Multivariate kernel regression in vector and product metric spaces

Marcia Schafgans*
London School of Economics

Victoria Zinde-Walsh†
McGill University and CIREQ

May 3, 2024

Abstract

Nonparametric kernel regression is widely used in econometrics and has been applied to models with cross-sectional, time series and panel data. As functional data analysis is gaining attention, our analysis extends to the presence of multiple functional regressors or a mixture of functional and scalar regressors. We address point-wise asymptotic properties of the estimators by relaxing restrictive smoothness assumptions on the conditioning distribution. Firstly, we establish the rate and asymptotic normality for the Nadaraya-Watson estimator for arbitrary conditioning distributions in R^q; for an absolutely continuous distribution we provide statistical guarantees for the standard rate and the asymptotic normality without requiring further smoothness. We demonstrate faster convergence associated with dimension reducing types of singularity, such as a fractal structure or a factor structure in the regressors. Second, the paper extends the limit theory of kernel functional regression to multivariate regression over a product of any number of metric spaces. Finite sample evidence confirms rate improvement due to singularity in regression over R^q. For functional regression the simulations underline the importance of accounting for multiple functional regressors.

JEL classicification: C14

Keywords: Nadaraya-Watson estimator; singular distribution; multivariate functional regression; small cube probability.

1 Introduction

This paper is centered on the nonparametric regression model

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

* *Corresponding author*: Economics Department, London School of Economics, Houghton Street, London WC2A 2AE, UK. E-mail address: m.schafgans@lse.ac.uk

† Economics Department, McGill University, 855 Sherbrooke St. W., Montreal, Quebec H3A 2T7, Canada. E-mail address: victoria.zinde-walsh@mcgill.ca.

where the regressor X can be a vector in R^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 for it. Y represents a scalar dependent variable, u denotes an unobserved error, and the conditional mean function $m(\cdot)$ satisfies some smoothness assumptions.

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

(*i*) If $\Xi = R^q$ and $X = (X^1, \cdots, X^q)$, 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.

Our **first aim** is to extend the applicability of kernel estimation to regression on R^q, where there is possible non-smoothness of the conditioning distribution due to mass points (e.g. in neonatal mortality in Arulampalam et al., 2017, as well as in income or subjective probability data), fractal structure (as in Mandelbrot, 1997 for daily prices in cotton market among others, also well documented for networks, urban growth, financial markets) or functional relations between regressors (e.g. in production functions, Ackerberg et al., 2015), by establishing conditions for asymptotic normality for general conditioning distributions.

(*ii*) If Ξ is a general univariate metric space the regression problem is known as functional regression (see, e.g. Ramsey and Silverman, 2005). We denote this setting by $\Xi = \Xi^{[1]}$ with $q = 1$. Estimation and inference techniques for kernel functional regression have been developed by Ferraty and Vieu (2004) and have been further explored in a growing body of literature (see Ferraty et al., 2006 and Masry, 2005); local linear kernel estimator has been considered here as well (e.g., Ferraty and Nagy, 2022, Hong and Linton, 2020).

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 a univariate functional regression. This is due to the fact that the vector X may have heterogeneous components, such as monetary and demographic variables, which require a multivariate kernel estimator with possibly different bandwidths for different components. This is commonplace in the literature, see for instance Huynh et al. (2009). Hence, case (i) cannot be reduced to case (ii).

(*iii*) If $X = (X^1, \cdots, X^q)$ is comprised of q heterogeneous components, with each X^l defined on a univariate Banach or metric measure space $\Xi_l^{[1]}$, a multivariate functional regression is called for. We assume that X lies in a product probability measure space $\Xi = \Xi^{[q]} \equiv \Xi_1^{[1]} \times \cdots \times \Xi_q^{[1]}$. The

metrics on $\Xi_l^{[1]}$, $\|.\|_l$, may differ for each of the q components of function spaces.

Our **second aim** is to provide asymptotic normality results to enable multiple functional regression on objects with components in heterogeneous metric spaces. Kurisu et al. (2023) also make a case for combining a random vector and a function to obtain an estimate for the propensity score used in estimating the average treatment effect. Our general framework makes it possible to examine multivariate models with multiple functional and usual predictors. This provides a basis for simultaneously evaluating the impact of the different predictors rather than comparing their performance in distinct models, as in Caldeira et al. (2020) and Ferraty and Nagy (2022).[1]

In the literature on asymptotic normality for kernel regression with $\Xi = R^q$ density is assumed to exist, while density does not exist in general metric measure spaces. In the univariate metric space, $\Xi = \Xi^{[1]}$, the probability measure is characterized by the small ball probability (e.g., see Ferraty and Vieu, 2006): for the ball $B(x,h) = \{X : \|x - X\| \leq h\}$ in $\Xi^{[1]}$. The probability measure is denoted $P_X(B(x,h))$. Characterizing the measure locally via a ball is insufficient when we wish to examine heterogeneous regressors in R^q without requiring the existence of density or, in general, in product metric measure spaces $\Xi^{[q]}$.

Kankanala and Zinde-Walsh (2023) introduced small cube probability for a cube[2] $C(x,h) = \left\{X \in R^q : \left|X^l - x^l\right| \leq h^l, l = 1, \cdots, q\right\}$ in R^q. With distribution function of X given by F_X the corresponding probability measure is given by

$$P_X(C(x,h)) = \int_{C(x,h)} dF_X. \tag{2}$$

Here, we use the small cube probability to extend the regression on univariate metric measure spaces to a product of different such spaces where the probability measure is defined (see, e.g. Pollard, 2001, Ch. 4 Section 3). We consider a small cube around $x = \left(x^1, \cdots, x^q\right)$ where for each component $B^l\left(x^l, h^l\right) = \left\{X^l : \left\|x^l - X^l\right\|_l \leq h^l\right\}$ denotes the small ball. For a vector $h = \left(h^1, \cdots, h^q\right)$ the cube is

$$C(x,h) = \left\{X : \left\|X^l - x^l\right\|_l \leq h^l\right\} = \left\{X : X^l \in B^l\left(x^l, h^l\right)\right\}. \tag{3}$$

and its 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 kernel regression estimator, can

[1] E.g., Caldeira et al. compares the functional model forecasting aggregate stock market excess return based on its history with regression models based on traditional predictors, while Ferrary and Nagy compare the performance two distinct functional regressors (growth velocity profiles from ages 1-10 to 5-8) for predicting adult height.

