_i - a) K(W_i(x))$$

to provide the Nadaraya-Watson (NW) estimator for the regression function $m(X)$ defined in (1)

$$\widehat{m}(x) = B_n^{-1}(x) A_n(x), \tag{6}$$

$$B_n(x) = \frac{1}{n}\sum_{i=1}^{n} K(W_i(x)); \; A_n(x) = \frac{1}{n}\sum_{i=1}^{n} K(W_i(x)) Y_i. \tag{7}$$

The NW estimator has the advantage that its limit properties can be established under very general assumptions on the model. In particular, for functional regression this estimator can be applied in any metric (even semi-metric) space, or product of such spaces.

2.3 The Local linear (LL) estimator

The local linear kernel approach to estimation places more restrictions on the regression function and on the domain of definition of the conditioning variables. In particular, in the conditioning space an inner product needs to be defined and for the regression function a derivative (gradient) needs to exist. In R^q the scalar product exists and for a differentiable $m(x)$ the local linear estimator is well defined and its properties were examined in the literature starting with Stone (1977). Here we extend the asymptotic normality for the LL estimator results for R^q by removing the usual smoothness assumptions on the distribution.

The theory and practice of the functional LL estimator is less well developed. When $\Xi^{[1]}$ is a Hilbert space and $m(x)$ is suitably differentiable, Ferraty and Nagy (2022) established asymptotic normality for the LL estimator. Here we point out that this result can be extended to $\Xi^{[q]}$ analogously to the extension proposed here for the NW estimator. While providing the general framework of the LL estimator for a function space $\Xi^{[q]}$, establishing the asymptotic normality of functional LL estimator lies beyond the scope of this paper.

To set up the local linear estimator consider a function space $\Xi^{[q]}$. Assume that each $\Xi_j^{[1]}$ is a space with a scalar product, e.g. a Hilbert space, then for any $v^j, w^j \in \Xi_j^{[1]}$ denote the inner product by $\langle v^j, w^j \rangle_j$. To extend this to a product space consider in $\Xi = \Xi^{[q]}$ the vectors $v = \left(v^1, ..., v^q\right)$, $w = \left(w^1, ..., w^q\right)$ with $v^j, w^j \in \Xi_j^{[1]}$. Then define the inner product for $\Xi^{[q]}$ as

$$\langle v, w \rangle = \sum_{j=1}^{q} \langle v^j, w^j \rangle_j,$$

also define the $(q+1)\times 1$ vector $\mathcal{X} = \left(1, x^1 - X^1, ...x^q - X^q\right)^T \in R\times\Xi^{[q]}$. Define the kernel weights as $K\left(W(x)\right) = \prod k_j\left(W^j\left(x\right)\right)$ with $W^j\left(x\right) = \left(h^j\right)^{-1}\left(x^j - X^j\right)$ for $\Xi = R^q$ and $W^j\left(x\right) = \left(h^j\right)^{-1}\left\|x^j - X^j\right\|_j$ in the product function space $\Xi = \Xi^{[q]}$. The local linear estimator $\hat{\beta}\left(x\right) = \left(\hat{\beta}^0\left(x\right), \hat{\beta}^1\left(x\right), ..., \hat{\beta}^q\left(x\right)\right)$ solves the generalized least squares problem:

$$\hat{\beta}\left(x\right) = \arg\min_{\beta\in R\times\Xi}\sum_{i=1}^{n}\left[\left(Y_i - \beta^0\left(x\right) - \sum_{j=1}^{q}\left\langle\mathcal{X}_i^j, \beta^j\left(x\right)\right\rangle_j\right)^2 K\left(W_i(x)\right)\right]. \quad (8)$$

The solution for $\beta^0\left(x\right)$ represents the local linear (LL) estimator of the conditional mean. When $\Xi = R^q$, $\beta^j\left(x\right)$ $(j = 1, ..., q)$ are scalars representing the values of the partial derivative of the conditional mean function at x, that is $(\beta^1(x), \cdots, \beta^q(x))^T = \nabla m(x)$. Here ∇g denotes the Jacobian vector for a function $g\left(x\right)$ of $x = \left(x^1, ..., x^q\right)^T \in R^q$ (where $(\cdot)^T$ denotes the transpose): $\nabla g\left(x\right) = \left(\frac{\partial}{\partial x^1}, ...\frac{\partial}{\partial x^q}\right)^T g\left(x\right)$. When $\Xi_j^{[1]}$ is a Hilbert space of functions the corresponding $\beta^j\left(x\right)$ is a functional derivative at x which is an element of that space, which we also denote by $m_j'(x)$. We next focus on $\Xi^{[q]} = R^p$.

For R^q the inner product $\sum_{j=1}^{q}\left\langle\mathcal{X}_i^j, \beta^j\left(x\right)\right\rangle_j$ in (8) is just the scalar product $\sum_{j=1}^{q}\mathcal{X}_i^j\beta^j\left(x\right)$. Then $\hat{\beta}(x) = (\mathcal{X}^T\mathcal{K}(x)\mathcal{X})^{-1}\mathcal{X}^T\mathcal{K}(x)Y$, where $\mathcal{K}(x)$ is an n dimensional diagonal matrix with $K\left(W_i\right)$ as its ith diagonal element.

Let

$$D = \begin{pmatrix} 1 & 0 \\ 0 & D_h \end{pmatrix}$$

where D_h is a $q\times q$ dimensional diagonal matrix: $D_h = diag\left\{h_1, \cdots, h_q\right\}$. Note that $D^{-1}\mathcal{X} = \mathcal{W}$ with the ith row given by $\mathcal{W}_{i,\cdot} = \left(1, W_i^T\right) = \left(1, W_i^1, \cdots, W_i^q\right)$. We can then express the LL estimator, $\hat{\beta}\left(x\right)$, as a solution to

$$\widetilde{B}_n\left(x\right)\hat{\beta}\left(x\right) - \widetilde{A}_n\left(x\right) = 0$$

with

$$\begin{aligned}\widetilde{B}_n\left(x\right) &= \left[\frac{1}{n}\sum_{i=1}^{n}K\left(W_i\right)_{i,\cdot}\mathcal{X}_{i,\cdot}\mathcal{X}_{i,\cdot}^T\right] \\ &= D\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}K\left(W_i\right) & \frac{1}{n}\sum_{i=1}^{n}\left(h\circ W_i\right)^T K\left(W_i\right) \\ \frac{1}{n}\sum_{i=1}^{n}\left(h\circ W_i\right)K\left(W_i\right) & \frac{1}{n}\sum_{i=1}^{n}W_iW_i^TK\left(W_i\right)\end{pmatrix}D; \quad (9)\end{aligned}$$

$$\widetilde{A}_n\left(x\right) = \left[\frac{1}{n}\sum_{i=1}^{n}K\left(W_i\right)\mathcal{X}_{i,\cdot}Y_i\right] = D\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}Y_iK\left(W_i\right) \\ \frac{1}{n}\sum_{i=1}^{n}W_iY_iK\left(W_i\right)\end{pmatrix}. \quad (10)$$

Here $a \circ b$ denotes the Hadamard product of two matrices or vectors, in particular, for $q-$dimensional vectors, $h, W \in R^q$ the vector $h \circ W$ has components $h^j W^j$, $j = 1, .., q$.

Identification of the components of $\hat{\beta}(x)$ requires invertibility of $\widetilde{B}_n(x)$. Note that with singularity it possible that several components of W_i are zero in some neighbourhood of x with probability 1; then $\widetilde{B}_n(x)$ will have reduced rank. Assuming invertibility we can express the estimator as

$$\hat{\beta}(x) = \widetilde{B}_n^{-1}(x) \widetilde{A}_n(x) \tag{11}$$

2.4 Assumptions

We operate under i.i.d. sampling although the results could easily extend to more general sampling schemes and assume a continuously differentiable regression function even though we could derive results under Holder differentiability or differentiability of the expectation of the regression function. The simplified assumptions serve to highlight the impact of the underlying distribution of the conditioning X, be it a random vector of a more complicated object, such a vector of functions. We use the generic notation L for a scalar lower bound and M for an upper bound, indicating with a subscript the function that is being bounded.

Assumption 2 *(a) The sample (Y_i, X_i) for $i = 1, ..., n$ is i.i.d. with $Y_i \in R; X_i \in \Xi^{[q]}$. (b) For $\mu_2(x) = E\left(u^2 | X = x\right)$ and $\mu_4(x) = E\left(u^4 | X = x\right)$ there is a bound $0 < L_\mu < \mu_i(x) < M_\mu < \infty, i = 2, 4$; $\mu_2(x)$ is continuous.*

The fourth order moment is not necessarily needed here, we just need some way to insure that the CLT applies. Boundedness of the second moment of the error and continuity is used substantively in the proof.

The properties of the regression function over the space Ξ are tied in with the properties of the conditioning space. Let d_1 and $d_2 \geq 0$ define two scalar functions with arguments $x \in \Xi^{[q]}$ and $\tilde{x}$. We say that $d_1(x, \tilde{x}) = O(d_2(x, \tilde{x}))$, if for any x, some h_0 and any $\tilde{x} \in C(x, h) \subset \Xi^{[q]}$ with $h < h_0$ there exists a positive constant B_x such that

$$|d_1(x, \tilde{x})| \leq B_x d_2(x, \tilde{x}).$$

Assumption 3 *(a) The function $m(x)$ on the space $\Xi^{[q]}$ is such that*

$$|m(x) - m(\tilde{x})| = O\left(\max_j \left\|x^j - \tilde{x}^j\right\|_j\right).$$

(b) The space $\Xi^{[q]}$ is such that there an inner product is defined and the function $m(x)$ is such that for some $\nabla m(x) = \left(m_1'(x), ..., m_q'(x)\right) \in \Xi^{[q]}$ at every x and some $\varepsilon > 0$ with

$$\tilde{x} \in \left\{\tilde{x} : \max_j \left\|x^j - \tilde{x}^j\right\|_j < \varepsilon\right\}$$

$$m(x) = m(\tilde{x}) + \sum_{j=1}^{q} \left\langle m'_j(x), x^j - \tilde{x}^j \right\rangle + O\left(\max_j \left\| x^j - \tilde{x}^j \right\|_j^2 \right).$$

Here $m'_j(x)$ is the functional derivative at x with respect to x^j.

Assumption 3(a) in R^q holds under differentiability of $m(x)$, or even with Lipschitz continuity. For Assumption 3(b) to apply the space $\Xi^{[q]}$ needs to be a linear function space; inner product $\langle \cdot, \cdot \rangle$ needs to be defined in $\Xi^{[q]}$, then $\nabla m(x)$ is an element of $\Xi^{[q]}$. In R^q the vector $\nabla m(x)$ represents partial derivatives of $m(x)$, in the general case $\nabla m(x) := (m'_1(x), \cdots, m'_q)^T$ is a vector of functional derivatives (where applicable); Assumption (b) is similar to the one in Ferraty, Nagy (2022).

In asymptotic analysis on R^q further smoothness assumptions are often made on $m(x)$, e.g. Theorem 2.2 in Li, Racine (2007) requires three times continuous differentiability of $m(x)$ with x being an interior point, but here no extra smoothness is required.

2.5 The conditioning probability measures

2.5.1 Measures on R^q

Borel probability measures on R^q are given by distribution functions. The Lebesgue decomposition represents the distribution F_X on R^q as a mixture of an absolutely continuous distribution, $F^{a.c.}$, a singular distribution (the distribution function is continuous but there is no function that integrates to it), F^s, and a discrete distribution, F^d :

$$\begin{aligned} F_X(x) &= \alpha_1 F^{a.c.}(x) + \alpha_2 F^s(x) + \alpha_3 F^d(x); \\ \alpha_l &\geq 0 \; l = 1,2,3; \; \sum\nolimits_{l=1}^{3} \alpha_l = 1. \end{aligned}$$

Clearly, if $\alpha_3 = 0$, then the distribution function $F_X(x)$ is a continuous function.

Denote by $\underline{h}$ the smallest, and by $\bar{h}$ the largest of the components in the vector h that defines the cube $C(x,h)$. If the distribution function is continuous then the small cube measure $P_X(C(x,h))$ goes to zero.

Remark 1 *In R^q the results obtained from considering the cube $P_X(C(x,h)) = F_X(x-h, x+h)$ apply equally if we were to consider the ball, $B(x,r)$ centered at x with a radius r (due to equivalence of norms). Indeed,*

$$B\left(x, \underline{h}/\sqrt{2}\right) \subset C(x,h) \subset B\left(x, \sqrt{2}\bar{h}\right) \subset C\left(x, \sqrt{2}\bar{h}\right).$$

Recall that support of the probability measure, P_X, associated with F_X is the set of points $x \in R^q$ where for any $r > 0$ we have for the ball $B(x,r)$

$$P_X(B(x,r)) > 0.$$

A well-known regularity condition, which applies among others to fractal measures, is Ahlfors regularity (see, e.g., Ahlfors, 1966). An Ahlfors-regular

distribution (Borel probability measure) on R^q is such that for some $C \geq 1, 1 \geq s > 0$ and $h > 0$ at every support point

$$C^{-1}h^{sq} \leq P_X(B(x,h)) \leq Ch^{sq}. \tag{12}$$

If F_X satisfies the Ahlfors regularity condition, then for some constants, $L_F \leq C^{-1}2^{-sq/2}$, $M_F \geq C2^{-sq/2}$ at every point of support we get the bounds

$$0 < L_F(2\underline{h})^{sq} \leq P_X(C(x,h)) \leq M_F\left(2\bar{h}\right)^{sq} < \infty. \tag{13}$$

Hence, the small cube probability is polynomial in h.
If x is an isolated mass point then (13) applies with $s = 0$. We define here a distribution class that extends the Ahlfors regularity condition to include $s = 0$.

Definition 1 *Denote the class of distributions that at point of support x satisfy* (13) *for some $0 \leq s \leq 1$ by $\mathcal{A}(s)$.*

Kankanala and Zinde-Walsh (2023) detail examples of distributions in the class $\mathcal{A}(s)$ that include absolutely continuous distributions, as well as singular: fractals, measures supported on a lower dimensional space e.g. reduced rank or multicollinear, measures on random vectors with some absolutely continuous and some discrete components.

Consider now a distribution that is represented as a mixture of distributions $F_l \subset \mathcal{A}(s_l)$:

$$F_X = \sum_l \alpha_l F_l;\ \sum_l \alpha_l = 1;\ \alpha_l > 0.$$

Then with $s = \inf\{s_l\}$ and $s' = \sup\{s_l\}$ for some bounded constants $L_F \geq 0,\ M_F > 0$,

$$L_F\underline{h}^{s'q} \leq P_X(C(x,h) \leq M_F\bar{h}^{sq}. \tag{14}$$

Assumption 4 below provides a class of measures on $\Xi^{[q]}$ that on R^q generalizes the class $\mathcal{A}(s)$ and allows for mixtures of distributions from such classes for different s.

Assumption 4 *Given any point $x \in \Xi^{[q]}$ in the support of P_X, the probability measure P_X is such that for some bandwidth sequence h with $\bar{h} \to 0$ as $n \to \infty$ and for some $0 < \varepsilon < 1$*

$$\text{(a) } nP_X(C(x,h) \to \infty; \tag{15}$$

$$\text{(b) } \frac{P_X(C(x,h)}{P_X(C(x,\varepsilon h)} < C_F < \infty. \tag{16}$$

with $C_F \geq 1$.

Definition 2 *Denote the class of measures that satisfy* (15, 16) *by $\mathcal{D}$.*

Examples of distributions in class $\mathcal{D}$ on R^q

1. Suppose that F_X belongs to a finite mixture of distributions from $\mathcal{A}(s)$ for different $s : F_X = \sum_{l=1}^{L} \alpha_l F_{s_l}$ with $0 \leq s_1 \leq ... \leq s_L \leq 1,\ \alpha_l > 0$, and $\sum_{l=1}^{L} \alpha_l = 1$. Every point in the support of F_X is a point of continuity if $s_1 > 0$. To verify that (16) of Assumption 4 is satisfied, we first note that for any class $\mathcal{A}(s_l)$ the condition holds with some corresponding $C_F = C(s_l)$ and probability measure P_{s_i} corresponding to the distribution F_{s_i}. As $P_X(C(x,h)$ can be represented as a sum of probabilities conditional on x being generated by a term F_{s_l} in the mixture times the marginals, we then apply (16) to every term in the sum

$$\begin{aligned} \sum_{l=1}^{L} \alpha_l P_{s_l}(C(x,h) &< \sum_{l=1}^{L} \alpha_l C(s_l) P_{s_l}(C(x,\varepsilon_i h) \\ &\leq \max C(s_l) \sum_{l=1}^{L} \alpha_l P_{s_l}(C(x,\bar{\varepsilon} h). \end{aligned}$$

with $\bar{\varepsilon} = \max(\varepsilon_l)$ and $C_F = \max C(s_l)$ the condition (16) therefore is shown to hold; (15) holds similarly.