[2] 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. There appears to be no established terminology for a product of balls in products of metric spaces.

be categorized based on assumptions made about the space Ξ, the small cube probability $P_X(C(x,h))$, and the kernel function $K(\cdot)$. Below, we outline the main results on pointwise asymptotic normality of the regression estimator.

A For $\Xi = R^q$, assuming a sufficiently smooth distribution of the regressors 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., see Racine and Li, 2007). This paper establishes the asymptotic normality of the kernel estimator for $\Xi = R^q$ for a general distribution function F_X. For kernels that are non-zero on their support boundary, such as the uniform, F_X could be arbitrary, while for kernels that are zero on their support boundary, such as the Epanechnikov, the probability measure corresponding to F_X is restricted to be doubling[3] to obtain asymptotic normality. 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 in R^q.

B The pointwise asymptotic normality for a Banach or metric space $\Xi = \Xi^{[1]}$ was established by Ferraty and Vieu (2006), Ferraty et al. (2007) and Geenens (2015) in the i.i.d. case. Masry (2005) and Delsol (2009) derived the limit distribution for a strongly mixing process. Ferraty and Nagy (2022) provided results for a local linear estimator in the Hilbert space Ξ. Hong and Linton (2020) derived limit results for the local linear estimator for a stationary as well as dynamic model with strong mixing. For the general case where Ξ represents a finite product of normed or metric spaces, i.e., $\Xi = \Xi^{[q]}$, we derive here the limit Gaussian distribution for kernel regression with corresponding multivariate kernel for strongly mixing stationary processes.

The derivations in the technical appendix of this paper provide moments and bounds for general multivariate local functions; these bounds generalize the results in the literature (e.g. in Ferraty and Vieu, 2006) and subsequent papers on univariate functional regression. These moment results are applicable to other contexts, such as local linear and local polynomial estimation in R^q or in products of suitable metric spaces, kernel estimation of distribution functions and conditional distributions in R^q as well as to kernel regression of objects in metric spaces on objects in products of spaces.

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 in R^q. Moreover, our simulation results highlight the behavior of kernel estimators for regression models with dependence on both a functional object in Ξ and a random variable.

[3] A doubling measure (e.g. Vol'berg, Konyagin, 1988) restricts how the small cube shrinks as $h \to 0$. Many well-known probability measures in R^q are doubling: any measure for absolutely continuous distribution, any fractal measure, mixtures of fractals are doubling.

The theoretical results of this paper make it possible to interpret numerous applications of kernel estimators on R^q in the literature more widely. Currently, the literature that makes use of kernel estimation, e.g. the applications reviewed in Li and Racine (2007) as well as more recent ones such as those in Lee et al. (2019) and Linton and Xiao (2019), rely on asymptotic results with assumptions on smoothness of density; as is shown here these assumptions could be relaxed for general a.c. distributions without affecting the limit rate and asymptotic normality or the computation of the estimators. Additionally, while in functional regression (e.g. Ferraty and Nagy, 2022, Caldeira et al. 2020) it is common to suggest that regression on functions or curves may perform better than standard regression on some variables or vectors commonly used as regressors, this this paper is the first to provide a limiting Gaussian distribution for kernel regression on any number of random regressors and different functional objects jointly.

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 in the product space under more general assumptions than considered in the literature. In Section 4 we give some brief remarks regarding implementation (bandwidth selection). We present the results of our simulations in Section 5 and conclude in Section 6. 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. A supplementary Appendix D provides more details of the simulation.

2 The set-up and Assumptions

This section provides the formula for the Nadaraya-Watson kernel estimator over $\Xi^{[q]}$, introduces some useful notation, and gives formal assumptions. The distributional assumptions are very general. In the multivariate metric space $\Xi^{[q]}$ given by the product of the metric (measure) spaces, we assume the existence of a probability measure that could be (but not necessarily is) the product measure.

Assumption 1 *Given the metric measure spaces $\Xi_l^{[1]}$ with corresponding sigma-algebras and probability measures P_{X^l}, $l = 1, \cdots, q$, assume that the sigma-algebra for $\Xi^{[q]} = \Xi_1^{[1]} \times \cdots \Xi_q^{[1]}$ is generated by the products of sets from sigma algebras for $\Xi_l^{[1]}$ and a probability measure P_X is defined on Ξ; the mapping of $X = \left(X^1, \cdots, X^q\right)$ into each of the components $X^l \in \Xi_l^{[1]}$ is measurable (P_{X^l}) with respect to the joint measure.*

In the product space we define a vector w as $\left(w^1, \cdots, w^q\right)^T$ where each component is in the corresponding space, thus for $\Xi = R^q$, w is a q-dimensional vector of reals, in $\Xi = \Xi^{[q]}$ each $w^l \in \Xi_l^{[1]}$, $l = 1, \cdots, q$. The bandwidth vector is

$h = \left(h^1, \cdots, h^q\right)^T \in R^q$ with $\underline{h} = \min\left\{h^1, \cdots, h^q\right\}$ and $\bar{h} = \max\left\{h^1, \cdots, h^q\right\}$. We use the same notation $\|\cdot\|$ for the absolute value 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 $\Xi^{[1]}$ a Banach space, or metric (semi-metric) in a metric space $\Xi^{[1]}$; where the meaning is not clear from the context we shall specify.

2.1 The kernel

We restrict the multivariate kernel functions on R^q that will be considered to have bounded support and be suitably differentiable in the interior.

Let $I_\xi \subset \{1, \cdots, q\}$ denote any subset of q consecutive integers; there are 2^q such subsets including the empty set $\varnothing$; denote by $q(\xi)$ the cardinality of the set $I_\xi = \left\{j_1, \cdots, j_{q(\xi)}\right\}$ with $j_1 < \cdots < j_{q(\xi)}$. Let $\prod_{j \in I_\xi}(\partial_j)$ denote an operator that, when applied to a differentiable function $g(z) = g\left(z^1, \cdots, z^q\right)$ at z, maps it to its partial derivative for $j_1 < \cdots < j_{q(\xi)}$, that is

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

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.

The following assumption is made on the kernel function.

Assumption 2 *(kernel) (a) The kernel function $K(w) = K\left(w^1, \cdots, w^q\right)$ is sufficiently differentiable.*

(b) $K(w)$ is non-negative; $K(w)$ is non-increasing for $w : w^j \geq 0,\ j = 1, ..., q$.

(c) $K(w)$ is either symmetric (with respect to zero) with support on $[-1, 1]^q$ or $K(w) = K_+(w)$ is supported on $[0, 1]^q$.

(d) $K(w)$ is a type I kernel: $K(w) > 0$ on its support.

Assumptions 2(a,b,c) are satisfied by the commonly employed product kernels that use the Epanechnikov and quartic kernel; they are also satisfied by the uniform kernel. Assumption 2(d) that requires a type I kernel is not usual for kernel regression on R^q, but common in functional regression. It is satisfied by the uniform kernel. The functional regression literature demonstrates that with kernels of type I asymptotic normality can be established in more general settings. However, commonly used kernels may not be of type I and thus here the asymptotic normality results are given separately to apply under Assumption 2(a,b,c) and under the full Assumption 2 which includes the requirement of type I kernel.

2.2 The Nadaraya-Watson estimator

The Nadaraya-Watson estimator for the regression function $m(x)$ defined in (1) evaluated at a point x (here comprised of potentially various functional or other heterogeneous components), which we shall denote by $\widehat{m}(x)$, is well known; it provides a locally weighted average of Y_i where observations X_i close to x are given more weight. On the Euclidean space R^q, it uses conventional local weights which are defined via a vector $W(x) = h^{-1}(x - X)$ with components $W^l(x) = \left(h^l\right)^{-1}(x^l - X^l),\ l = 1, \cdots, q$ where h^l is a bandwidth, assigned to the l^{th} component of the random vector X, that determines the amount of smoothing that is done. The functional regression version of the kernel estimator at x in a univariate metric space $\Xi = \Xi^{[1]}$ with metric $\|.\|$ utilizes local weights based on $W(x) = h^{-1}\|x - X\|$ where h is a scalar bandwidth. On the multivariate metric space $\Xi = \Xi^{[q]} \equiv \Xi_1^{[1]} \times \cdots \times \Xi_q^{[1]}$, we consider $W(x) = h^{-1}\|x - X\|$ with $W^l(x) = \left(h^l\right)^{-1}\left\|x^l - X^l\right\|_l$, where $\|\cdot\|_l$ is the metric in $\Xi_l^{[1]},\ l = 1, \cdots, q$ and bandwidths appear for each component. Generically then

$$W(x) = \begin{cases} h^{-1}(x - X) & = \left(\left(h^1\right)^{-1}(x^1 - X^1), \cdots, \left(h^q\right)^{-1}(x^q - X^q)\right) \quad \text{on } \Xi = R^q \\ h^{-1}\|x - X\| & = \left(\left(h^1\right)^{-1}\left\|x^1 - X^1\right\|_1, \cdots, \left(h^q\right)^{-1}\|x^q - X^q\|_q\right) \quad \text{on } \Xi = \Xi^{[q]}. \end{cases} \tag{4}$$

The Nadaraya-Watson estimator, which we shall refer to as NW henceforth, is given by

$$\widehat{m}(x) = B_n^{-1}(x) A_n(x), \tag{5}$$

$$B_n(x) = \frac{1}{n}\sum_{i=1}^{n} K(W_i(x)); \ A_n(x) = \frac{1}{n}\sum_{i=1}^{n} K(W_i(x)) Y_i. \tag{6}$$

where $K(W(x)) = K(W^1(x), \cdots, W^q(x))$ is a multivariate kernel function. The kernel function $K(\cdot)$ and bandwidth vector h are chosen in a way that provides desirable properties for the kernel estimator. In the metric space $\Xi = \Xi^{[1]}$ kernel function $K(\cdot)$ is defined for a univariate nonnegative argument $K_+(\cdot)$; we extend the functional regression to the case $\Xi = \Xi^{[q]}$, where different bandwidths could appear for the different components $W^l(x), l = 1, \cdots, q$.

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

$$\arg\min_{a \in R} \sum_{i=1}^{n} (Y_i - a)^2 K(W_i(x)). \tag{7}$$

Our focus here is in establishing the limit limit properties of the NW estimator; a similar approach can be applied to other kernel estimators such as the local linear estimator.

2.3 Additional Assumptions

Consider the process $\{(X_i, Y_i)\}$. An i.i.d sequence would provide the simplest characterization, but we will permit strong mixing which makes it possible to extend the results to time series data. Denote by $\mathcal{F}_a^b$ the sigma algebra generated by $\{(X_i, Y_i)\}_{i=a}^b$. Define

$$\alpha(l) = \sup_t \sup_{A \in \mathcal{F}_{-\infty}^t; B \in \mathcal{F}_{t+l}^\infty} |P(AB) - P(A)P(B)|.$$

Recall that the process is strong mixing if $\alpha(l) \to 0$ as $l \to \infty$.

Assumption 3 *(a) The sequence (Y_i, X_i) for $i = 1, \cdots, n$ with $Y_i \in R; X_i \in \Xi^{[q]}$ is stationary and strong mixing with $\alpha(l)$ that satisfies for some $\zeta > 0$*

$$\alpha(s) < Cs^{-\kappa}; \quad \kappa > \frac{2(2+\zeta)}{\zeta}; \ \zeta > 0.$$

(b) $E(u|X = x) = 0$; $\mu_2(x) = E\left(u^2|X = x\right)$ satisfies $0 < L_{\mu_2} < \mu_2(x) < M_{\mu_2} < \infty$, $\mu_2(x)$ is continuous in the neighborhood of x.

(c) $E|Y_i|^{2+\zeta} < \infty$.