2. Consider a point x^* in the support of F_X where the distribution F_X is discontinuous, a mass point. $F_X(x) = \alpha I(x \geq x^*) + (1-\alpha) F^c(x)$, with the distribution function F^c continuous. Suppose that $F^c(x)$ satisfies Assumption 4, then at any point of support $x \neq x^*$, with $h < \|x - x^*\|$ Assumption 4 holds. At the point x^*

$$nP_X(C(x^*,h) \geq n\alpha,$$

thus (15) holds regardless of h. We have that $P_X(C(x^*,h)$ is bounded and $P_X(C(x^*,\varepsilon h) > \alpha$, thus (16) also holds for $C_F \geq 2/\alpha$ (regardless of h) at x^*.

2.5.2 Measures on metric spaces

A particular class of probability measures for which the results in functional NW estimation are established has small ball probability at a point x of the form

$$P_X(B(x,h)) = C_1 h^{\gamma} \exp\left(-C_2/h^{\beta}\right) \tag{17}$$

with positive C_1, γ, as well as at least non-negative C_2 and β. When $\beta = 0$ these small ball probabilities satisfy Assumption 4. However with C_1, C_2, β positive constants the probability measure is such that the small ball probability declines exponentially in h; then (16) does not hold. In an infinite dimensional space $\beta > 0$ is a natural condition. If $\beta = 0$ then locally the support is likely finite dimensional. The distribution could be supported on a finite-dimensional manifold in the function space.

Ferraty and Vieu (2006) and Ferraty et al. (2007) explored conditions similar to the ones in (15, 16). The condition (15) is maintained in all cases, but (16) may not always hold. Below we specify the relation of this condition to the ones given in the literature.

Following Ferraty et al. (2007, p 270) write

$$\tau_h(\varepsilon) = \frac{P_X(B(x, \varepsilon h))}{P_X(B(x, h))}.$$

Assumption H_3 in Ferraty et al. require that as $h \to 0$ the functions $tau_h(\varepsilon)$ converge:

$$\tau_h(\varepsilon) \to \tau_0(\varepsilon) \quad \text{for every } \varepsilon, \tag{18}$$

where the limit could be a regular function or a delta-function $\delta(\varepsilon = 1)$. Proposition 1 of that paper provides examples of distributions for which $\tau_0(\varepsilon) > 0$. A similar condition (Lemma 4.4) in Ferraty and Vieu (2006) is that

$$\int_0^v P_X(B(x, u))\, du > CvP_X(B(x, v)) \tag{19}$$

holds for some $C > 0$, $v_0 > 0$ and any $v < v_0$. The following lemma demonstrates that condition (16) (Assumption 4(b)) is a necessary condition for (18) to hold with $\tau_0(\varepsilon) > 0$ and a necessary condition for (19) to hold. Assumption 4(b) therefore is somewhat more general. As our results will show it is also sufficient for asymptotic normality without the need of a kernel of type I.

Lemma 1 *The condition* (16) *of Assumption 4 is a necessary condition (a) for (18) with* $\tau_0(\varepsilon) > 0$ *and (b) for (19) to hold.*

The lemma situates the condition (16) in relation to (18) and (19). In particular, as will be shown below, it determines when kernels of type II are useful. This is important for the numerous results that routinely use such kernels.

3 Asymptotic normality of kernel estimators

The technical results of Appendix A provide bounds on various moments that are instrumental in establishing distributional limits. The novel results in the Appendix apply in R^q and product $\Xi^{[q]}$ spaces. The bounds depend on the bandwidth h through the small cube probability, $P_X(C(x, h))$. When the measure P_X is given by an absolutely continuous distribution function, F_X, in R^q the limit results provide the same rates as those established in the literature under further differentiability assumptions on the distribution function, without relying on that assumption, furthermore, the limit results here do not require absolute continuity. When $\Xi = \Xi^{[1]}$ the results are the same as in the functional regression literature, however, we extend the functional regression results to $\Xi^{[q]}$.

3.1 NW estimator in product space

Given a sample $\{(X_i, Y_i)\}_{i=1}^n$ the Nadaraya-Watson estimator of the regression function $m(x)$ is defined for a kernel function $K(.)$ of Assumption 1(a) and a bandwidth (dependent on n), h, as

$$\widehat{m}(x) = \frac{A_n(x)}{B_n(x)}; \tag{20}$$

$$A_n(x) = \frac{1}{n}\sum_{i=1}^n K(W_i(x)) Y_i;\ B_n(x) = \frac{1}{n}\sum_{i=1}^n K(W_i(x)), \tag{21}$$

with $W_i(x) = h^{-1}(x - X_i)$ on R^q and $W_i(x) = h^{-1}\|x - X_i\|$ on the metric space $\Xi^{[q]}$. As the sample size changes the bandwidths go to zero. To simplify notation we do not indicate the dependence of the bandwidth on n.

For $\Xi = R^q$ the denominator, $B_n(x)$ is proportional to the usual kernel density estimator, given by $h^{-q}B_n(x)$, at point x. When continuous density, $f_X(x)$ exists, then the estimator $h^{-q}B_n(x)$ consistently estimates $f_X(x)$, but if density does not exist, $h^{-q}B_n(x)$ diverges to infinity.

Consistency of the NW estimator $\widehat{m}(x)$ over a metric space $\Xi = \Xi^{[1]}$ was established for general distribution classes in Gyorfi et al. (2002). The limit distribution for the estimator was examined in Ferraty et al. (2007) and Geenens (2015). The bandwidth sequence there is univariate but could depend on the point x. The results were established under the full Assumption 1 on the kernel. The usual kernels such as the Epanechnikov were not considered because it was argued that the small ball probability $P_X(B(x,h))$ could be of the form $C_1 h^\gamma \exp\left(-C_2/h^\beta\right)$ with C_1, C_2, γ, β positive constants that does not satisfy Assumption 4(b) required for consistency when $K(1) = 0$. The examples of the distributions, such as those in class $A(s)$, where small cube probability $P_X(C(x,h))$ are polynomial in h were considered to be more suitable for a finite dimensional rather than infinite dimensional context.

Here we consider the space Ξ, that could be R^q, or the functions (or metric) space $\Xi^{[1]}$, but also could be some product of metric spaces, $\Xi^{[q]}$.

Theorem 1 *Under either of the following sets of assumptions (i) Assumptions 1 (a-d), 2, 3 (a) and 4(a), or (ii) Assumptions 1 (a-c), 2, 3(a) and 4(a,b)*

$$\frac{EB_n(x)}{\sqrt{varA_n(x)}}\left(\widehat{m}(x) - m(x) - bias(\widehat{m}(x)\right) \to_d Z \sim N(0,1).$$

The point-wise consistency and asymptotic normality of the Nadaraya-Watson estimator with undersmoothing follows and is summarized in the following theorem.

Theorem 2 *Under the conditions of Theorem 1 and for h such that* $\frac{\bar{h}}{\sqrt{varA_n(x)}} \to 0$

$$\frac{EB_n(x)}{\sqrt{varA_n(x)}}\left(\widehat{m}(x) - m(x)\right) \to_d Z \sim N(0,1).$$

Remarks. 1. The limit result shows that for a distrbution in $\mathcal{A}(s)$ if $s = 1$, as is the case when density exists, the standard convergence rate $n^{1/2}h^{q/2}$ applies. However, this rate holds even when the density is discontinuous, thus even without the usual smoothness assumptions made in the literature we can provide statistical guarantees for the rate.

2. If there is singularity at the point x with $s < 1$, then the rate of convergence could be faster than in the absolutely continuous case $\left(n^{1/2}h^{sq/2} > n^{1/2}h^{q/2}\right)$, mitigating somewhat the "curse of dimensionality".

3. If x is a mass point then convergence of the estimator at that point is at the parametric rate $n^{1/2}$.

4. The values of h that achieve the trade-off between the bias and variance for Theorem 2 to hold at a point x may vary over a very restrictive range. Thus e.g. over the class $\mathcal{A}(s)$ the rate for h would have to be in the range $\left(n^{-\frac{2}{sq}}, n^{-\frac{2}{sq+\epsilon}}\right)$ for some $\epsilon > 0$.

5. It is possible that as $n \to \infty, h \to 0$ there is a sequence $\alpha(n,h) \to \infty$ such that a limit $\sigma^2_{\widehat{m}(x)}$ for $\alpha(n,h)\frac{varA_n(x)}{(EB_n(x))^2}$ exists. Then

$$\sqrt{\alpha(n,h)}\left(\widehat{m}(x) - m(x)\right) \to_d N\left(0, \sigma^2_{\widehat{m}(x)}\right)$$

However, there is no guarantee that a limit exists and as $n \to \infty$ the "standard deviation" may oscillate between the upper and lower bound on the ratio. Existence of a limit would require an additional assumption such as the one below.

Assumption 5 *As* $n \to \infty$, $h \to 0$

$$\begin{aligned} \frac{EB_n(x)}{P_X(C(x,h))} &\to \bar{B}_1(x); \\ \frac{n \cdot varA_n(x)}{P_X(C(x,h))} &\to \bar{A}_2(x). \end{aligned}$$

Assumption 5 holds e.g. for fractal distributions on R^q for which $\lim h^{-sq}F_X(x-h, x+h)$ exists for $h \to 0$ (e.g. with a non-zero density at x and $s = 1$).

Define now

$$\begin{aligned} \alpha(n,h) &= nP_X(C(x,h)); \\ \sigma^2_{\widehat{m}(x)} &= \frac{\bar{A}_2}{\bar{B}_1^2} \end{aligned}$$

with $C(x,h)$ as defined in (4).

Theorem 3 *Under the conditions of Theorem 2 and Assumption 5 with* $\alpha(n,h) \to \infty$

$$\sigma^2_{\widehat{m}(x)} = \lim_{n\to\infty} \alpha(n,h)\frac{var(A_n(x))}{E(B_n(x))^2}$$

and for h such that $\alpha(n,h)\bar{h}^2 \to 0$

$$\sqrt{\alpha(n,h)}\left(\widehat{m}(x) - m(x)\right) \to_d N\left(0, \sigma^2_{\widehat{m}(x)}\right).$$

Assumption 5 holds for fractal distributions on R^q in the $\mathcal{A}(s)$ class for which $\lim h^{-sq} F_X(x-h, x+h)$ exists for a scalar sequence $h \to 0$. In particular, an absolutely continuous distribution satisfies this assumption.

Thus for an absolutely continuous distribution the kernel estimator satisfies asymptotic normality with the standard rate since there $P_X(C(x,h) = F_X(x-h, x+h) = O(h^q)$. This holds even when the density has kinks.

The novel contribution here is to extend asymptotic normality results to the product space $\Xi_i^{[q]}$ extending the results to multivariate functional regression as well as multivariate regression with components in different functional and vector spaces.

3.2 Local linear estimator in R^q

Similarly to the local constant NW estimator, results for local linear estimators in R^q were obtained for smooth conditioning distributions (see, e.g., textbook Li and Racine, 2007, Theorem 2.7). The extensions in Lee et al. (2019) and in Linton and Xiao (2019) rely on smoothness of the conditioning distributions as well. Using the results on the moments bounds, we extend these results to classes of distributions such as class $\mathcal{D}$.

For functional regression Ferraty and Nagy (2022) provide limit theory on a hilbert space $\Xi^{[1]}$. For product spaces $\Xi^{[q]}$ those results could be extended, but such an extension is beyond the scope of this paper.

Here we focus on extending the limit theory for LL estimators in R^q to a wide class of probability measures.

Consider the expressions $(11, 9, 10)$. In R^q the local linear LL estimator vector, $\hat{\beta}(x) = \left(\hat{\beta}^0(x), \hat{\beta}^1(x) ... \hat{\beta}^q(x)\right)'$ represents the estimator of the regression function, $m(x)$ as $\hat{\beta}^0(x)$ and each component $\hat{\beta}^j(x)$ estimates the partial derivative $\frac{\partial}{\partial x^j} m(x)$. The expectations and variances for the matrix elements in those formulae satisfy the conditions under which bounds can be derived by applying the general results of Appendix A. Thus in the proofs we show that $E\widetilde{B}_n(x)$ as well as the variances (and higher moments) of the components of $\widetilde{B}_n(x)$ are bounded by $P_X(C(x,h))\bar{M}_{EB^m}$ and, with Assumption 3, the components of the matrix

$$var\widetilde{A}_n(x) = E\left[\widetilde{A}_n(x) - E\widetilde{A}_n(x)\right]\left[\widetilde{A}_n(x) - E\widetilde{A}_n(x)\right]'$$

are bounded from above by some $n^{-1}P_X(C(x,h))\bar{M}_{EAA'}$. We make an additional assumption.

Assumption 6 *The matrix $E\widetilde{B}(x)$ is invertible; the matrix $var\widetilde{A}(x)$ is positive definite.*

Remark here that full rank of the matrices cannot be taken for granted when the underlyng distribution is not absolutely continuous.

The moments have a rate that is determined by the small cube probability, in R^q. The following assumption requires that limits exist to simplify the expressions.

Assumption 7 *As* $n \to \infty$, $h \to 0$

$$\frac{E\widetilde{B}_n(x)}{P_X\left(C\left(x,h\right)\right)} \to \bar{B}_1\left(x\right);$$
$$\frac{n \cdot var\widetilde{A}_n\left(x\right)}{P_X\left(C\left(x,h\right)\right)} \to \bar{A}_2\left(x\right).$$

We express the final result in terms of the target values of the function $m\left(x\right)$ and its vector of partial derivatives $\nabla m\left(x\right)$ by expressing the vector $\hat{\beta}\left(x\right)$ as $\begin{pmatrix} m\left(x\right) \\ \widehat{\nabla m}\left(x\right) \end{pmatrix}$. By maintaining Assumption 3(b) bias control is achieved.

Theorem 4 *Under either of the following sets of assumptions (i) Assumptions 1 (a-d), 2, 3 (b) and 4(a), or (ii) Assumptions 1 (a-c), 2,3(b) and 4(a,b) and Assumption 7 with* $\alpha\left(n,h\right) = nF_X\left(x-h, x+h\right) \to \infty$ *for* h *such that* $D\sqrt{\alpha\left(n,h\right)}\bar{h} \to 0$

$$D\sqrt{\alpha\left(n,h\right)}\left(\begin{pmatrix} \hat{m}\left(x\right) \\ \widehat{\nabla m}\left(x\right) \end{pmatrix} - \begin{pmatrix} m\left(x\right) \\ \nabla m\left(x\right) \end{pmatrix}\right) \to_d N\left(0, \bar{B}_1\left(x\right)^{-1}\bar{A}_2\left(x\right)\bar{B}_1\left(x\right)^{-1}\right).$$

This provides the limit normality (with undersmoothing) for the local linear estimator in R^q under a general class of distributions, $\mathcal{D}$.

4 Simulations

4.1 Univariate (Point mass example)

In this section we consider univariate conditional expectations $m(x)$ under **non-standard distributional assumptions** on the regressor. In particular we will allow for mass points and mixtures with peaked normals where derivatives could vastly exceed those for the standard cases considered in the literature. Various variables may exhibit such features such as subjective probability, firm's earnings, income, job tenure, household expenditure, working hours, age at retirement, and neonatal mortality (heaped duration data) (Arulampalam et al, 2017). See also Compiani and Kitamura (2016) on use of mixtures in econometric models.

Here, we consider the conditional mean function[2]

$$m(X) = \sin(2.5X)$$

[2]Additional details of the simulation and analysis under a different conditional mean functions can be found in a supplementary document.

a graph of which is presented in Figure 1.

Include Figure 1[3]

To allow for mass points, we consider a distribution used by Jun and Song (2019):

$$F_X(x) = pF^d(x) + (1-p)\Phi(x)$$

for $p = 0$, 0.01, 0.1 and 0.2 where $F^d(x)$ is the discrete uniform distribution function with $D = -1, 0, 1$ and $\Phi\cdot$ is the standard Gaussian distribution function; Here F_X is a distribution which permits mass points when $p \neq 0$. Alongside, we consider the trinormal mixture used in Kotlyarova et al. (2016), whose density is given by

$$f_X(x) = 0.5\phi(x + 0.767) + 3\phi\left(\frac{x + 0.767 - 0.8}{0.1}\right) + 2\phi\left(\frac{x + 0.767 - 1.2}{0.1}\right)$$

where ϕ is the standard Gaussian density function. This a.c. distribution represents features (high derivatives) that are more reflective of unsmooth densities we want to emphasize here. Illustrative graphs of these distributions are given in Figure 2.