(d) For $i \neq j$ the function $\mu(x_1, x_2) = E(|u_i u_j| \ |X_i = x_1, X_j = x_2)$ is continuous in a neighborhood of (x, x).

(d) The conditional expectation $E(|Y_i Y_j| \, |X_i, X_j) \leq C_{YY} < \infty$ for all i, j.

The assumptions are standard and similar to those in Masry (2005), Delsol (2009), and Hong and Linton (2020).

Assumption 4 *The function $m(x)$ on the space $\Xi^{[q]}$ is such that*

$$|m(x) - m(z)| = M_{\Delta m} \max_l \left\|x^l - z^l\right\|_l^\delta; \ \delta > 0.$$

Assumption 4 requires Holder continuity of $m(x)$; with $\delta = 1$ would follow from differentiability or Lipschitz continuity (when this applies, e.g., for R^q).

In the above assumptions, and below, L, M denote lower and upper bounds of functions where the subscript typically denotes the function whose bounds are provided. The bounds could depend on the point x. For pointwise analysis no further restrictions are necessary, however, uniform results would necessitate further restrictions. If the bounds are constants independent of the point, then uniform properties can be obtained.

2.4 The conditioning probability measures

2.4.1 Measures on R^q

Borel probability measures on R^q are given by distribution functions and the support of the probability measure, P_X, is the set of points $x \in R^q$ where for any $r > 0$ we have for the ball $B(x, r)$

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

For the multivariate set-up, we rely on the cube (instead of small ball) measure

$$P_X(C(x, h)).$$

All rate results we obtain in R^q that consider the cube measure $P_X(C(x, h))$ when all components of h go to zero ($\bar{h} \to 0$) apply equally if we were to consider the ball, $B(x, r)$ centered art x where the rate at which the radius r goes to zero is the same as for h. Indeed

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

By the Lebesgue decomposition, the distribution F_X on R^q can be represented 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 :

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

The discrete part typically manifests itself in the data as mass points. There are many variables with purely discrete measurements, such as completed years of education, or categorical variables. There are also continuously distributed variables that exhibit mass points, e.g. Arulampalam et al. (2017), for neonatal mortality and Olson (1998), for weekly hours worked. In a multivariate setting, as soon as at least one variable is continuously distributed mass points do not arise and the joint distribution is a continuous function. However, if some variables are discrete or exhibit mass points, the distribution can no longer be absolutely continuous and has singular components (see, e.g. Kankanala and Zinde-Walsh, 2023, for a discussion). In many kernel based applications at least one of the variables is assumed continuous and an index model is assumed (single index as in Ichimura, 1993, and Klein and Spady, 1993, or multiple index as in Donkers and Schafgans, 2008) to avoid singularity as well as reduce dimensionality of the model. In a general multivariate distribution the presence of singularity achieves reduction of dimension (see, e.g. examples 2-4 in Kankanala and Zinde-Walsh, 2023) that will have a similar beneficial effect on convergence of the kernel estimator.

An example is provided by the class of Ahlfors-regular distributions (Ahlfors, 1966) on R^q is such that for some $C \geq 1, 1 \geq s > 0$ and (assuming equal components for h) scalar $h > 0$

$$C^{-1}h^{sq} \leq P_X(C(x,h)) \leq Ch^{sq}. \tag{8}$$

at every support point. For an absolutely continuous distribution with a bounded density, $s = 1$, and the so-called Hausdorf dimension of $C(x,h)$ is q, while if $s < 1$ (for example, as a result of some k components being discrete, or of a common factor structure with $q-k$ factors), the dimension is reduced to $q-k$, since then $s = \frac{q-k}{q}$. The more general characterizations provided in the next subsection also apply in R^q.

2.4.2 Measures on metric spaces and products

In functional regression on a measure metric space $\Xi^{[1]}$ when no restrictions are placed on the decline of the small ball measure pointwise asymptotic normality is established for kernels of type I. To be able to achieve consistency for a kernel that may not be type I several sufficient conditions were proposed.

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. (2007) requires that as $h \to 0$ the functions $\tau_h(\varepsilon)$ converge:

$$\tau_h(\varepsilon) \to \tau_0(\varepsilon) \quad \text{for every } \varepsilon, \tag{9}$$

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

$$\int_0^v P_X(B(x,u))\,du > CvP_X(B(x,v)) \tag{10}$$

holds for some $C > 0$, $v_0 > 0$ and any $v < v_0$. These conditions were referred to in various subsequent papers on functional regression, e.g. Hong and Linton (2020).

Here we generalize these conditions to apply on $\Xi^{[q]}$ by considering a cube rather than a ball; we also provide a modified condition that is necessary for either (9) or (10) to hold, but at the same time is sufficient for convergence without requiring type I kernel.

Assumption 5 *Given any point $x \in \Xi^{[q]}$ in the support of P_X, the probability measure P_X is such that for some h_0 and all h with $\bar{h} > h_0$ and for some $0 < \varepsilon < 1$, there is a constant $1 < C_\varepsilon < \infty$ such that*

$$\frac{P_X(C(x,h))}{P_X(C(x,\varepsilon h))} < C_\varepsilon < \infty. \tag{11}$$

Definition 1 *$\mathcal{D}$ is the class of probability measures that satisfies* (11).

This definition is equivalent to the doubling property that states that (11) applies with $\varepsilon = 1/2$. Indeed for any ε there are positive integers $\kappa_1, \kappa_2 : \varepsilon \geq 2^{-\kappa_1}$ and $2^{-1} \geq \varepsilon^{\kappa_2}$. If the measure is doubling for constant $C_{1/2}$, then (11) holds for $C_\varepsilon = C_{1/2}^{\kappa_1}$; if (11) holds, then the constant for doubling is $C_{1/2} = C_\varepsilon^{\kappa_2}$. We introduce the form (11) in case there is a preference for some ε.

The next lemma shows that the condition (11) holds whenever (9) or (10) hold.