Include Figure 2

We simulated 500 random samples of 1000 observations (in supplemental material we also consider $n = 100$ and $n = 500$) from the model

$$y = \sin(2.5X) + \sigma\varepsilon,$$

where $\{\varepsilon_i\}$ is drawn independently from $\{X_i\}$ from the standard normal distribution and select σ to yield a given signal to noise ratio, snr, here selected to equal one (in supplemental material we consider alternative signal to noise ratios).

We obtain the standard leave-one-out cross validated bandwidth for the NW local constant estimator using the NP package in R (Hayfield and Racine, 2008) with the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)1(u^2 \leq 1)$ (a type II kernel). To evaluate the sensitivity of our results to the bandwidth selected (and evaluate the effect of under- and oversmoothing) we apply different multipliers to the cross-validated bandwidth (specifically 0.5, 0.75, 1, and 1.25).

The nonparametric fit is evaluated at a grid of 100 points x; the grid is drawn to be representative of the true conditioning distribution (evaluates F_X^{-1} on a uniform grid on $[0, 1]$). The bias, standard deviation and root mean squared error (RMSE) over the 500 simulations are presented in Figure 3. Their details (provided across three columns) are displayed based on using the cross-validated bandwidth h_{cv} in red, two undersmoothed bandwidths $0.5h_{cv}$ in green and $0.75h_{cv}$ in blue, and an oversmoothed bandwidth $1.25h_{cv}$ in purple.

Include Figure 3

[3]The figures and tables can be found in the Annex at the end of the paper.

The different rows of panels reflect the different conditioning distributions under consideration. As expected, at all points, the undersmoothed bandwidth results in a smaller bias and larger variance, where the standard deviation increases at points where the distribution is more sparse. The crossvalidated and oversmoothed bandwidth provide reasonable performance in that this bias is small. The impact of increasing the probability of mass points is marked, in particular the impact on the standard deviation around the point mass points becomes more pronounced, reflecting the fact that at these points the rate of convergence is much faster (parametric rate). While, as expected, the bias is close to zero at the mass points, the impact of the mass points is non-negligible when evaluating points slightly above/below (Need to decide whether grid points should be bounded away from mass points.) The RMSE result summarizes these two results which underlie the importance of considering the different rates of convergence when constructing point and uniform confidence bounds (an issue we will address in a separate paper). The bottom panel, where we consider the trinormal mixture, displays an interesting pattern where small standard derivations are observed in regions where the derivative of the trinormal mixture is large; the bias and standard deviation are fairly stable when x takes values from say [-1,0.5] where the distribution does not have large derivatives, while the standard deviation increases again where the distribution is more sparse. This shows that we find a similar pattern for an a.c. distribution at points where density derivatives are large as we do for singular points for non-absolutely continuous conditioning distribution.

4.2 Bivariate (with singularity: X_1 and X_2 on unit circle)

In this section we consider bivariate regression models in the presence of a singularity. In particular we consider the setting where the regressors lie on a circle. This example may find its origin in Hotelling's (1929) spatial model of horizontal differentiation which assumes that each consumer has an 'ideal' variety identified by his location on the unit circle (bounded product space). This boundedness of the product space underlies a positive relationship between market size and the price elasticity. A large market (in the sense of a large population), leads to more varieties being produced implying a more crowded product space and more substitution between goods. See also Desmet and Parente (2010).

Here we consider the following bivariate conditional mean function

$$m(X) = X_1 + X_2$$

in the presence of the deterministic relation between X_1 and X_2 given by the unit circle, that is

$$X_1^2 + X_2^2 = 1.$$

The bivariate distribution $F_X(x_1, x_2)$ is uniform on the circle. To obtain our random sample, we draw ϕ from $U[0, 2\pi]$, and evaluate $x_1 = \cos(\phi)$ and $x_2 =$

$\sin(\phi)$. In Figure 4, we present the conditional mean function.

Include Figure 4

We simulated 500 random samples of 1000 observations from the model

$$y = X_1 + X_2 + \sigma\varepsilon$$

where the error ε is drawn independently from the standard normal distribution. We select σ to yield a a signal to r=noise ratio of 1.ensure which results in an average signal to noise ratio equalling 1.0. Other signal to noise ratios and sample sizes are considered in the supplementary material. The leave-one-out cross validated bandwidths for the NW local constant estimator, as expected, are comparable for both arguments x_1 and x_2 given its symmetric formulation. We consider the effect of different bandwidths by multiplying the vector h_{cv} by 1.25, 1, 1.75 and 1.5 as before (results where we vary the multiplicative factors by component are available in the supplemental material).

In order to evaluate the performance of the NW local regression estimator in the presence of this singularity, it is important to recognize that the deterministic relationship between X_1 and X_2 implies a reduced dimension which can be achieved by a single index model (Ichimura, 1993), which is the true model (oracle) here. To implement the single index estimator, we follow the approach used in Hardle et al. (1993) which jointly minimizes a (weighted) least-squares cross-validation function with respect to the parameters and (scalar) bandwidth; the leave-one-out cross-validated bandwidth were obtained using the NP package in R with the Epanechnikov product kernel. On average the estimate on x_2 in the single index (with coefficient on x_1 normalized to 1) is close to 1 as expected. For given estimate on x_2 we evaluate the performance of changing the bandwidth h by using the same multiplicative factors on h_{cv}.

In Table 1, the average mean absolute error (MAE) and average mean squared error (MSE) over the 500 simulations are presented for both estimators. They reveal that the overall performance of the estimators is comparable, so that the benefits of the the reduced dimension achieved by the single index model is appropriately reflected in the NW kernel regression estimator.

Include Table 1

As in the univariate setting, we analyse the bias, standard deviation and RMSE of the nonparametric estimators over a grid of 100 (x_1, x_2) points based on equidistant values of ϕ on $[0, 2\pi]$. In Figure 5, we provide these details (provided across three columns) for different bandwidths. The top two panels provide the results for the NW local constant estimator with ϕ and x_1 (with $x_2 \geq 0$) on the horizontal axis respectively. The bottom two panels provide the results for the single index estimator, again displayed with either ϕ or x_1 on the horizontal axis.

Include Figure 5

The standard deviation of the estimates show a much large variation over the values of (x_1, x_2) based on the single index model, while exhibiting a more

stable profile in the local kernel regression estimator. The lowest standard deviation for the single index estimator shows when both (x_1, x_2) are positive (or by symmetry when both $(x_1, x_2$ are negative), which is also where the biases are largest. Overall, the simulations suggest that the improved convergence due to singularity provides results that are comparable (and sometimes superior) to the true single index model (oracle).

4.3 Bivariate (with reduced dimension: X_1 and X_2 on straight line)

In this section we consider bivariate regression models in the presence of a reduced dimensionality for sub-populations. The example that we may envision here could be where y is demand, x_1 income of the husband, and x_2 income of the wife, but where e.g., for purposes of tax minimization, their combined income is given by $x_1 + x_2 = d(k)$ for $k = 1, 2, 3$. In this set-up, essentially we only have one absolutely continuous regressor together with the (ordered) discrete random variable $d(k)$ instead of two absolutely continuous regressors, where the latter may not be obvious from the original data. It is important to recognize that in the presence of such reduced dimensionality, the rate of convergence is determined by the lower number of continuous regressors only, see also Li and Racine (2004).

We discuss other examples, where the reduced dimensionality for sub-populations (with unobserved heterogeneity) would render this a useful framework to consider.

Here we consider the following bivariate conditional mean function

$$m(X) = \log(X_1) + \log(X_2)$$

in the presence of the reduced dimensionality with

$$X_1 + X_2 = d(k), \text{ with } k = 1, 2, 3$$

for given values of $d(k)$, specifically $d(1) = 4, d(2) = 6, d(3) = 7$ and $x_1 > 0, x_2 > 0$. The probability of an observation belonging to a sub-population with $k = 1, 2, 3$ is given by $p_1 = 0.5, p_2 = 0.3$, and $p_3 = 0.2$ respectively and x_1 is drawn from the uniform distribution: $U[1, 3]$. The conditional mean function, is presented in Figure 6.

Include Figure 6

We simulated 500 random samples random samples of 1000 observations from the model

$$y_1 = \log(X_1) + \log(X_2) + \sigma\varepsilon$$

where ε is drawn independently from $\{X_i\}$ from a standard normal distribution and σ is chosen to provide a signal to noise ratio equalling one (in the supplemental material other signal to noise ratios are considered).

Here we implement the local kernel regression estimator in the presence of the reduced dimensionality of the regressors twice: first without recognizing the reduced dimensionality aspect (which we label as NP(c) as it treats both regressors as absolute continuous), and once recognizing its presence (which we label as NP(d) to reflect the associated discrete component). To implement NP(d) we use the generalized product kernel on $(x_1, d(k))$ using the Wang and van Ryzin (1981) kernel for the ordered discrete regressor. As the model would permit the single index model specification when we use $(\log(x_1), \log(x_2))$ as our conditioning variables, we consider this estimator here as well (which we label as SI).

In Table 2, we present the Mean Absolute Error (MAE) and Mean Squared Error (MSE) using these methods

Include Table 2

They show a comparable performance across the estimators. The overall performance of the local kernel regression estimator that does recognize the reduced dimension of the conditioning variables (NP(d)) suggests a slight improvement over the local kernel regression estimator that does not (NP(c)). Nevertheless, the results are comparable to the benefits associated with the reduced dimension achieved by the single index model (which does require one to recognize that $(\log(x_1), \log(x_2))$ are are the conditioning variables (oracle).

In Figure 7-9, we present the bias, variance and RMSE of the nonparametric NP(c), NP(d) and SI estimators over a grid for x_1 separately for each sub-population using a selection of multiples of bandwidths centred around $(h_{cv,1}, h_{cv,2})$, where $h_{cv,j}$ is the cross validated bandwidth for x_j.

Include Figures 7-9

The standard deviation and RMSE over x_1 and values of $d(k)$ show a comparable pattern which is fairly stable across the range of x_1 and reveals increases at the boundary values with x_1 taking values close to 1 or 3 (not dissimilar for different values of $d(k)$). While the bias clearly is smallest when using the undersmoothed bandwidth, the impact of varying the bandwidths on the bias appears to be stronger for NP(c), the model that does not recognize the dimension reduction, than NP(d) (unsure what explanation this may have, if any). Finally, the single index which uses $\log(x_1)$ and $\log(x_2)$ as conditioning values exhibits less bias at the lower value of x_1 than displayed by the NP estimators; pattern on the standard deviation does reveal dissimilarities depending on the value of $d(k)$ unlike the NP estimators.

To be completed.

4.4 Bivariate (with reduced dimension: X_1 and X_2 perfect collinear)

In this section we briefly discuss the setting where the regressors are perfect multicollinear. Simulations reveal that cross-validation has the tendency to "smooth

out", that is, remove the perfect collinear (and hence irrelevant) regressor by selecting the bandwidth on one of the regressors in orders of magnitude larger than the standard deviation of the regressor. These simulations extend the results discussed in Hall, et al (2007), to the setting of perfect multicollinearity.

5 Conclusions

To be completed

A The general technical results

This Appendix provides derivations for moments and bounds on the moments that are used in establishing the limit properties of the estimators.

Derivation of moments standard in the kernel regression literature on R^q typically relies on the existence of density and utilizes an expansion of the density function. Here we provide derivations that do not use density and apply to regression in R^q as well as to multiple functional regression on $\Xi^{[q]}$. Our results provide calculation for the expectation relative to the measure P_X of a function $\psi\left(\frac{x-X}{h}\right)$ defined in the vicinity of x. A common example of $\psi(.)$ is a kernel function or its power.

For $W = \left(W^1, ..., W^q\right)$ with components $W^j = \frac{x^j - X^j}{h^j}$ consider a random function of X in the vicinity of $x : \psi(W) =$

$$\psi\left(\frac{x-X}{h}\right) = \psi\left(\frac{x^1 - X^1}{h^1}, ..., \frac{x^q - X^q}{h^q}\right) \tag{A.1}$$

In particular, it could be a product function (such as a product kernel):

$$\psi\left(\frac{x-X}{h}\right) = \prod_{j=1}^{q} \psi_j\left(\frac{x^j - X^j}{h^j}\right). \tag{A.2}$$

Consider the set of the indices $\{1, 2, ..., q\}$; there are 2^q subsets of this set including the empty set, $\varnothing$. Denote each subset by I_ξ; $\xi = 0, 1, ..., 2^q - 1$ with $I_0 = \varnothing$. The indices ξ are ordered such that the indices are non-decreasing in the cardinality of the set and are ordered lexicographically for each cardinality. Let $q(\xi)$ denote the cardinality of the subset I_ξ. The complement of the subset I_ξ is denoted by I_ξ^c. Thus, for instance, for $q = 3$, there are 8 subsets: $I_0 = \varnothing; I_1 = \{1\}, I_2 = \{2\}, I_3 = \{3\}, I_4 = \{1, 2\}, I_5 = \{1, 3\}; I_6 = \{2, 3\}, I_7 = \{1, 2, 3\}$.

Let $\prod_{j \in I_\xi}(-\partial_j)$ denote an operator that, when applied to a differentiable function $g(z) = g\left(z^1, ..., z^q\right)$ at z, maps it to its partial derivative for $j_1 < ... <$

$j_{q(\xi)} \in I_\xi$, times $(-1)^{q(\xi)}$ that is

$$\left(\prod_{j\in I_\xi}(-\partial_j)\right) g(z) = (-1)^{q(\xi)} \frac{\partial^{q(\xi)}}{\partial_{j_1}...\partial_{j_{q(\xi)}}} g(z).$$

We call a function $g(z)$ "sufficiently differentiable" if for any set I_ξ the derivative $\left(\prod_{j\in I_\xi}\partial_j\right) g(z)$ exists at any point on the interior of its support, is continuous and extends continuously to the boundary.

Using the delta-function operator, $\delta\left(z^j = a\right)$ that applied to a continuous function $g\left(z^1, \cdots, z^j, \cdots, z^q\right)$ sets the jth component to a scalar value a, we define the operator $\Delta_{a,\xi}$:

$$\Delta_{a,\xi} g(z) = \prod_{j\in I_\xi^c} \delta\left(z^j = a\right) \prod_{j\in I_\xi} (-\partial_j)\, g(z); \tag{A.3}$$

with typically $a = -1$ or 1 denoting a point on the boundary.

We deal with three situations. First, the following Lemma A.1 provides the expectation of $\psi(W) = \psi\left(\frac{x-X}{h}\right)$ given in (A.1) with respect to the probability measure on R^q given by a distribution function F_X on R^q, for a sufficiently differentiable random function ψ with support $[-1,1]^q$. Second, Corollary A.1 to Lemma A.1 details this result for a symmetric function. Third, Lemma A.2 provides the expectation of a a sufficiently differentiable function

$$\psi(W) = \psi_+\left(\frac{\left\| x^1 - X^1\right\|_1}{h^1}, \cdots, \frac{\| x^q - X^q\|_q}{h^q}\right) \tag{A.4}$$

with support on $[0,1]^q$ relative to a measure on $\Xi^{[q]}$.

The moments are expressed via sums of Lebesgue integrals on the support $S^{q(\xi)}$, that involve probability measures of some sets, $\widetilde{C}$, $P_X(\widetilde{C})$, where $\widetilde{C}$ are either small cubes or finite unions of small cubes inside the cube $C(x,h)$:

$$\mathcal{I}_\xi\left(x; S^{q(\xi)}, P_X(\widetilde{C}), \Delta_{a,\xi}\psi\right) = \int_{S^{q(\xi)}} P_X(\widetilde{C}) \Delta_{a,\xi}\psi(v)\, dv(\xi) \tag{A.5}$$

with $dv(\xi) = \prod_{j\in I_\xi} dv^j$.

Lemma A.1 *Suppose that the function $\psi(w)$ is sufficiently differentiable with support on $[-1,1]^q$ and P_X is given by a distribution function $F_X(.)$ on R^q. Then with $P_X\left(\widetilde{C}\right)$ expressed as $F_X(x-h, x+\lambda_\xi \circ h)$ where λ_ξ is a vector with components:*

$$\{\lambda_\xi\}^j = v^j, \text{ if } j \in I_\xi, \text{ otherwise } \{\lambda_\xi\}^j = 1, \tag{A.6}$$

the expectation is

$$\begin{aligned} & E\left(\psi\left(\frac{x-X}{h}\right)\right) \\ = & \sum_{\xi=0}^{2^q-1} \mathcal{I}_\xi\left(x;[-1,1]^{q(\xi)}, F_X(x-h, x+\lambda_\xi \circ h), \Delta_{-1,\xi}\psi\right) \end{aligned} \tag{A.7}$$

Corollary A.1 *If in addition to the conditions of Lemma A.1 the function $\psi(\cdot)$ is symmetric around zero in every argument, then for $\xi \in I_\xi$ the corresponding $\widetilde{C} = C(x, \lambda_\xi \circ h)$ and*

$$\begin{aligned} & E\left(\psi\left(\frac{x-X}{h}\right)\right) \\ = & \sum_{\xi=0}^{2^q-1} \mathcal{I}_\xi\left(x;[0,1]^{q(\xi)}, P_X\left(C(x, \lambda_\xi \circ h)\right), \Delta_{1,\xi}\psi\right). \end{aligned} \tag{A.8}$$

We see that symmetry significantly simplifies the moment expression, in particular in this case (A.5) involves a measure of a small cube in each integral.