Lemma 1 *The condition* (11) *of Assumption 5 is a necessary condition (a) for (9) to hold with $\tau_0(\varepsilon) > 0$ and (b) for (10) to hold.*

A particular class of probability measures considered in functional regression (e.g. Ferraty and Vieu, 2006) has small ball probability at a point x of the form

$$P_X(B(x,h)) = C_1 h^\gamma \exp\left(-C_2/h^\beta\right) \tag{12}$$

with positive C_1 and γ as well as at least non-negative C_2 and β. When either C_2 or β is zero, the measure is in class $\mathcal{D}$. With both C_2 and β positive constants the probability measure is such that the small ball probability declines exponentially in h; then (11) does not hold and the measure is not in $\mathcal{D}$. Here kernels of type I are required. The exponential rate of decay of the small ball probability is shown to lead to slow convergence for the estimators (curse of dimensionality). There are ways to mitigate the curse of dimensionality arising from such exponential decay. It is common to apply finite dimensional approximation of these functionals as suggested in Gasser et al. (1998). Indeed, the case where functional data can be accurately approximated in a finite dimensional space is not rare (correspond to observation of smooth curves with common shape) as noted by Ferraty and Nagy (2022).

We next provide some examples for measures in class $\mathcal{D}$.

1. Consider $\Xi^{[q]} = \Xi_1^{[1]} \times \cdots \Xi_q^{[1]}$. If the measure P_{X_l} on each $\Xi_l^{[1]}$ satisfies the Ahlfors regularity condition with order s for all $l = 1, \cdots, q$, then the measure P_X on $\Xi^{[q]}$ has bounds

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

   at every point of support. If x is an isolated mass point then (13) applies with $s = 0$. Such a measure is in $\mathcal{D}$.

2. Suppose that P_X belongs to a finite mixture of distributions that satisfy (13) for different s : $P_X = \sum_{l=1}^L \alpha_l P_{s_l}$ with $0 < s_1 \leq \cdots \leq s_L \leq 1$, $\alpha_l > 0$, and $\sum_{l=1}^L \alpha_l = 1$. Every point in the support of P_X is a point of continuity if $s_1 > 0$. To verify that (11) is satisfied, we first note that for any s_l the condition holds with some corresponding $C_\varepsilon = C(s_l)$ and probability

measure P_{s_i}. As $P_X(C(x,h)$ can be represented as a sum of probabilities conditional on x being generated by a term P_{s_l} in the mixture, times the marginals, we then apply (11) to every term in the sum

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

With $\bar{\varepsilon}=\max\left(\varepsilon_l\right)$ and $C_{\bar{\varepsilon}}=\max C\left(s_l\right)$ the condition (11) therefore is shown to hold.

3. Consider on R^q a mass point x^* in the support of F_X where the distribution F_X is discontinuous. $F_X(x)=\alpha I\left(x\geq x^*\right)+\left(1-\alpha\right)F^c(x)$, with the distribution function F^c continuous. Suppose that $F^c\left(x\right)$ satisfies (11), then at any point of support $x\neq x^*$, with $h<\left\|x-x^*\right\|$ (11) holds. At the point x^*

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

We have that $P_X(C(x^*,h)$ is bounded and $P_X(C(x^*,\varepsilon h)>\alpha$, thus (11) also holds for $C_\varepsilon\geq 2/\alpha$ (regardless of h) at x^*.

2.4.3 Joint measure

Consider now the dependence structure of the data sequence; by Assumption 3 it is stationary and mixing. Denote the product space $\Xi^{[2q]}=\Xi^{[q]}\times\Xi^{[q]}$; the measure on this product space has marginals P_X on each $\Xi^{[q]}$, see, e.g., Pollard (2001).

Let X_s and X_t denote two distinct random observations on $\Xi^{[q]}$. In the case of independence the joint measure $P_{s,X}\left(C\left(x,h\right)\times C\left(x,h\right)\right)=\Pr\left(X_s\in C\left(x,h\right),\right.$ $\left.X_t\in C\left(x,h\right)\right)$ is just a product of the P_X measures of the cube. But with dependence an additional assumption needs to be made on how the joint measure relates to the measure of the small cube. We provide the same assumption as, e.g., Masry (2005) and Hong and Linton (2020), which we apply to the small cube (rather than small ball) probability.

Assumption 6 *The joint measure* $P_{s,t}\left(C\left(x,h\right)\times C\left(x,h\right)\right)$ *is such that for some* $0<M_{FF}<\infty$

$$\sup_{s\neq t}P_{s,t}\left(C\left(x,h\right)\times C\left(x,h\right)\right)\leq M_{FF}\left(P_X\left(C\left(x,h\right)\right)\right)^2. \tag{14}$$

The assumption is trivially satisfied under independence.

3 Asymptotic normality of kernel estimator

Consider the NW estimator as given by (5) . As the sample size increases the bandwidths are assumed 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 the density, $f_X(x)$, exists and is continuous, the estimator $h^{-q}B_n(x)$ consistently estimates $f_X(x)$, but if the 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), and uniform convergence rates were derived by Ferraty, Laksaci, Tadj and Vieu (2010). The limit distribution for the estimator was examined in Masry (2005), Ferraty et al. (2007), Delsol (2009), and Geenens (2015). These papers emphasized the type I kernel for the infinite dimensional function space and allowed for exponentially decaying small ball probability.

Our theorem below establishes an asymptotic normality result under a wide variety of assumptions over regressors in the space Ξ, that could be R^q, or metric space $\Xi^{[1]}$, but also could be some product of metric spaces, $\Xi^{[q]}$.

The key to deriving the distributional results is the computation of moments for multivariate functions $\psi(W)$ of the random vector W that takes values in R^q. We consider $\psi(W)$ for $W = \left(W^1, \cdots, W^q\right)$ where for $X \in R^q$ the component W^l is defined as $(h^l)^{-1}(x^l - X^l)$, while for $X \in \Xi^{[q]}$ the component W^l is defined as $(h^l)^{-1}\left\|x^l - X^l\right\|_l$.

The technical results of Appendix A provide moments and bounds on the various moments that apply in R^q and product $\Xi^{[q]}$ spaces. The general moments expressions provided in Appendix A are applicable to the different functions of interest, such as kernels, powers and products of kernels and their products with conditional moments.

For example, the bounds for the expectation of the function $g(X)K^m\left(\frac{x-X}{h}\right)$, where $g(\cdot)$ is a continuous function, are provided in terms of the small cube probability as

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

with constants $L_{EgK^m}(x)$ and $M_{EgK^m}(x)$ defined at x. For $g(\cdot) = 1$ and $m = 1$ this defines the bounds on the expectation of $B_n(x)$, defined in (6). We provide its lower bound, which we denote as $L_{EB}P_X(C(x,h))$, with appropriate conditions for L_{EB} to be strictly positive; this is crucial to ensure that the denominator is such that it exists and the limit does not blow up. Type I kernel automatically ensures that $L_{EB} > 0$, but for kernels such as Epanechnikov the bound requires an extra assumption such as Assumption 5.

Once the moments and the bounds are derived, the proofs of asymptotic normality proceed along the same lines as those in Masry (2005), or Hong and Linton (2020), for the stationary case. The mixing Assumption 3 provides

weak enough dependence that the expectations and variances of the numerator and denominator of the NW estimator have the same leading terms as under independence ensuring the bounds enjoy the same rates.

Theorem 1 provides the limit normality under two alternative types of conditions: (i) applies to kernels of type I kernel without imposing further constraints on the conditioning distribution, and (ii) does not impose the type I kernel but imposes Assumption 5 (equivalent to the doubling measure).

Theorem 1 *Under either of the following sets of assumptions (i) Assumptions 1-4 and 6 or (ii) Assumptions 1, 2(a-c), 3-6 for* $h \to 0$ *as* $n \to \infty$ *such that* $nP_X\left(C\left(x,h\right)\right) \to \infty$

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

The rate for $E(B_n(x))^2/varA_n(x)$ is shown to be $nP_X(C(x,h)$ and the rate for the bias of $\widehat{m}\left(x\right)$ in $\Xi^{[q]}$ is established in the Theorem below as $O\left(\bar{h}^{\delta}P_X\left(C(x,h)\right)\right)$. The point-wise consistency and asymptotic normality of the Nadaraya-Watson estimator with undersmoothing is summarized in the theorem below.

Theorem 2 *Under the conditions of Theorem 1*

(a) the bias of $\widehat{m}\left(x\right)$ *in* $\Xi^{[q]}$ *is* $O\left(\bar{h}^{\delta}P_X\left(C(x,h)\right)\right)$*;*

(b) for h *such that* $\frac{\bar{h}^{\delta}}{\sqrt{varA_n(x)}} \to 0$

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

where δ *is defined in Assumption 4.*

Remarks.

1. The moments bounds under the assumptions of Theorems 1 and 2 imply that $EB_n(x)$ has the same rate as $P_X(C(x,h))$ while $varA_n(x)$ declines at the rate $\frac{P_X(C(x,h))}{n}$. **The rate** for $\frac{EB_n(x)}{\sqrt{varA_n(x)}}$ therefore equals $n^{1/2}P_X\left(C\left(x,h\right)\right)^{1/2}$ (this diverges by assumption of Theorem 1).

2. The limit result shows that for a distribution on R^q, when density exists, the standard convergence rate $n^{1/2}h^{q/2}$ applies on R^q since then $P_X\left(C\left(x,h\right)\right) = O\left(h^q\right)$. This rate holds even when the density is discontinuous. Without the usual smoothness assumptions made in the literature, we therefore provide **statistical guarantees for the rate**.

3. If there is singularity at point x and (13) is satisfied with $s < 1$, then the rate of convergence is $n^{1/2}h^{sq/2}$, which is 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". When x is an isolated mass point then convergence of the estimator at that point is at the parametric rate $n^{1/2}$.

4. 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

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

Suitable assumptions on the distribution, such as H_3 in Ferraty et al. (2007) and Condition 3(i) in Masry (2005) and similar ones in subsequent papers provide conditions on the probability measure on $\Xi^{[1]}$ that are sufficient for the convergence. Generally, one needs to ensure that the limits in the following assumption hold.

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

$$\begin{aligned} \frac{EK(W_i(x))}{P_X(C(x,h))} &\to \bar{B}_1(x); \\ \frac{var\left(K(W_i(x))Y_i\right)}{P_X(C(x,h))} &\to \bar{A}_2(x). \end{aligned}$$

Assumption 7 holds e.g. for fractal distributions on R^q for which $\lim h^{-sq}P_X(C(x,h))$ exists for $h \to 0$. An absolutely continuous distribution with a continuous density also satisfies this assumption.

Define now

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

Theorem 3 *Under the conditions of Theorem 2 and Assumption 7 with $\alpha(n,h) \to \infty$ and for h such that $\alpha(n,h)\bar{h}^{2\delta} \to 0$*

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

This limit extends the results that were obtained in the literature on kernel estimation in R^q under more restrictive assumptions on the distribution F_X. For functional regression with X in a metric space $\Xi^{[1]}$ our assumptions are comparable to those of Ferraty and Vieu (2007), Masry (2005), and subsequent papers. We provide a somewhat more general condition in Assumption 5 under which asymptotic normality holds for kernels that could be zero on the boundary. Theorem 2 is a general result and provides a useful basis for inference, while Theorem 3 requires an additional assumption such as Assumption 7. Finally, for multivariate regression on $\Xi^{[q]}$ with an arbitrary finite q the general limit distribution characterization is provided.

4 Implementation

Estimation of the function $m(x)$ requires a selection of the kernel, $K(\cdot)$, and bandwidth, h. As is clear from the results here and the literature, type I kernel (such as the uniform) is preferred but other kernels can also deliver asymptotic rates provided the small cube probability does not decline exponentially fast. Selection of a bandwidth vector sequence impacts convergence of the estimator. Pointwise convergence may imply the need to vary the bandwidth depending on the point, whereas some overall convergence criterion, such as e.g. Mean Integrated squared error (MISE) may have as a solution a unique bandwidth for all points.

The choice of a unique bandwidth could be based on a criterion such as MISE applied for a particular known distribution (e.g. Gaussian) replacing the unknown underlying distribution measure leading to a "rule-of-thumb" bandwidth. Such a choice was used e.g. in Hong and Linton (2020) for functional regression for an infinite autoregressive process where such a bandwidth was chosen on the basis of an i.i.d. Gaussian process.