The lemma below applies to a functional or metric product space $\Xi^{[q]}$.

Lemma A.2 *Suppose that the function $\psi(W) = \psi_+\left(\frac{\|x^1-X^1\|_1}{h^1}, ..., \frac{\|x^q-X^q\|_q}{h^q}\right)$ where $\psi_+(z)$ defined on $[0,1]^q$ is sufficiently differentiable and P_X is a probability measure defined on $\Xi^{[q]}$ and $\widetilde{C} = C(x, \lambda_\xi \circ h)$ with λ_ξ defined in (A.6) and*

$$\begin{aligned} & E\left(\psi_+\left(\frac{\left\|x^1-X^1\right\|_1}{h^1}, ..., \frac{\|x^q-X^q\|_q}{h^q}\right)\right) \\ = & \sum_{\xi=0}^{2^q-1} \mathcal{I}_\xi\left(x;[0,1]^{q(\xi)}, P_X\left(C(x, \lambda_\xi \circ h)\right), \Delta_{1,\xi}\psi\right). \end{aligned} \tag{A.9}$$

A simplified expression for functions that take zero value on the boundary is provided in the next corollary. We say that a function $\psi(.)$ in the Lemmas A.1, A.2 and Corollary A.1 is zero on the boundary if for any w with at least one w^j with $\left|w^j\right| = 1$ the value $\psi(w)$ is zero. In particular, the expression is applicable to commonly used kernels, such as product Epanechnikov or quartic kernels; in addition to symmetry these kernels take a zero value on the boundary. It this case $I_{\xi^c} = \emptyset$ and the only non-zero term in the sums in (A.7-A.9) corresponds to $\xi = 2^q - 1$.

Corollary A.2 *If $\psi(w)$ satisfies the Corollary A.1 to Lemma A.1 or satisfies Lemma A.2 and is zero at the boundary, then*

$$E\left(\psi\left(\frac{x-X}{h}\right)\right) = \int_{[0,1]^q} P_X(C(x, h \circ v))(-1)^q \partial\psi(v)\, dv \tag{A.10}$$

The lemmas (and corollaries) provide expressions for moments for the functions arising in the local constant and local polynomial kernel estimators. Most importantly, they can be used to establish bounds on moments in terms of small cube probability. The lemma below gives a general expression for the upper bound which depends on h via the small cube probability.

Lemma A.3 *Under Corollary A.1 of Lemma A.1 or Lemma A.2*

$$\left| E\left(\psi\left(\frac{x-X}{h} \right) \right) \right| \leq P_X(C(x,h)) M_{E\psi}(x),$$

where

$$M_{E\psi}(x) = 2^q \max_{0 \leq \xi \leq 2^q - 1} \int_{[0,1]^{q(\xi)}} |\Delta_{1,\xi} \psi(v)| \, dv(\xi)$$

The expressions simplify for functions that take zero value on the boundary.

In part (a) of the next lemma we provide a condition on the probability measure that ensures the existence of a positive lower bound for such functions; the bound is expressed via the small cube probability. For functions in the Lemmas A.1, A.2 and Corollary A.1, that are not zero on the boundary, no restrictions on the probability measure are required; lower bounds in (b) do not require any additional conditions on the probability measure. Kernels that are not zero at 1 (in the univariate case) are called type I kernels, an example is the uniform kernel. Because of the fact that no restriction is needed to establish the lower bound as proportional to the small cube probability, Ferraty et al. (2007) and Hong and Linton (2020) (in the univariate case) advocate the use type I kernels. However, results for kernels that are zero on the boundary are of interest because such kernels are often used in the literature.

Lemma A.4 *If $\psi(w)$ is non-increasing for $w > 0$ and $\psi(w)$ satisfies the Corollary A.1 to Lemma A.1 or satisfies Lemma A.2*
(a) If ψ is zero at the boundary and the measure satisfies

$$\frac{P_X(C(x,h))}{P_X(C(x,\varepsilon h))} < C_F < \infty,$$

for some $0 < \varepsilon < 1$ then there is the lower bound:

$$E\left(\psi\left(\frac{x-X}{h} \right) \right) \geq P_X(C(x,h)) L_{E\psi}$$

where

$$L_{E\psi} = \frac{1}{C_F} \int_{[\varepsilon,1]^q} (-1)^q \partial\psi(v) \, dv > 0.$$

(b) If ψ satisfies

$$\psi(1, ..., 1) > 0,$$

then there is the lower bound:

$$E\left(\psi\left(\frac{x-X}{h}\right)\right) \geq P_X(C(x,h))L_{E\psi}$$

where

$$L_{E\psi} = \psi(1,...,1) > 0$$

Next, consider moments for a product of the local random function $\psi(w)$ with some bounded continuous function $g : \Xi \to R$. Consider here the sets $\widetilde{C}$ that are either small cubes or unions of small cubes in Ξ. Define for a bounded continuous function $g(x)$

$$\Omega_g(\widetilde{C}) = \int_{\widetilde{C}} g(z)dP_X(z) \tag{A.11}$$

For a small cube the function $\Omega_g(C(x,h))$ could take negative values, so considered as a function of the cube it defines a signed measure. Boundedness and continuity of the function g implies that the measure is absolutely continuous with respect to P_X.

The next lemma gives the expression for the moment of the product and bounds on the moment. The upper bound for this moment is expressed as a multiple of the small cube probability. The lower bound can be defined similarly, and need not be positive.

Lemma A.5 *(a) Under the conditions of Corollary A.1 of Lemma A.1 or Lemma A.2 for $\psi(.)$, for a bounded continuous function $g(x)$ the moment*

$$E\left[g(x)\left(\psi\left(\frac{x-X}{h}\right)\right)\right] = \sum_{\xi=0}^{2^q-1} \mathcal{I}_\xi(x;[0,1]^{q(\xi)}, \Omega_g(C(x,\lambda_\xi \circ h)), \Delta_{1,\xi}\psi)$$

with $\Omega_g(C(x,\lambda_\xi \circ h))$ given by (A.11) with $\widetilde{C} = C(x,\lambda_\xi \circ h)$
(b) The moment is bounded

$$\left|E\left[g(x)\left(\psi\left(\frac{x-X}{h}\right)\right)\right]\right| \leq M_{Eg\psi}P_X(C(x,h)),$$

with

$$M_{Eg\psi} = \sup|g(x)| M_{E\psi};$$

(c) Under the conditions of (a) or (b) of Lemma A.4

$$\left|E\left[g(x)\left(\psi\left(\frac{x-X}{h}\right)\right)\right]\right| \geq L_{Eg\psi}P_X(C(x,h)),$$

with

$$L_{Eg\psi} \geq \max\{0, (\inf g(x)) L_{E\psi}\}$$

where for (a)

$$L_{E\psi} = \frac{1}{C_F} \int_{[\varepsilon,1]^q} (-1)^q \partial\psi(v)\, dv > 0;$$

under (b)

$$L_{E\psi} = \psi(1, ..., 1) > 0$$

Since the function $\psi\left(\frac{x-X}{h}\right)$ is non-zero only over a small neighbourhood of x the $\sup|g(x)|$ as well as $(\inf g(x))$ could be taken over a $h-$neighbourhood of x. When $\inf g(x) > 0$, the lower bound is positive.

B Proofs of general technical results

Proof of Lemma A.1
The function $\psi(w^1, \cdots, w^q)$ can be written as

$$\psi(w^1, w^2, \cdots, w^q) = \psi(-1, -1, \cdots, -1) + \sum_{i=1}^{q} \int_{-1}^{w^i} \partial_{v^i}\psi(-1, -1, \cdots, v^i, \cdots, -1)dv^i \tag{B.1}$$

$$+ \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \int_{-1}^{w^i} \int_{-1}^{w^j} \partial_{v^i}\partial_{v^j}\psi(-1, \cdots, v^i, \cdots, -1, \cdots, v^j, \cdots, -1)dv^i dv^j$$

$$+ \cdots$$

$$+ \int_{-1}^{w^1} \cdots \int_{-1}^{w^q} \partial_{v^1}\partial_{v^2}\cdots\partial_{v^q}\psi(v^1, v^2, \cdots, v^q)dv^1 dv^2 \cdots dv^q$$

$$= \psi(-1, -1, \cdots, -1) + \sum_{\xi=1}^{2^q-1} \left\{ \int_{S_v^{q(\xi)}} \left(\left[\prod_{j \in I_\xi^c} \delta\left(v^j = -1\right) \prod_{j \in I_\xi} \partial_j \right] \psi\left(v^1, ...v^q\right) \right) \prod_{j \in I_\xi} dv^j \right\}$$

$$= \sum_{\xi=0}^{2^q-1} \left\{ \int_{S_v^{q(\xi)}} (-1)^{q(\xi)} \Delta_{-1,\xi}\psi\left(v^1, ...v^q\right) dv(\xi) \right\} \tag{B.2}$$

where $S_v^{q(\xi)} = \prod_{j \in I_\xi} \left[-1, w^j\right]$, $\Delta_{-1,\xi}$ is defined in A.3, and $dv(\xi) = \prod_{j \in I_\xi} dv^j$.

To each term of (B.2) with $v^j = \frac{x^j - t^j}{h^j}$, apply a change of variables: $t^j = x^j - v^j h^j$ for all $j \in I_\xi$. Here

$$dv(\xi) := \prod_{j \in I_\xi} dv^j = \prod_{j \in I_\xi} \left(-h^j\right)^{-1} dt^j =: (-1)^{q(\xi)} \prod_{j \in I_\xi} \left(h^j\right)^{-1} dt(\xi).$$

The limits of integrals change with $-1 \to x^j + h^j; w^j \equiv \frac{x^j - X^j}{h^j} \to X^j$. Since the function $\psi(w) = \psi(\frac{x-X}{h})$ is zero outside of the set

$$I(x - h \leq X \leq x + h) = \prod_{i=1}^{q} I(x^i - h^i \leq X^i \leq x^i + h^i),$$

$X^j \leq x^j + h^j$, and we apply a reversal in the limits of integral (hence an additional $(-1)^{q(\xi)}$) to obtain

$$\begin{aligned}
&\psi\left(\tfrac{x^1-X^1}{h^1}, \tfrac{x^2-X^2}{h^2}, \cdots, \tfrac{x^q-X^q}{h^q}\right) \\
= &\sum_{\xi=0}^{2^q-1}\left\{(-1)^{q(\xi)}\int_{S_t^{q(\xi)}}\left[\Delta_{-1,\xi}\psi\left(\frac{x^1-t^1}{h^1},...\frac{x^q-t^q}{h^q}\right)\right]\prod_{j\in I_\xi}(h^j)^{-1}dt(\xi)\right\} I(x-h\leq X\leq x-h),
\end{aligned}$$

where $S_t^{q(\xi)} = \prod_{j\in I_\xi}[X^j, x^j + h^j]$.

The expectation is

$$\begin{aligned}
&E\left(\psi(\frac{x-X}{h})\right) \\
= &\sum_{\xi=0}^{2^q-1}(-1)^{q(\xi)}\int_{R^q}\left\{\int_{S_t^{q(\xi)}}\Delta_{-1,\xi}\psi\left(\frac{x^1-t^1}{h^1},...\frac{x^q-t^q}{h^q}\right)\prod_{j\in I_\xi}(h^j)^{-1}dt(\xi)\right\} I(x-h\leq X\leq x-h)dP_X(X) \\
= &\sum_{\xi=0}^{2^q-1}(-1)^{q(\xi)}\int_{\widetilde{S}_t^{q(\xi)}}\left(\Delta_{-1,\xi}\psi\left(\frac{x^1-t^1}{h^1},...\frac{x^q-t^q}{h^q}\right)\right)\left\{\int_{S_X^q(t(\xi))}dF_X(X)\right\}\prod_{j\in I_\xi}(h^j)^{-1}dt(\xi) \qquad \text{(B.3)}
\end{aligned}$$

where the first expression uses linearity of the expectation operator and the second uses the Fubini theorem and recognizes that the domain of integration in the curly brackets depends on $t^j,\ j\in I_\xi$. In particular as for every $j\in I_\xi$

$$x^j - h^j \leq X^j \leq t^j$$

and $t^j \in [X^j, x^j + h^j]$, $S_X^q(t(\xi)) = \prod_{j\in I_\xi}[x^j - h^j, t^j]\prod_{j\in I_\xi^c}\left[x^j - h^j, x^j + h^j\right]$ (incorporating the requirement $I(x-h\leq X\leq x+h)$). This provides

$$\int_{S_X^q(t(\xi))}dF_X(X) = F_X(x-h, x+\lambda_\xi(t)\circ h).$$

where

$$\{\lambda_\xi(t)\}^j = \frac{t^j - x^j}{h^j},\ \text{if } j\in I_\xi,\ \text{otherwise } \{\lambda_\xi(t)\}^j = 1.$$

For the limits of the integral with respect to $t(\xi)$ we use $\widetilde{S}^{q(\xi)} = \prod_{j\in I_\xi}[x^j - h^j, x^j + h^j]$.

The last displayed expression (B.3) then becomes

$$\sum_{\xi=0}^{2^q-1}(-1)^{q(\xi)}\int_{\widetilde{S}_t^{q(\xi)}}\left(\Delta_{-1,\xi}\psi\left(\frac{x^1-t^1}{h^1},...\frac{x^q-t^q}{h^q}\right)\right)F_X(x-h, x+\lambda_\xi(t)\circ h)\prod_{j\in I_\xi}\left(h^j\right)^{-1}dt(\xi)$$

After applying a change of variables with $t^j = x^j - v^j h^j$, this yields

$$\sum_{\xi=0}^{2^q-1}(-1)^{q(\xi)}\int_{[-1,1]^{q(\xi)}}\left(\Delta_{-1,\xi}\psi\left(v^1,...,v^q\right)\right)F_X(x-h, x+\lambda_\xi\circ h)dv(\xi)$$

where

$$\{\lambda_\xi\}^j = v^j, \text{ if } j \in I_\xi, \text{ otherwise } \{\lambda_\xi\}^j = 1$$

The change of variables uses $dv(\xi) = \left(h^j\right)^{-1} dt(\xi)$ together with a reversal in the limits of integration as before. Therefore,

$$\begin{aligned}
& E\left(\psi\left(\frac{x-X}{h}\right)\right) \\
= & \sum_{\xi=0}^{2^q-1} \mathcal{I}_\xi(x;[-1,1]^{q(\xi)}, F_X(x-h, x+\lambda_\xi \circ h), \Delta_{-1,\xi}\psi) \\
= & \sum_{\xi=0}^{2^q-1} \int_{[-1,1]^{q(\xi)}} F_X(x-h, x+\lambda_\xi \circ h) \prod_{j\in I_\xi^c} \delta\left(v^j = -1\right) \prod_{j\in I_\xi} (-\partial_j)\psi(v)\, dv(\xi).
\end{aligned}$$

■

Proof of Corollary A.1
In the expression obtained in Lemma A.1 substituting symmetric $\psi(.)$ results in the operator $\delta\left(v^j = -1\right)$ providing the same result as $\delta\left(v^j = 1\right)$; we have

$$\sum_{\xi=0}^{2^q-1} \int_{[-1,1]^{q(\xi)}} F_X(x-h, x+\lambda_\xi \circ h) \prod_{j\in I_\xi^c} \left(\delta\left(v^j = 1\right)\right) \prod_{j\in I_\xi} (-\partial_j)\psi(v)\, dv(\xi),$$

where $dv(\xi) = \prod_{j\in I_\xi} dv^j$ and recall

$$\{\lambda_\xi\}^j = v^j, \text{ if } j \in I_\xi, \text{ otherwise } \{\lambda_\xi(v)\}^j = 1.$$

Without loss of generality let

$$dv(\xi) = dv^{j_1} dv^{j_2} \cdots dv^{j_{q(\xi)}}.$$

Next, we integrate with respect to v^{j_1} (meanwhile holding $v^{j_2}, \cdots, v^{j_{q(\xi)}}$ constant)

$$\begin{aligned}
& \int_{-1}^{1} F_X(x-h, x+\lambda_\xi \circ h) \prod_{j\in I_\xi^c} \delta\left(v^j = 1\right) \prod_{j\in I_\xi} (-\partial_j)\psi(v)\, dv^{j_1} \\
= & \int_{-1}^{0} F_X(x-h, x+\lambda_\xi \circ h) \prod_{j\in I_\xi^c} \delta\left(v^j = 1\right) \prod_{j\in I_\xi} (-\partial_j)\psi(v)\, dv^{j_1} \\
& + \int_{0}^{1} F_X(x-h, x+\lambda_\xi \circ h) \prod_{j\in I_\xi^c} \delta\left(v^j = 1\right) \prod_{j\in I_\xi} (-\partial_j)\psi(v)\, dv^{j_1}. \quad \text{(B.4)}
\end{aligned}$$

In the first integral, on the rhs of (B.4) apply a change of variable $v^{j_1} = -z^{j_1}$; define $\lambda_\xi(j_1)$ the same as λ_ξ for every component, except for j_1^{th} where it is

$-z^{j_1}$; let $v\left(-z^{j_1}\right)$ represent v with v^{j_1} replaced with $-z^{j_1}$, then this term can be written as

$$-\int_0^1 F_X(x-h, x+\lambda_\xi(j_1)\circ h)\prod_{j\in I_\xi^c}\left(\delta\left(v^j=1\right)\right)\prod_{j\in I_\xi}\left(-\partial_j\right)\psi\left(v\left(-z^{j_1}\right)\right)(-1)\,dz^{j_1}.$$

where the minus from interchanging the limits of integration and the minus arising from the change of variables cancel out. By symmetry of $\psi(.)$, $\left(-\partial_{j_1}\right)\psi\left(v\left(-z^{j_1}\right)\right) = -\left(-\partial_{j_1}\right)\psi\left(v\left(z^{j_1}\right)\right)$, where $v\left(z^{j_1}\right)$ represent v with v^{j_1} replaced with z^{j_1}. The first integral then becomes

$$-\int_0^1 F_X(x-h, x+\lambda_\xi(j_1)\circ h)\prod_{j\in I_\xi^c}\left(\delta\left(v^j=1\right)\right)\prod_{j\in I_\xi}\left(-\partial_j\right)\psi\left(v\left(z^{j_1}\right)\right)dz^{j_1}.$$

By changing the notation in the second integral on the rhs of (B.4) only, writing z^{j_1} in place of v^{j_1} to express it as $v\left(z^{j_1}\right)$, the sum of the two integrals in (B.4) becomes

$$\int_0^1 \left\{F_X(x-h, x+\lambda_\xi\circ h)-F_X(x-h, x+\lambda_\xi(j_1)\circ h)\right\}\prod_{j\in I_\xi^c}\left(\delta\left(v^j=1\right)\right)\prod_{j\in I_\xi}\left(-\partial_j\right)\psi\left(v\left(z^{j_1}\right)\right)dz^{j_1}.$$

Next, consider the integral with respect to v^{j_2} (meanwhile holding $v^{j_3},\cdots,v^{j_{q(\xi)}}$ constant). That is, we evaluate

$$\int_{-1}^1\left\{\int_0^1\left\{F_X(x-h, x+\lambda_\xi\circ h)-F_X(x-h, x+\lambda_\xi(j_1)\circ h)\right\}\prod_{j\in I_\xi^c}\left(\delta\left(v^j=1\right)\right)\right.$$
$$\left.\prod_{j\in I_\xi}\left(-\partial_j\right)\psi\left(v\left(z^{j_1}\right)\right)dz^{j_1}\right\}dv^{j_2}.$$

A similar substitution $z^{j_2}=-v^{j_2}$, with $v\left(z^{j_1}, z^{j_2}\right)$, $\lambda_\xi(j_2)$ and $\lambda_\xi(j_1, j_2)$ similarly defined provides the integral as

$$\int_0^1\int_0^1\left\{F_X(x-h, x+\lambda_\xi\circ h)-F_X(x-h, x+\lambda_\xi(j_2)\circ h)\right.$$
$$\left.+F_X(x-h, x+\lambda_\xi(j_1, j_2)\circ h)-F_X(x-h, x+\lambda_\xi(j_1)\circ h)\right\}$$
$$\times\prod_{j\in I_\xi^c}\delta\left(v^j=1\right)\prod_{j\in I_\xi}\left(-\partial_j\right)\psi\left(v\left(z^{j_1}, z^{j_2}\right)\right)dz^{j_1}dz^{j_2}.$$

Continuing this until $z^{j_{q(\xi)}}$ yields

$$\int_0^1 \cdots \int_0^1 \left\{ \underbrace{\int_{x_1-h_1}^{x_1+h_1} \int_{x_2-h_2}^{x_2+h_2} \cdots}_{j\in I_\xi^c} \underbrace{\int_{x_{j_1}-z_{j_1}h_{j_1}}^{x_{j_1}+z_{j_1}h_{z_{j_1}}} \cdots \int_{x_{j_{q(\xi)}}-z_{j_{q(\xi)}}h_{j_{q(\xi)}}}^{x_{j_{q(\xi)}}+z_{j_{q(\xi)}}h_{z_{j_{q(\xi)}}}}}_{j\in I_\xi} dF_X \right\}$$

$$\times \prod_{j\in I_\xi^c} \delta\left(v^j=1\right) \prod_{j\in I_\xi} (-\partial_j)\,\psi\left(v(z^{j_1}, z^{j_2}, \cdots, z^{j_{q(\xi)}})\right) dz(\xi)$$

where $dz(\xi) = dz^{j_1} dz^{j_2} \cdots dz^{j_{q(\xi)}}$. Simply changing the notation, with $dv(\xi) = dv^{j_1} dv^{j_2} \cdots dv^{j_{q(\xi)}}$ and $v(v^{j_1}, v^{j_2}, \cdots, v^{j_{q(\xi)}}) = v$, yields

$$\int_0^1 \cdots \int_0^1 \underbrace{\int_{x_1-h_1}^{x_1+h_1} \int_{x_2-h_2}^{x_2+h_2} \cdots}_{j\in I_\xi^c} \underbrace{\int_{x_{j_1}-v_{j_1}h_{j_1}}^{x_{j_1}+v_{j_1}h_{v_{j_1}}} \cdots \int_{x_{j_{q(\xi)}}-v_{j_{q(\xi)}}h_{j_{q(\xi)}}}^{x_{j_{q(\xi)}}+v_{j_{q(\xi)}}h_{v_{j_{q(\xi)}}}}}_{j\in I_\xi} dF_X$$

$$\times \prod_{j\in I_\xi^c} \delta\left(v^j=1\right) \prod_{j\in I_\xi} (-\partial_j)\,\psi(v)\,dv(\xi)$$

$$= \int_{[0,1]^{q(\xi)}} F_X(x-\lambda_\xi\circ h, x+\lambda_\xi\circ h) \prod_{j\in I_\xi^c} \delta\left(v^j=1\right) \prod_{j\in I_\xi} (-\partial_j)\,\psi(v)\,dv(\xi)$$

$$= \int_{[0,1]^{q(\xi)}} P_X(C(x,\lambda_\xi\circ h) \prod_{j\in I_\xi^c} \delta\left(v^j=1\right) \prod_{j\in I_\xi} (-\partial_j)\,\psi(v)\,dv(\xi)$$

$$= \int_{[0,1]^{q(\xi)}} P_X(C(x,\lambda_\xi\circ h)\Delta_{1,\xi}\psi(v)\,dv(\xi)$$

This concludes the proof.

■

Proof of Lemma A.2

The probability measure P_X on $\Xi = \Xi^{[q]}$ given $x \in \Xi$ defines a distribution F_{Z_+} in R^q with support on R_+^q, a non-negative multivariate quadrant, given by the measurable mapping $\Xi \to R_+^q$ with $X = \left(X^1, ..., X^q\right)$ mapped into a random vector

$$Z_+ = \left(Z_+^1, ..., Z_+^q\right);\ Z_+^i = \left\|x^i - X^i\right\|_i.$$

Then $X = x$ transforms into $Z_+ = X - x = 0$. If $r \in R^q$ has non-negative components, the F_Z measure of the cube $C(0,r)$ is concentrated in the non-negative quadrant and is given by

$$F_{Z_+}(C(0,r)) = P_X(C(x,r)) = P_X(\prod_{i=1}^q B(\left\|x^i - X^i\right\|_i \le r^i)_i).$$

We can consider on R^q the symmetric function $\psi(.)$ with support in $[-1,1]^q$, based on the given $\psi_+(.)$, defined as $\psi\left(v^1, ..., v^q\right) = \psi_+\left(\left\|v^1\right\|, ..., \|v^q\|\right)$. Thus

the symmetric function $\psi\left(\frac{x-X}{h}\right)$ of Corollary A.1 can be written as $=\psi\left(\frac{Z}{h}\right)$ and its values are given by $\psi\left(\frac{Z}{h}\right)=\psi_{+}\left(\frac{Z_{+}}{h}\right)$. The result from Corollary A.1 then written with $x=0$ and the probability measure corresponding to the distribution $F_{Z_{+}}$

$$P_{Z_{+}}\left(C(0,\lambda_{\xi}\circ h)\right)=F_{Z_{+}}\left(C(0,\lambda_{\xi}\circ h)\right)$$

becomes

$$E\left(\psi\left(\frac{Z}{h}\right)\right)=\sum_{\xi=0}^{2^{q}-1}\mathcal{I}_{\xi}(0;[0,1]^{q(\xi)},F_{Z_{+}}\left(C(0,\lambda_{\xi}\circ h)\right),\Delta_{1,\xi}\psi).$$

Recognizing that $E_{Z_{+}}\left(\psi_{+}\left(\frac{Z_{+}}{h}\right)\right)=E_{Z_{+}}\left(\psi\left(\frac{Z_{+}}{h}\right)\right)$ and transforming back to the original probability measure we get

$$E\left(\psi\left(\frac{\left\|x^{1}-X^{1}\right\|_{1}}{h^{1}},...,\frac{\left\|x^{q}-X^{q}\right\|_{q}}{h^{q}}\right)\right)=\sum_{\xi=0}^{2^{q}-1}\mathcal{I}_{\xi}(x;[0,1]^{q(\xi)},P_{X}\left(C(x,\lambda_{\xi}\circ h)\right),\Delta_{1,\xi}\psi).$$

■

Proof of Corollary A.3.
(A.10) arises immediately from the sums in the expressions for the expectation since any term ξ for which $I_{\xi}^{c}\neq\varnothing$ is zero. ■

Proof of Lemma A.3
Since the components of $\lambda_{\xi}\circ h$ are between zero and components of h,

$$P_{X}\left(C(x,\lambda_{\xi}\circ h)\right)\leq P_{X}\left(C(x,h)\right).$$

We detail the bound for the expression of Lemma A.1; similar derivations provide it under Lemma A.2. Consider the expression

$$\sum_{\xi=0}^{2^{q}-1}\int_{[-1,1]^{q(\xi)}}F_{X}(x-h,x+\lambda_{\xi}\circ h)\prod_{j\in I_{\xi}^{c}}\left(\delta\left(v^{j}=-1\right)\right)\prod_{j\in I_{\xi}}\left(-\partial_{j}\right)\psi\left(v\right)dv\left(\xi\right);$$

this can be bounded by

$$\begin{aligned}&\sum_{\xi=0}^{2^{q}-1}\int_{[-1,1]^{q(\xi)}}F_{X}(x-h,x+h)\left|\prod_{j\in I_{\xi}^{c}}\left(\delta\left(v^{j}=-1\right)\right)\prod_{j\in I_{\xi}}\left(-\partial_{j}\right)\psi\left(v\right)\right|dv\left(\xi\right)\\ \leq\;& F_{X}(x-h,x+h)2^{q}\max_{0\leq\xi\leq2^{q}-1}\int_{[-1,1]^{q(\xi)}}\left|\prod_{j\in I_{\xi}^{c}}\left(\delta\left(v^{j}=-1\right)\right)\prod_{j\in I_{\xi}}\left(-\partial_{j}\right)\psi\left(v\right)\right|dv\left(\xi\right).\end{aligned}$$

Recall that $F_X(x-h, x+h) = P_X(C(x,h))$. Thus under Lemma A.1 or Lemma A.2

$$\left| E\left(\psi\left(\frac{x-X}{h}\right)\right)\right| \leq P_X(C(x,h)) M_{E\psi}(x),$$

where

$$M_{E\psi}(x) = 2^q \max_{0\leq\xi\leq 2^q-1} \int_{S^{q(s)}} |\Delta_{a,\xi}\psi(v)|\, dv(\xi).$$

■

Proof of Lemma A.4. (a) The lower follows after substituting the lower bound, $C_F^{-1} P_X(C(x,h))$ for $P_X(C(x,\varepsilon h))$ in (A.10).
(b) Highlights the term $\psi(1,..,1) > 0$ in the expression for the expectation which appears when $\xi = 0$. As all other terms are non-negative, this term is sufficient for the lower bound. ■

Proof of Lemma A.5. (a) The derivation is identical to that in Lemma A.1 with the only difference that the dF_X or dP_X is replaced by $d\Omega_X$ in all the derivations providing the result.
(b) The upper bound follows from the boundedness of the function $g(x)$.
(c) The lower bound in the case of positive $g(x)$ follows; when $g(x)$ can take non-positive values, the zero lower bound on the absolute value of the moment holds.

■

C Proofs of lemmas and theorems

The proofs of the lemmas and theorems in the text heavily rely on the results in Appendix A.

Proof of Lemma 1.
(a) Given (18) with $\tau_0(\varepsilon) = C(\varepsilon) > 0$ we have that for small enough h

$$P_X(B(x,\varepsilon h)) > \tilde{C} P_X(B(x,h)),$$

where $\tilde{C} = C(\varepsilon) + \delta$, $\delta > 0$ and (16) holds.

(b) Since $P_X(B(x,u))$ is non-decreasing and continuous in u

$$\int_0^v P_X(B(x,u))\, du = P_X(B(x,\tilde{s}v))\, v,$$

where $\tilde{s} < 1$. From (19) it follows that $P_X(B(x,\tilde{s}v))\, v > CvP_X(B(x,v))$, thus for $\varepsilon = \tilde{s}$ (16) holds. ■

A useful consequence of Assumption 4 that will be helpful to bound ratios for the general case is provided in the following auxiliary Lemma.

Lemma C.1 *Under Assumption 4 for any* $0 < \varepsilon < 1$

$$\frac{P_X\left(C(x,h)\right)}{nP_X\left(C(x,\varepsilon h)\right)^2} \to 0.$$

Proof of Lemma C.1.
First using (16) followed by (15)

$$\begin{aligned} & \frac{P_X\left(C(x,h)\right)}{nP_X\left(C(x,\varepsilon h)\right)^2} \\ \leq \quad & C_F^2 \frac{P_X\left(C(x,h)\right)}{nP_X\left(C(x,h)\right)^2} \to 0. \end{aligned}$$

■

The next lemma and its corollary provide bounds used in the proofs below, which implement the results in Appendix A to our kernel function: that is $\psi(\cdot) = K(\cdot)$.

Lemma C.2 *Given Assumption 2 and a kernel* K *satisfying 1(a-c) the moment satisfies*

$$L_{EK^m} P_X\left(C(x,h)\right) \leq E\left(K^m\left(\frac{x-X}{h}\right)\right) \leq M_{EK^m} P_X\left(C(x,h)\right).$$

(a) Under Assumption 4(a,b) the bounds are

$$L_{EK^m} = \frac{1}{C_F}\int_{[\varepsilon,1]^q} (-1)^q \partial K^m(v)\, dv;\ M_{EK^m} = \int_{[0,1]^q} (-1)^q \partial K^m(v)\, dv$$

for some $0 < \varepsilon < 1$.
(b) Under Assumption 4a and the addition of Assumption 1(d) on the kernel K*, the bounds on the moments are*

$$L_{EK^m} = (-1)^q K^m(1);\ M_{EK^m} = \sup_{[0,1]^q} \sum_{\xi=0}^{2^q-1} \left[\prod_{j\in I_\xi^c} \delta\left(v^j = 1\right) \prod_{j\in I_\xi} (-1)^{q(\xi)} \partial_j\right] K^m(v).$$

Lemma C.2 shows that the standard kernels, widely used in estimation in R^q, such as Epanechnikov, can provide bounded moments for the kernel function and thus prove to be useful when Assumption 4(b) that requires (16) holds. When Assumption 4(b) is not satisfied we require the additional assumption on the kernel 1(d). For odd m, $\partial K^m(v) < 0$ for any $v \geq 0$, thus the lower bound L_{EK^m} is always positive. This is important since such functions appear in the denominator of the estimator.