The least squares cross-validation (CV) method of bandwidth selection is a popular approach to obtaining a unique bandwidth for all points. This method was used by Ferraty and Vieu (2007), Hong and Linton (2020), and Ferraty and Nagy (2022) among many others. The CV approach to selection of the bandwidth vector has the advantage well documented in regression on R^q (Hall et al., 2007) of providing the ability to discern the possible redundant regressors by selecting increasing bandwidths for those components of the regressor vector.

Each approach to bandwidth selection has advantages and disadvantages. In particular, the distinction between a unique bandwidth and a point-wise bandwidth is important, in that when the interest is in the behavior of the conditional mean function (say, smoothness vs wiggliness, or lack of dependence on some component of X) then a unique bandwidth sequence has a better chance of providing an estimator that is informative of those features, and a point-wise bandwidth may be misleading.

We do not pursue this important investigation here, instead relying on the methods widely used in the literature. In particular, in the next section we concentrate on a CV bandwidth sequence. In the simulations we also evaluate performance at bandwidths around the CV bandwidth by considering several multiples of it to gain some insight into the stability of results over different bandwidth choices.

Aside from the estimator of conditional mean, estimators of variance and mean squared error are needed to evaluate the performance of the estimator. While in the literature on kernel regression on R^q, the leading term of the limit variance is expressed via the density function, typically in the actual implementation the corresponding estimators do not make use of those expressions, instead estimating the variance directly from the data and possibly with bootstrap (see Hall and Horowitz, 2013, and Ferraty, van Keilegom and Vieu, 2010). This is consistent with our asymptotic results that do not rely on existence of density.

5 Simulations

This section provides various simulations that show the finite sample performance of NW estimator under singularity in a point mass univariate setting and in various multivariate settings, including functional regressors. We provide here the main results. Additional details and features are in the supplementary Appendix D.

5.1 Univariate (Point mass example)

In this example we consider a conditioning distribution with mass points. Alongside we consider an a.c. distribution which represents features (high derivatives) that are more reflective of unsmooth densities we want to emphasize here.

Illustrative graphs of the distributions F_X we consider here are given in Figure 1. The distribution with mass points considered, following Jun and Song (2019), is given by

$$F_X(x) = pF^d(x) + (1-p)\Phi(x) \quad \text{with } p = 0.2,$$

where F^d is the discrete uniform distribution function with $D = \{-1, 0, 1\}$; Φ is the standard Gaussian distribution function. The a.c. counterpart, reflective of unsmooth densities, is the trinormal mixture considered 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 denotes the standard Gaussian density function. [4]

Include Figure 1

We simulated 500 random samples $\{(y_i, X_i)\}_{i=1}^n$ using the model

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

for different sample sizes. The error $\{\varepsilon_i\}_{i=1}^n$ is drawn independently of the regressor and has a standard Gaussian distribution; σ is selected to yield a given signal to noise ratio, snr, here selected to equal one. We use the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)1(u^2 \leq 1)$ and obtain the leave-one-out crossvalidated bandwidth.

In Figure 2, we evaluate the NW nonparametric fit locally. The figure shows the root mean squared error (RMSE) over the 500 replications with $n = 1,000$ at the cross-validated bandwidth.

Include Figure 2

Pointwise convergence rates of the NW estimator are influenced non-trivially by the underlying distribution. To gain insight into such local differences, we

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

analyse the pointwise RMSE results with cross validated bandwidth at a coarse grid of points across samples of size n equal to 50, 100, 200, 400, 800, 1600, 3200 based on 500 replications from the above DGP. To obtain empirical rates of convergence we run a regression of $\log(RMSE)$ on $\log(n)$ and a constant. The coefficient on $\log(n)$ is the "realized" rate of convergence; for example if $RMSE \propto n^{-2/5}$ (univariate kernel regression with smooth density and second order kernel) then $\log(RMSE) = \alpha_0 + \alpha_1 \log(n)$ and α_1 can be estimated and should be close to -0.4. A comparison is made to the 'oracle' estimator, which applies OLS to the known DGP where α_1 is expected to be close to -0.5.[5]

In Table 1, illustrative results are provided for the conditioning distribution with mass points and the trinormal conditioning distribution on a set of support points.

Include Table 1

For the conditional distribution with mass points, the NW estimator with cross-validated bandwidth performs remarkably well at points sufficiently far from our mass points (faster than the expected rate of -0.4), while the deterioration at mass points, compared to the achievable parametric rate, is small. The empirical rate at mass points is close to -0.5 when the bandwidth is set equal to zero. The deterioration of the empirical convergence rate for points close to the mass points (within the small ball probability measure under cross validated bandwidth) is more severe due to the boundary weight associated with mass in the neighborhood. Clearly, the use of smaller bandwidths will reduce its impact. The convergence rates for the tri-normal distribution, are reflective of usual smooth nonparametric regression even at points with high derivatives.

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

In this section we consider a bivariate regression model in the presence of a singularity. In particular, we consider the setting where the regressors lie on a unit circle, that is

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

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). See also Desmet and Parente (2010).

Subject to this singularity, we simulated 500 random samples $\{(y_i, X_{1i}, X_{2i}\}_{i=1}^n$ from the model

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

for different sample sizes. We assume $F_X(x_1, x_2)$ is uniform on the unit circle. To obtain our random sample we draw ϕ from $U[0, 2\pi]$ and use $(X_1, X_2) = (\cos(\phi), \sin(\phi))$. The error $\{\varepsilon_i\}_{i=1}^n$ is drawn independently of the regressors and

[5]The authors thank Jeff Racine for suggesting this insightful exercise. See also Hall and Racine (2015).

has a standard Gaussian distribution; σ is selected to yield a given signal to noise ratio, here selected to equal one.

The deterministic relation between X_1 and X_2 implies a reduced dimension; it is common to reduce dimension by employing a single index model (Ichimura, 1993), which is valid here. We implement this estimator as in Hardle et al. (1993).