Proof of Lemma C.2.
The bounds are obtained by substituting $K^m(.)$ for $\psi(.)$ in Lemma A.3 and

using the fact that $\left[\prod_{j\in I_\xi^c}\delta\left(v_j=1\right)\prod_{j\in I_\xi}(-1)\partial_j\right]K^m\left(v\right)$ is non-negative for odd m on $[0,1]^q$. ■

Recall, for some bounded continuous function $g\left(x\right)$, we defined

$$\Omega_g(C(x,h)) = \int\limits_{C(x,h)} g(z)dF_X(z) \tag{C.1}$$

where $C(x,h)$ is the small cube.

Corollary C.1 *Under the conditions of Lemma C.2 for a bounded function $g\left(x\right)$ the moment*

$$E\left[g\left(X\right)\left(K\left(\frac{x-X}{h}\right)\right)^m\right] = \int_{[0,1]^q}\Omega_g\left(C\left(x,h\circ v\right)\right)(-1)^q\partial K^m\left(v\right)dv$$

with $\Omega_g\left(C\left(x,h\circ v\right)\right)$ given by (A.11) is bounded as

$$L_{EgK^m}P_X\left(C\left(x,h\right)\right) \leq \left|E\left[g\left(X\right)\left(K\left(\frac{x-X}{h}\right)\right)^m\right]\right| \leq M_{EgK^m}P_X\left(C\left(x,h\right)\right),$$

where the bounds are

$$L_{EK^mg} = \inf g\left(x\right)L_{EK^m};\ M_{EK^m} = \sup\left|g\left(x\right)M_{EK^m}\right|;$$

the upper and lower bounds on $g\left(x\right)$ can be taken over the $h-$neighbourhood of x.

The corollary follows directly from Lemma **C.2.** Note that here the lower bound does not have to be positive.

Proof of Theorem 1.
The proof proceeds by first examining the denominator of the estimator, then establishing the asymptotic normality for the numerator and finally combining. We first formulate the needed results in the forms of two Lemmas and a Proposition and then provide the proofs.

The moments for $B_n\left(x\right)$ are expressed via moments of the kernel function:

$$\begin{aligned} EB_n(x) &= EK\left(W_i(x)\right); \\ EB_n^2(x) &= \frac{1}{n}EK\left(W_i(x)\right)^2 + \frac{n-1}{n^2}\left(EK\left(W_i(x)\right)\right)^2. \end{aligned}$$

The following Lemma makes it possible to replace $B_n\left(x\right)$ by $EB_n\left(x\right)$ and $\widehat{m}\left(x\right)$ by $\widetilde{m}\left(x\right) = \frac{A_n(x)}{EB_n(x)}$ when examining the limit properties of the estimator.

Lemma C.3 *Under either of the following sets of assumptions (i) Assumptions 1 (a-d), 2, and 4(a), or (ii) Assumptions 1 (a-c), 2, and 4(a,b)*
(a)

$$\frac{B_n(x)-EB_n(x)}{EB_n(x)} = o_p\left(1\right)$$

(b)

$$\widehat{m}(x) = \widetilde{m}(x)(1 + o_p(1))$$

where $\widetilde{m}(x) = \dfrac{A_n(x)}{EB_n(x)}$.

Next, we consider the numerator of $\widehat{m}(x)$:

$$A_n(x) = \frac{1}{n}\sum_{i=1}^{n} \zeta_{in} \text{ with } \zeta_{in} = K(W_i(x))Y_i.$$

Each term, ζ_{in}, is a function of x, but to simplify we suppress in notation this dependence. The next lemma expresses the moments for the numerator of the estimator at x.

Applying the definition in (C.1), we define $\Omega_m(x-h, x+h)$ for $g(x) = m(x)$ and $\Omega_{m^2+\mu_2}(x-h, x+h)$ for $g(x) = m(x)^2 + \mu_2(x)$.

Lemma C.4 *Under the conditions of Lemma C.3 and Assumption 3(a)*

(a)

$$E\zeta_{in} = \int_{[0,1]^q} \Omega_m(C(x, h \circ v))(-1)^q \partial K(v)\, dv;$$

(b)

$$E\zeta_{in}^2 = \int_{[0,1]^q} \Omega_{m^2+\mu_2}(C(x, h \circ v))(-1)^q \partial K^2(v)\, dv.$$

Using Lemma C.4 we can write the expressions for $var\zeta_{in} = E\zeta_{in}^2 - (E\zeta_{in})^2$ and establish the moments for the numerator of $\widehat{m}(x)$ as

$$EA_n(x) = E\zeta_{in};\ varA_n(x) = \frac{1}{n}\left(E\zeta_{in}^2 - (E\zeta_{in})^2\right)$$

by substituting the formulae from the Lemma.

The moments involve the measures $\Omega_g(C(x, h \circ v))$ for the (bounded) functions $m(x), m(x)^2 + \mu_2(x)$. The lower and upper bounds are expressed via the corresponding measures $P_X(C(x,h))$; details are in Corollary C.1 of Appendix C. Hence $varA_n(x)$ as $n \to \infty$, $h \to 0$ possesses the same rate as $n^{-1}P_X(C(x,h))$. Note that if x is a mass point, the parametric rate holds.

Next, define

$$\widetilde{\zeta}_{in} = \frac{1}{\sqrt{n}} \frac{\zeta_{in} - E\zeta_{in}}{\sqrt{var\zeta_{in}}}.$$

Proposition C.1 *Under the conditions of Lemma C.4 as* $n \to \infty$, $h \to 0$

$$\sum_{i=1}^{n} \widetilde{\zeta}_{in} \to_d Z \sim N(0,1).$$

The asymptotic normality result provides asymptotic normality for the numerator, $A_n(x)$. To establish asymptotic normality of the ratio, $\frac{A_n(x)}{B_n(x)} = \widehat{m}(x)$ first recall that by Lemma C.3 the asymptotic distribution for $\widehat{m}(x)$ is the same as for $\frac{A_n(x)}{EB_n(x)}$. The asymptotic variance is then provided by $\lim_{n\to\infty} \frac{varA_n(x)}{(EB_n(x))^2}$. Noting that $\hat{m}(x) - m(x) = \hat{m}(x) - E(\hat{m}(x)) + bias(\hat{m}(x))$ we obtain the result of Theorem 1.

Next, we provide the proofs of the Lemmas and Proposition.

Proof of Lemma C.3.
(a) Using the result from Lemma C.2 applied with $m - 1, 2$ we obtain

$$\begin{aligned} & var\left(\frac{B_n(x) - EB_n(x)}{EB_n(x)}\right) \\ = & \frac{\frac{1}{n}EK^2 - \frac{1}{n}(EK)^2}{(EK)^2} \\ \leq & \frac{M_{EK^2}}{n(L_{EK}^2) P_X(C(x,h))}. \end{aligned}$$

Hence, we have that

$$\frac{B_n(x) - EB_n(x)}{EB_n(x)} = O_p\left(\frac{1}{nP_X(C(x,h))}\right)$$

and thus is $o_p(1)$ by Assumption 4(a).
(b) We have that

$$\begin{aligned} \widehat{m}(x) &= \frac{A_n(x)}{EB_n(x) + (B_n(x) - EB_n(x))} \\ &= \frac{A_n(x)}{EB_n(x)}\left(1 + \frac{1}{EB_n(x)}(B_n(x) - EB_n(x))\right)^{-1} \\ &= \widetilde{m}(x)(1 + o_p(1)), \end{aligned}$$

where the last inequality is obtained by using $\frac{B_n(x)-EB_n(x)}{EB_n(x)} = o_p(1)$ that permits expansion of $\left(1 + \frac{1}{EB_n(x)}(B_n(x) - EB_n(x))\right)^{-1}$ into a geometric progression with the common ratio of $o_p(1)$. ■

Proof of Lemma C.4.

(a) Using the expressions from Corollary C.1

$$\begin{aligned} E\zeta_{in} &= E\left(E\zeta_{in}|X_i\right) && \text{(C.2)} \\ &= E\left[K\left(\frac{x-X_i}{h}\right)\left(m\left(X_i\right)+E(u_i|X_i)\right)\right] \\ &= \int K\left(\frac{x-X}{h}\right) m\left(X\right) dF_X && \text{(C.3)} \\ &= \int_{[0,1]^q} \Omega_m(C\left(x, h\circ v\right))\partial K\left(-v\right) dv, \end{aligned}$$

where to obtain the expression we use $E(u|X)=0$ and the definition of $\Omega_m\left(.,.\right)$.

(b) Similarly to (a)

$$\begin{aligned} E\zeta_{in}^2 &= E\left(E\left[K\left(\frac{x-X_i}{h}\right)^2\left(m\left(X_i\right)+u_i\right)^2|X_i\right]\right) && \text{(C.4)} \\ &= \int K\left(\frac{x-X}{h}\right)^2\left(m\left(X\right)^2+\mu_2\left(X\right)\right) dF_X \\ &= \int_{[0,1]^q} \Omega_{m^2+\mu_2}(C\left(x, h\circ v\right))\partial K^2\left(-v\right) dv \\ &\quad + \int_{[0,1]^q} \Omega_{m^2+\mu_2}(C\left(x, h\circ v\right))\partial K^2\left(-v\right) dv \\ &= \int_{[0,1]^q} \Omega_{m^2+\mu_2}(C\left(x, h\circ v\right))\partial K^2\left(-v\right) dv. \end{aligned}$$

■

Proof of Proposition C.1.
We verify the conditions of the Lyapunov CLT. The first three conditions: (i) $E\widetilde{\zeta}_{in}=0$; (ii) $\sum_{i=1}^n var\left(\widetilde{\zeta}_{in}\right)=1$; (iii) $\widetilde{\zeta}_{in}$ and $\widetilde{\zeta}_{jn}$ are independent, hold. All that remains is to verify the Lyapunov condition $\sum_{i=1}^n E\left|\widetilde{\zeta}_{in}\right|^{2+\delta}\to 0$; we do it for the fourth moments $(\delta=2)$.

We have

$$
\begin{aligned}
& E\left(\zeta_{in}-E\zeta_{in}\right)^4 \qquad\qquad (C.5)\\
= \;& E\zeta_{in}^4-4E\zeta_{in}^3E\zeta_{in}+6E\zeta_{in}^2\left(E\zeta_{in}\right)^2-4E\zeta_{in}\left(E\zeta_{in}\right)^3+\left(E\zeta_{in}\right)^4\\
= \;& \int \partial K^4\left(-v\right)\Omega_4(C\left(x,h\circ v\right))dv\\
& -4\int \partial K^3\left(-v\right)\Omega_3(C\left(x,h\circ v\right))dv\int \partial K\left(-v\right)\Omega_3(C\left(x,h\circ v\right))dv\\
& +6\left[\int \partial K^2\left(-v\right)\Omega_2(C\left(x,h\circ v\right))dv\right]^2\\
& -3\left[\int \partial K\left(-v\right)\Omega_1(C\left(x,h\circ v\right))dv\right]^4
\end{aligned}
$$

where $\Omega_r=E(Y^r|X=x)=E((m(X)+u)^r|X=x)$ for $r=1,\cdots,4$. We note $\Omega_1\equiv\Omega_m$, for $g_1=m$, whereas with reference to (A.11), for Ω_2 the $g_2=m^2+\mu_2$, for Ω_3 the $g_3=m^3+3m\mu_2+\mu_3$, and for Ω_4 the $g_4=m^4+6m^2\mu_2+4\mu_3+\mu_4$, with $\mu_3=E(u^3|X=x)$.

By definition of $\widetilde{\zeta}_{in}$, then

$$
\sum_{i=1}^{n}E\left|\widetilde{\zeta}_{in}\right|^4=\frac{1}{n}\frac{E\left(\zeta_{in}-E\zeta_{in}\right)^4}{\left(var\zeta_{in}\right)^2}
$$

Applying the bounds from Lemma A.5 (a) to each factor of each term in $(C.5)$ we express the upper bound on the numerator.

$$
E\left(\zeta_{in}-E\zeta_{in}\right)^4\leq M_1P_X\left(C\left(x,h\right)\right)+M_2P_X\left(C\left(x,h\right)\right)^2+M_3P_X\left(C\left(x,h\right)\right)^4,
$$

where $M_i,i=1,2,3$ combines the upper bounds M_{EgK^i} with appropriately defined g functions for each term in $(C.5)$. This bound can be expressed as $M_{num}P_X\left(C\left(x,h\right)\right)$.

Correspondingly, using $(C.4),(C.2)$ we can bound from below the denominator

$$
var\zeta_{in}=E\zeta_{in}^2-\left(E\zeta_{in}\right)^2. \qquad (C.6)
$$

Indeed, from Lemma A.5 (b)

$$
E\zeta_{in}^2\geq L_{E(m^2+\mu_2)K^2}P_X\left(C\left(x,h\right)\right);
$$

and from (a)

$$
\begin{aligned}
\left(E\zeta_{in}\right)^2 &\leq \left[M_{EmK}P_X\left(C\left(x,h\right)\right)\right]^2\\
&\leq M_{EmK}^2P_X\left(C\left(x,h\right)\right)^2.
\end{aligned}
$$

Then $var\zeta_{in}$ is bounded below by $L_{den}P_X\left(C\left(x,h\right)\right)$. When the measure of the small ball goes to zero at x (as when the distributon function on R^q is continuous at x), then for small enough h we have that $L_{den}>0$.

Then

$$\frac{1}{n}\frac{E\left(\zeta_{in}-E\zeta_{in}\right)^{4}}{\left(var\zeta_{in}\right)^{2}}\leq\frac{M_{num}}{nL_{den}}$$

and

$$\frac{1}{n}\frac{E\left(\zeta_{in}-E\zeta_{in}\right)^{4}}{\left(var\zeta_{in}\right)^{2}}=o\left(1\right).$$

Thus the Lyapunov condition is verified for the case that the small ball probability goes to zero. With a mass, α_x, at x consider $P_XC\left(x,h\right))=P_X\left(X=x\right)+$

$P_{\tilde{X}}C\left(x,h\right))$ where $\tilde{X}=\begin{cases} X & \text{for } X\neq x \\ 0 & \text{for } X=x \end{cases}$. Then for the measure $P_X\left(X=x\right)$ a standard $\sqrt{n}$ asymptotic normality holds. ■

Proof of Theorem 2
The result follows on combining Theorem 1 with evaluation of the bias. To evaluate the bias of $\widehat{m}\left(x\right)$, $E\left(\frac{A_n(x)}{B_n(x)}\right)-m\left(x\right)$. In the product space $\Xi=\Xi^{[q]}$ the following lemma applies:

Lemma C.5 *Under the conditions of Lemma C.4 the bias of* $\widehat{m}\left(x\right)$ *in* $\Xi^{[q]}$ *is* $O\left(\bar{h}P_X\left(x,h\right)\right)$.

Proof of Lemma C.5.
We make use of the following expansion of $\frac{1}{B}$ around $\frac{1}{EB}$:

$$\frac{1}{B}=\frac{1}{EB}-\frac{1}{\left(EB\right)^{2}}\left(B-EB\right)+O\left(\frac{\left(B-EB\right)}{EB}\right)^{2}.$$

To evaluate the bias $E\widehat{m}\left(x\right)-m\left(x\right)\equiv E\frac{A_n(x)}{B_n(x)}-m\left(x\right)$, we substitute the expansion of $\frac{1}{B_n(x)}$ around $\frac{1}{EB_n(x)}$, yielding

$$\text{Bias}(\widehat{m}(x))=E\frac{A_n-m\left(x\right)B_n}{B_n-EB_n+EB_n}=E\frac{EA_n-m\left(x\right)B_n}{EB_n}\left(1-\frac{\left(B_n(x)-EB_n(x)\right)}{EB_n(x)}+...\right),$$

thus the leading term is $\frac{EA_n-m(x)EB_n}{EB_n}$ since $var\left(\frac{B_n(x)-EB_n(x)}{EB_n(x)}\right)=o\left(1\right)$ by Lemma 2.

Then consider $EA_n-m\left(x\right)EB_n=E\zeta_{in}-m\left(x\right)EB_n=$

$$\int K\left(\frac{x-X}{h}\right)\left(m\left(X\right)-m\left(x\right)\right)dF_X.$$

Then by Assumption 3(a)

$$\left|E\left(K\left(\frac{x-X}{h}\right)\left(m\left(X\right)-m\left(x\right)\right)\right)\right|\leq O\left(\bar{h}\right)\left|EK\left(\frac{x-X}{h}\right)\right|.$$

This is then $O\left(\bar{h}P_X\left(C\left(x,h\right)\right)\right)$. The result follows. ■

Proof of Theorem 3
The result follows by substituting the limits in Assumption 5 into the result of Theorem 2. ■

Proof of Theorem 4
The proof relies on the following Lemma that provides in (a) the bounds on the matrix components, in (b) and (c) justifies using the estimator that replaces the original random matrix $\hat{B}\left(x\right)$ by its expectaion and in (d) applies the central limit theory to obtain asymptotic normality.

Lemma C.6 *Under the conditions of Lemma C.2 and Assumptions 3(b) and 6 (a) the components are bounded*

$$\begin{aligned}
&E\left\{\widetilde{B}_n(x)\right\}_{lm} < P_X\left(C\left(x,h\right)\right)\bar{M}_{EB};\\
&E\left\{\widetilde{B}_n(x)\right\}_{lm}\left\{\widetilde{B}_n(x)\right\}_{l'm'} < P_X\left(C\left(x,h\right)\right)\bar{M}_{EB^2};\\
&\left|\left\{var\widetilde{A}_n\left(x\right)\right\}_{lm}\right| < n^{-1}P_X\left(C\left(x,h\right)\right)\bar{M}_{EAA'};
\end{aligned}$$

(b)

$$\left[E\widetilde{B}_n(x)\right]^{-1}\left[\widetilde{B}_n(x)-E\widetilde{B}_n(x)\right]=o_p\left(1\right)$$

(c) as $n\to\infty$

$$D\left(\hat{\beta}\left(x\right)\right)=D\left(\tilde{\beta}\left(x\right)\right)\left(1+o_p\left(1\right)\right)$$

where $D\left(\tilde{\beta}\left(x\right)\right)=\left[E\widetilde{B}(x)\right]^{-1}\widetilde{A}\left(x\right);$
(d)

$$E\widetilde{B}_n(x)\left(var\widetilde{A}_n\left(x\right)\right)^{-1/2}D\left(\hat{\beta}\left(x\right)-\beta\left(x\right)-bias(\hat{\beta}\left(x\right))\right)\to_d Z\sim N\left(0,I_{q+1}\right).$$

To conclude the proof all that is needed is to establish the rate for the bias and the limit expressions.

Proof of Lemma C.6

(a) The moments of the scalar matrix elements of $\tilde{B}_n\left(x\right)$ and their products involve moments of the functions $K\left(\frac{x-X}{h}\right), K\left(\frac{x-X}{h}\right)\left(x^i-X^i\right); K\left(\frac{x-X}{h}\right)\left(x^i-X^i\right)\left(x^j-X^j\right)$ and products of such functions. To be completed.

References

[1] Ahlfors, Lars (1966): "Lectures on Quasiconformal Mappings," Princeton University Press.

[2] Ang, A. and D. Kristensen (2012) "Testing conditional factor models," *Journal of Financial Economics*, **106**, 132-156.

[3] Arulampalam, W., V. Corradi and D. Gutknecht (2017) "Modeling heaped duration data: An application to neonatal mortality," *Journal of Econometrics*, **200**, pp. 363-377.

[4] Baele, L. and J.M. Londono (2013) "Understanding industries betas," *Journal of Empirical Finance*, **22**, 30–51.

[5] Barrientos-Marin, J. , Ferraty, F. and Vieu, P.(2010) "Locally modelled regression and functional data", *Journal of Nonparametric Statistics*, 22: 5, 617 — 632

[6] Calonico, S., M.D. Cattaneo, and M.H. Farrell (2018) "On the effect of bias estimation on coverage accuracy in nonparametric inference," *JASA*,**113**, 767-779.

[7] Chagny and Roche (2016)"Adaptive estimation in the functional nonparametric regression model," *Journal of Multivariate Analysis*, **146**, 105–118.

[8] Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. (2014) "Anti-Concentration and honest, adaptive confidence bands," The Annals of Statistics 42, no. 5, 1787–1818. http://www.jstor.org/stable/43556344.

[9] Compiani, G. and Y. Kitamura (2016) "Using mixtures in econometric models: A brief review and some new results," *Econometrics Journal*, **19**, pp. C95-C127.

[10] Desmet, K. and S.L. Parente (2010) "Bigger is better: Market size, demand elasticity, and innovation," *International Economic Review*, **51**, pp. 319-333.

[11] Donkers, A.C.. and M. Schafgans (2008) "Estimation and specification of semiparametric index models," *Econometric Theory*, **24**, 15844–1606.

[12] Fan, J. and I. Gijbels (1992) *Local Polynomial Modelling and Its Applications,* Chapman and Hall, London.

[13] Febrero-Bande M. and M.O. de la Fuente (2012) "Statistical computing in functional data analysis: The R package fda.usc," *Journal of Statistical Software*, **51**, Issue 4.

[14] Ferraty F. and P. Vieu (2006) "Nonparametric functional data analysis: Theory and Practice,"Springer, New York.

[15] Ferraty F., A. Mas, and P. Vieu (2007) "Nonparametric regression on functional data: Inference and practical aspects,"*Australian & New Zealand Journal of Statistics*, **49**, pp. 267–286.

[16] Ferraty F., Nagy S. 2022, "Scalar-on-function local linear regression and beyond," *Biometrika*, 109, 2, pp. 439–455

[17] Geenens, G. (2015) "Moments, errors, asymptotic normality and large deviation principle in nonparametric functional regression," *Statistical Probability Letters*, **20**, pp. 3-18.

[18] Gyorfi, L., M. Kohler, A. Krzyzak and H. Walk (2002) "A Distribution-Free Theory of Nonparametric Regression,"Springer.

[19] Hall, P. and J. Horowitz (2013) "A simple bootstrap method for constructing nonparametric confidence bands for functions," *Annals of Statistics*, **41**, pp. 1892-1921.

[20] Hardle, W., P. Hall and H. Ichimura (1993) "Optimal smoothing in single-index models," *The Annals of Statistics*, **21**, pp. 157-178.

[21] Hardle, W. and Marron (1991) "Bootstrap simultaneous error bars for nonparametric regression," *The Annals of Statistics*, **19**, 778–796.

[22] Hayfield, T. and J.S. Racine (2008) "Nonparametric Econometrics: The np package," *Journal of Statistical Software*, **27**, Issue 5.

[23] Hillberry, R. and D. Hummels (2008) "Trade responses to geographic frictions: A decomposition using micro-data," *European Economic Review*, **52**, 527-550.

[24] Hong,S. and O. Linton (2020), "Nonparametric estimation of infinite order regression and its application to the risk-return tradeoff," *Journal of Econometrics*, **219**, pp. 389-424

[25] Hotelling, H. (1929) "Stability in competition," *The Economic Journal*, **39**, pp. 41-57.

[26] Huynh, K.P. and D.T. Jacho-Chavez (2009) "Growth and Governance: A nonparametric analysis," *Journal of Comparative Economics*, 131-143.

[27] Ichimura, H. (1993) "Semiparametric least squares (SLS) and weighted SLS estimation of single index models," *Journal of Econometrics*, **58**, 71–120.

[28] Kankanala, S. and V. Zinde-Walsh (2023) "Kernel-weighted specifiaion testing under general distributions," arXiv:2204.01683v2, *Bernoulli*, in print.

[29] Klein, R.W. and R.H.Spady (1993) "An efficient semiparametric estimator for binary choice models," *Econometrica*, **61**, 387-421.

[30] Kotlyarova, Y, M.M.A. Schafgans, and V. Zinde-Walsh (2016) "Smoothness: Bias and efficiency of non-parametric kernel estimators," in *Advances in Econometrics: Essays in Honor of Aman Ullah*, Vol. 36, G. Gonzales-Rivera, R.C. Hill and T.-H. Lee, eds.

[31] La Torre, D, S. Marsiglio and F. Privileggi (2011) "Fractals and self-similarity in economics: The case of a stochastic two-sector growth model," *Image Analysis and Stereology*, **30**, pp. 143-131.

[32] Lee M., Mukherjee D. and A. Ullah (2019) "Nonparametric estimation of the marginal effect in fixed-effect panel data models," *Journal of Multivariate Analysis* **171**, 53-67.

[33] Li, Q. and J.S. Racine (2007) "Nonparametric econometrics: Theory and practice," Princeton University Press.

[34] Lin, Z., H.-G. Müller, and F. Yao, "Mixture inner product spaces and their application to functional data analysis, *The Annals of Statistics,* **46**, pp. 370-400.

[35] Linton O., Z. Xiao (2019) "Efficient estimation of nonparametric regression in the presence of dynamic heteroskedasticity," *Journal of Econometrics*,**213**, pp. 608-631.

[36] Mackay, J. M. and J. T. Tyson (2010) "Conformal dimension. Theory and application" Amer. Math. Soc., Providence, RI, 2010.

[37] Mandlebrot (1968)

[38] Mastromarco, C. and L. Simar (2015) "Effect of FDI and time on catching up: New insights from a conditional nonparametric frontier analysis," *Journal of Applied Econometrics* **30**, 826–847.

[39] Nadaraya, E. (1965) "On non-parametric estimates of density functions and regression curves," *Theory of Probability & Its Applications*, **10**, 186–190.

[40] Nikol'skii, S.M. (2001) [1994], "Sobolev space," Encyclopedia of Mathematics, EMS Press.

[41] Racine, J. and Q. Li (2004) Nonparametric estimation of regression function with both categorical and continuous data, 119(1), 99–130.

[42] Severn K.E., Dryden I.L., Preston S.P. (2021) "Non-parametric regression for networks," *Stat*, 10 (1), pp. e373 1-11.

[43] Stone, C.J. (1982) "Optimal global rates of convergence for nonparametric regression," *Annals of Statistics*, **10**, 1040–1053.

[44] Stone, C.J. (1977) "Consistent nonparametric regression," *Annals of Statistics*, **5**, 595–645.

[45] Watson, G. S. (1964). "Smooth regression analysis," *Sankhya: The Indian Journal of Statistics*, Series (1961-2002), 26(4), 359– 372.

ANNEX Figures and Tables – Kernel estimation in regression on vector and function spaces

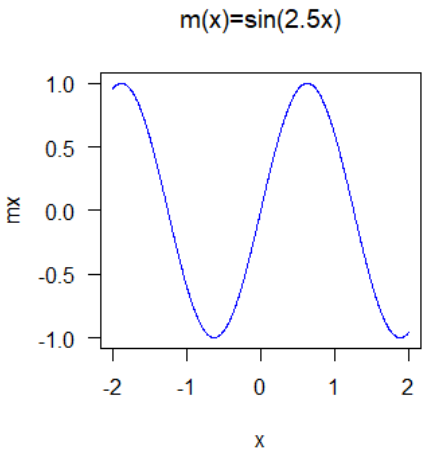


Figure 1 Univariate conditional mean

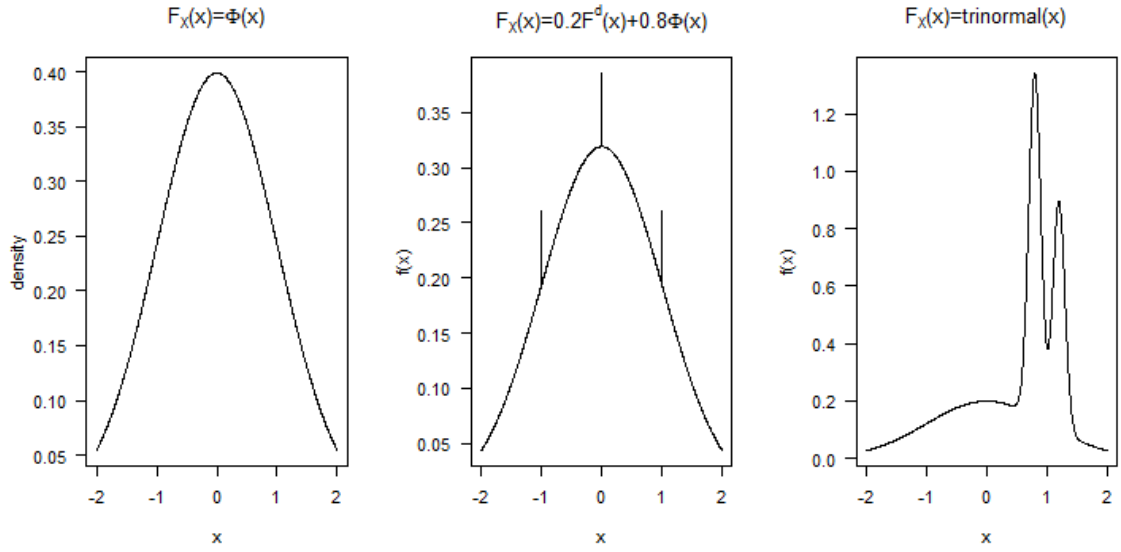


Figure 2 Distribution of conditioning variables (point mass and high derivatives)

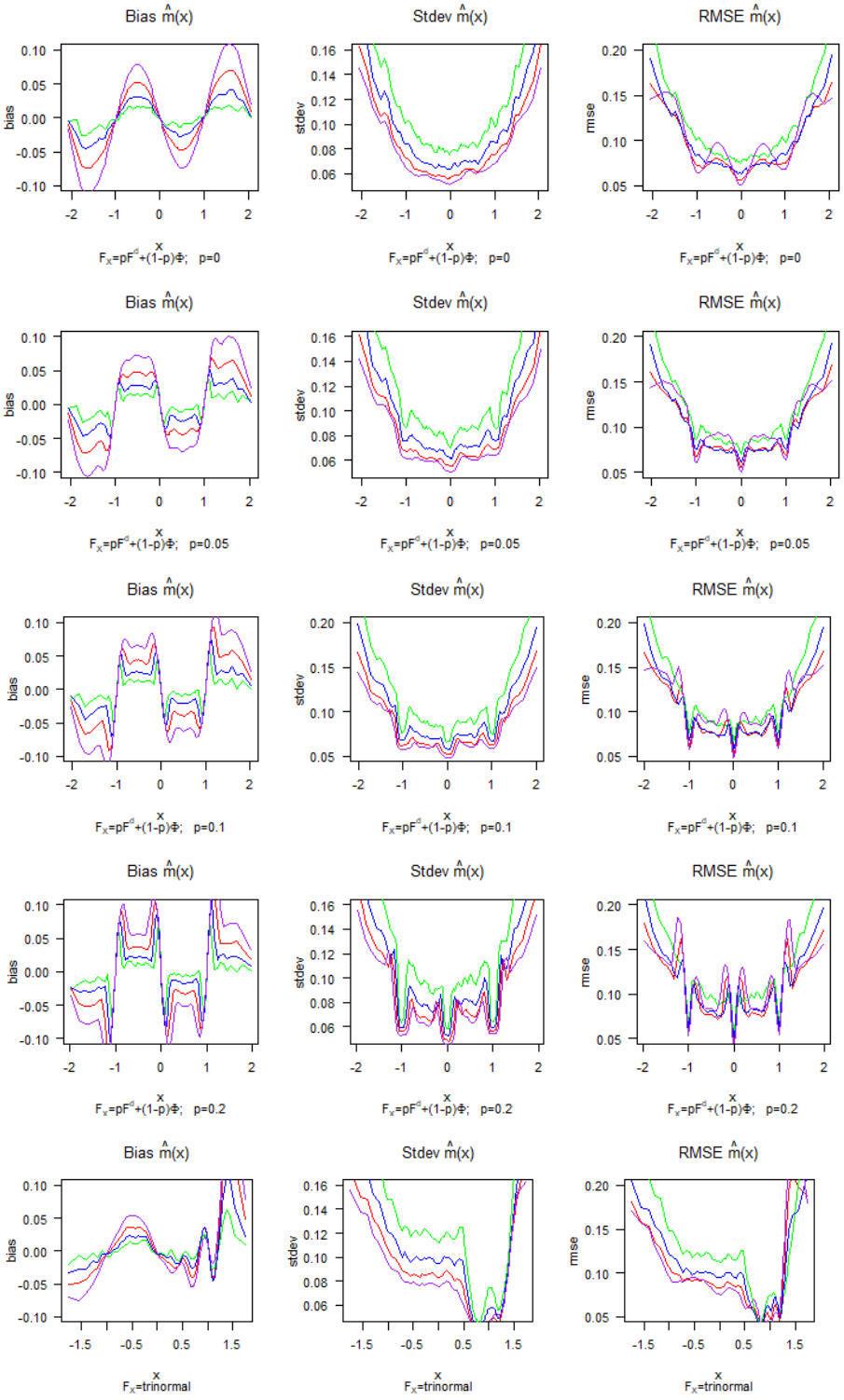
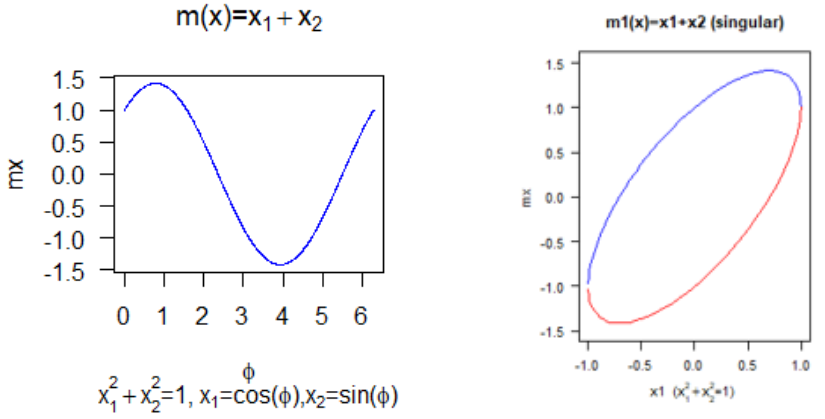


Figure 3 Bias, standard deviation and RMSE of nonparametric estimates over 500 simulations by distribution F_X and bandwidth. The top panel is the base setting where F_X is standard gaussian, in the panels 2-4 F_X has mass points (with increasing probability), and in the 5th panel F_X is the trinormal distribution. The green and blue lines presents results under undersmoothing (0.5hcv and 0.75hcv respectively), the purple line oversmoothing (1.25hcv), and the red line presents the cross-validated result.



(a) (b)

Figure 4 Bivariate conditional mean with singularity (a) m(x) with φ on the horizontal axis, (b) m(x) with x_1 on the horizontal axis; the blue line represents the setting where $x_2 \leq 0$, the red line where $x_2 \geq 0$, with $x_2=0$ the lines intersect.

Table 1 Aggregate performance of the NW local constant and single index estimator by bandwidth (as multiples of the cross-validated bandwidths)

	(NW) local constant estimator		(SI) Single Index estimator	
Multiplier	**MAE**	**MSE**	**MAE**	**MSE**
0.50	0.8044	1.0152	0.8027	1.0110
0.75	0.8021	1.0114	0.8011	1.0071
1.00	0.8011	1.0070	0.8002	1.0050
1.25	0.8019	1.0107	0.8008	1.0065

Note: The MAE is given by $\sum_{i=1}^{n}|y_i - \hat{m}(x_i)|$ and the MSE is given by $\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2$ where the leave-one-out estimator for $m(x_i)$ is used. The rows reflect the choice of different bandwidths indicated by a multiplying factor on the cross-validated bandwidth.

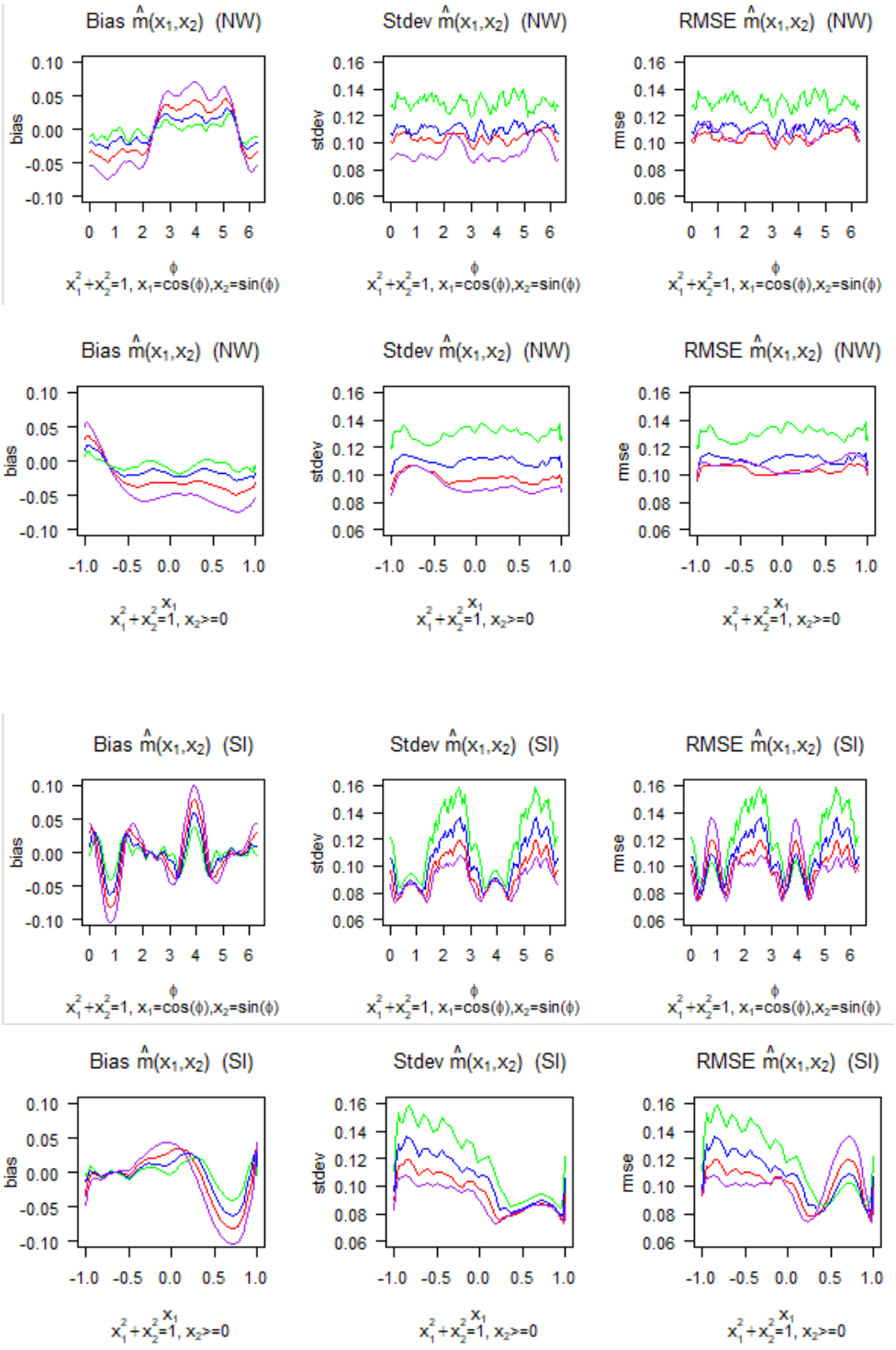


Figure 5 Bias, standard deviation and RMSE of nonparametric estimates over 500 simulations by bandwidth. The top two panels present the results for the Nadaraya-Watson (NW) local constant estimator (with φ and x_1 on the horizontal axis, respectively) and the bottom two panels present the results for the Single Index (SI) estimator. The green and blue lines presents results under undersmoothing (0.5hcv and 0.75hcv respectively), the purple line oversmoothing (1.25hcv), and the red line presents the cross-validated result.

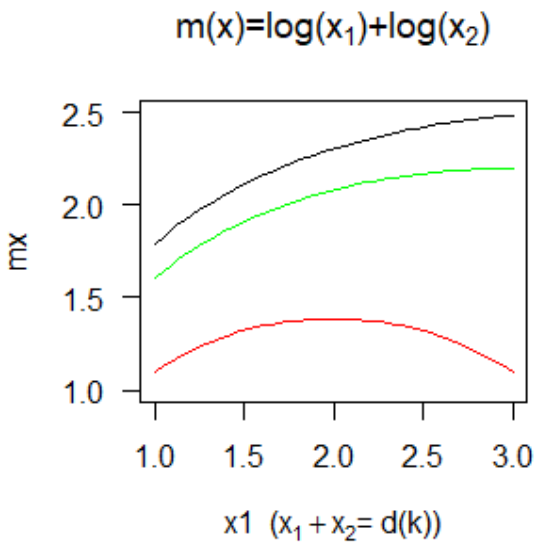


Figure 6 Bivariate model with reduced dimensionality in the conditioning variables. The red line represents d(1)=3, the green line d(2)=6 and the black line d(3)=7.

Table 2 Aggregate performance of the NW local constant and single index estimator by bandwidth (as multiples of the cross-validated bandwidths)

	(NW) local constant estimator; Based on X1 and X2		(NW) local constant estimator; Based on X1 and d(k)		(SI) Single Index estimator; Based on logX1 and logX2	
Multiplier	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
0.50	0.3226	0.1636	0.3198	0.1607	0.3213	0.1622
0.75	0.3216	0.1625	0.3196	0.1606	0.3205	0.1615
1.00	0.3211	0.1621	0.3196	0.1606	0.3201	0.1611
1.25	0.3221	0.1631	0.3202	0.1611	0.3205	0.1614

Note: The MAE is given by $\sum_{i=1}^{n}|y_i - \hat{m}(x_i)|$ and the MSE is given by $\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2$ where the leave-one-out estimator for m(x_i) is used. The rows reflect the choice of different bandwidths indicated by a multiplying factor on the cross-validated bandwidth.

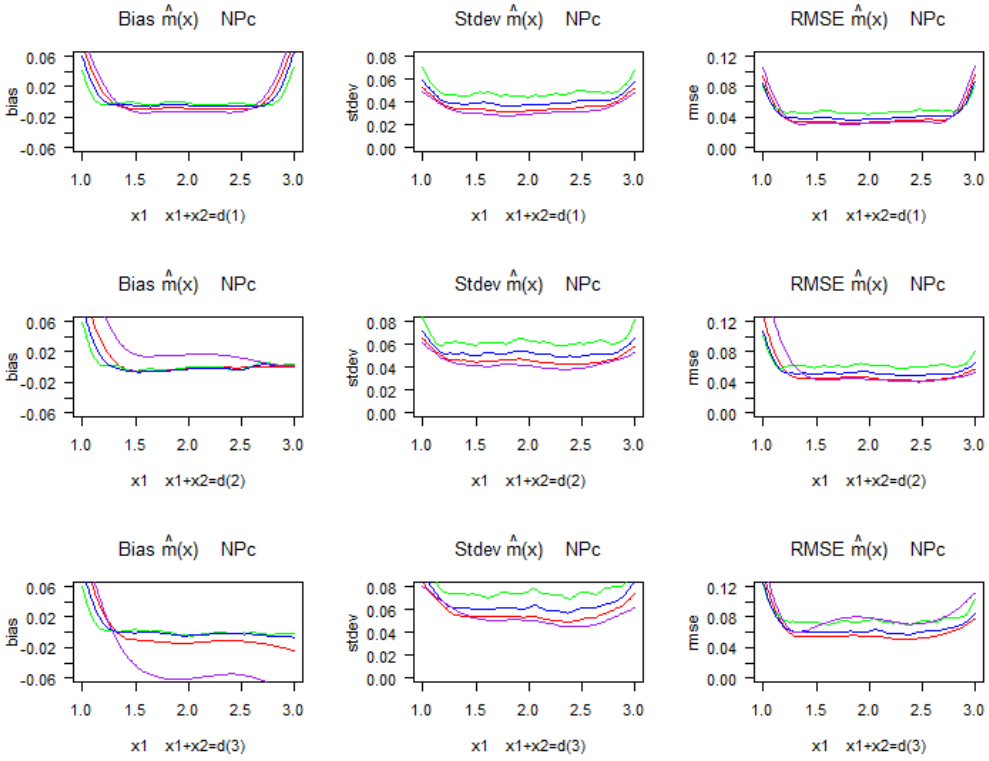


Figure 7 NW kernel regression estimator (based on X1 and X2)

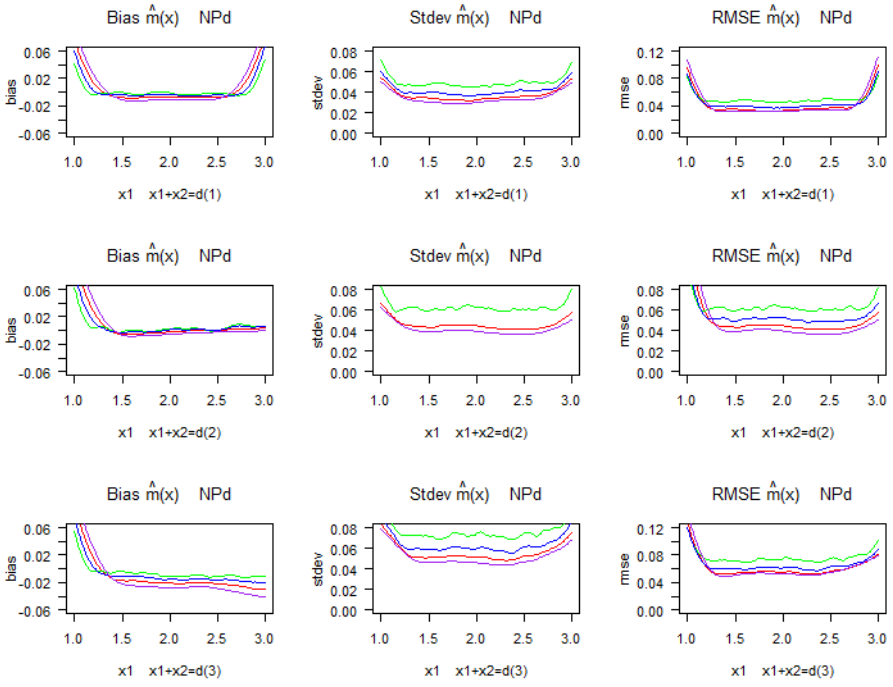


Figure 8 NW kernel regression estimator (based on X1 and d(k))

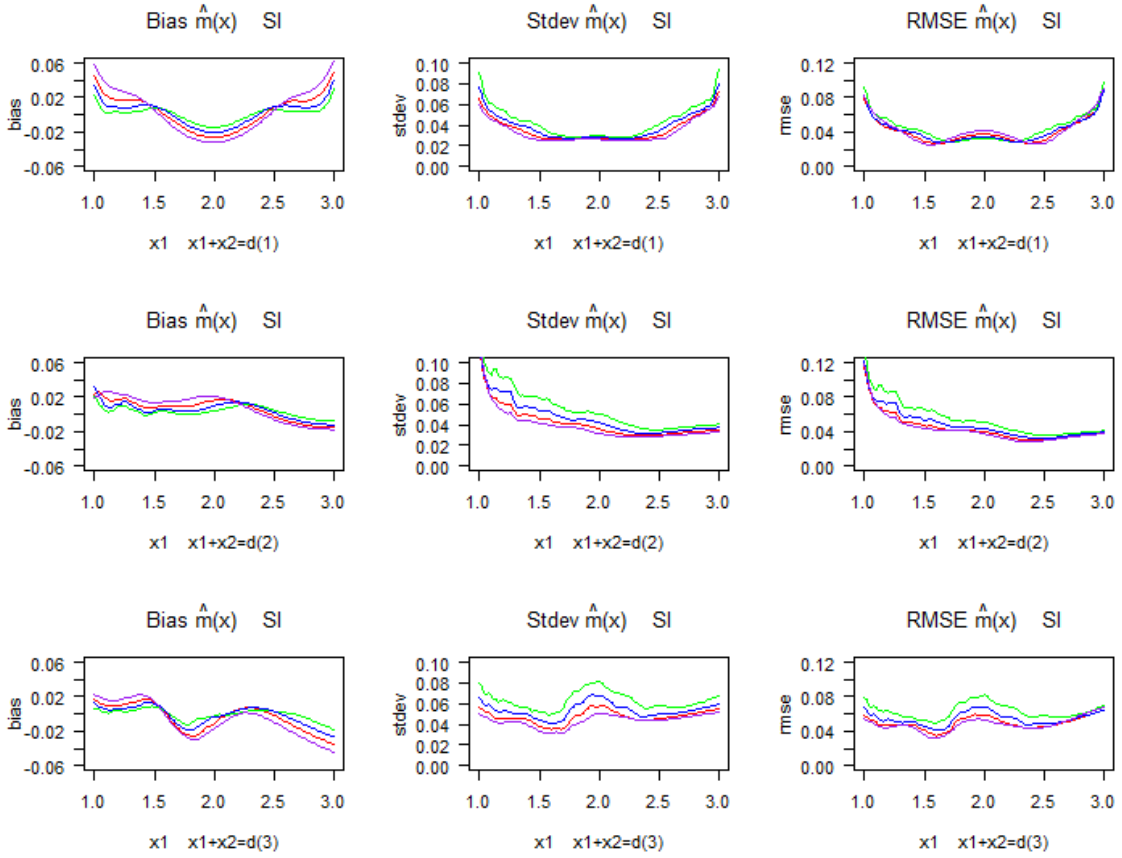


Figure 9 SI estimator (based on logX1 and logX2)