We evaluate the average RMSE over 500 simulations at a grid of 100 (x_1, x_2) points based on equidistant values of ϕ on $[0, 2\pi]$ for $n = 1,000$. Figure 4 shows the corresponding curves at the cross-validated bandwidth for the NW and single index (SI) estimators. The results reveal a comparable (and sometimes superior) performance of the NW estimator relative to the single index estimator performance.

Include Figure 4

When evaluating the pointwise convergence rate, we find that at points on the unit circle, the NW estimator typically achieves a faster empirical rate of convergence than -0.4, faster than that of the SI estimator at these points. Illustrative results hereof are provided in Table 2. For comparison we also consider simulations with X_1 and X_2 drawn independently on $U[-1, 1]$. There, as expected, the rate of convergence is slower than SI where RMSE $\propto n^{-2/(q+4)}$ with $q = 2$. The results support the rate reduction due to singularity.

Include Table 2

5.3 Bivariate (with mass of (X_1, X_2) on 1-dimensional line)

Here we consider a model where $m(X) = \log(X_1) + \log(X_2)$ with conditioning variables X_1 and X_2 satisfying $X_1 + X_2 = d(k)$, with fixed $d(k)$ corresponding to $k = 1, 2, 3$. An example could be where X_1 and X_2 represent incomes of the husband and wife and, for tax purposes, their combined income is set at some $d(k)$. This is equivalent to one continuous and one discrete regressor, however, such equivalence may not be easily deduced from the multivariate data. The rate of convergence is determined by the lower number of continuous regressors only, see also Racine and Li (2004).

We simulated 500 random samples $\{(y_i, X_{1i}, X_{2i}\}_{i=1}^n$ using the model

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

for different sample sizes with the additive error chosen as in previous simulations. The probability of an observation belonging to a sub-population with $k = 1, 2, 3$ is set equal to 0.5, 0.3, and 0.2 respectively and $d(1) = 4, d(2) = 6, d(3) = 7$; X_1 is drawn from the uniform distribution: $U[1, 3]$.

We implement the NW estimator in this set-up twice: first using X_1 and X_2 as conditioning regressors (NW.c) and second using X_1 and $D = d(k)$ as conditioning regressors (NW.d). For the categorical (ordered) discrete conditioning regressor D we use weights proposed by Wang and van Ryzin (1981) in place of the Epanechnikov kernel.[6]

[6]Details of the geometric weights used are provided in the supplementary Appendix D.

In Table 3 we provide illustrative results comparing the empirical rate of converge of the NW.d and NW.c to the oracle estimator at a coarse grid.

Include Table 3

The reduced dimensionality is reflected in the estimates of the pointwise rate of convergence which are around -0.40 rather than the slower rate of -0.33 the presence of two continuous regressors would suggest ($q = 2$). Interestingly, the estimate of the empirical rate for NW.c is slightly faster than NW.d (sampling variability ignored). The results also compare favourably to the SI estimator which requires in this case the knowledge that the conditioning variables are given by $(\log(X_1), \log(X_2))$ here.

5.4 Bivariate (in the presence of a functional regressor)

In this section we consider the presence of a functional regressor in a multivariate setting.

We consider the following bivariate conditional mean function $m(X) = m_1(X_1) + X_2$, where X_1 is a functional regressor and $X_2 \in R$ is an additional control (random scalar) which may be correlated with $m_1(X_1)$. Following Ferraty et al. (2007), the functional regressor is defined as

$$X_1(t) = \sin(wt) + (a + 2\pi)t + b, \quad t \in (-1, 1)$$

with a and b drawn from $U(-1, 1)$ and w drawn from $U(-\pi, \pi)$ and

$$m_1(X_1) = \int_{-1}^{1} |X_1'(t)|(1 - \cos(\pi t))dt.$$

For X_2 we first consider a N(0,1) random variable which is drawn independently of X_1. To permit the additional regressor to be correlated with the functional regressor X_1, we also generate the regressor $X_2 = m_1(Z)$ where $Z(t)$ is a functional regressor similar to $X_1(t)$ with (a, b, w) replaced by (a', b', w') where the correlation between (a', b', w') and (a, b, w) is given by ρ (and set equal to 0 and 0.8).

We use discretized curves X_1 and Z where $-1 = t_1 < t_2 < \cdots, t_{101} = 1$ for our simulation. Thus, we obtain 500 random samples $\{(y_i, \{X_{i1}(t_j)\}_{j=1}^{101}, X_{i2})\}$ with sample size 250, for the model

$$y = m_1(X_1) + X_2 + \sigma\varepsilon$$

with $n = 250$. ε is drawn from $N(0, 1)$ and σ is selected to yield a signal to noise ratio equal to one.

The weights that define the NW estimator in this multivariate setting is given by

$$W(x) = \left((h_1)^{-1} \|x_1 - X_1\|_1, (h_2)^{-1} (x_2 - X_2) \right),$$

where we use the L_2 norm between first order derivatives of the curves of the functional regressor X_1 as in Ferraty et al. (2007), that is

$$\|x_1 - X_1\|_1 = \sqrt{\int_{-1}^{1} (x_1'(t) - X_1'(t))^2 \, dt}.$$

We use a product kernel with kernel $K(u) = 1 - u^2$ defined on $[0, 1]$ for the functional regressor and the Epanechnikov kernel defined on $[-1, 1]$ for X_2. We obtain the leave-one-out cross-validated bandwidth vector $(h_{1,cv}, h_{2,cv})$.

In the supplementary Appendix D, we show what impact excluding either the functional or the additional control has on the residual sum of squares. When X_2 is given by an independently drawn $N(0, 1)$ regressor, the loss of ignoring X_2 is more severe than dropping the function regressor. When using $X_2 = m_1(Z)$ as additional regressor, the loss of dropping either regressor is stronger when $\rho = 0$ than $\rho = 0.8$ as expected. The impact on the residual sum of squares of dropping either is similar in this setting. In Table 4, we provide the RMSE of the NW estimator at the cross-validated bandwidths for our models together with the misspecified RMSE where either the functional or control is dropped.

Include Table 4

To evaluate the pointwise performance of the NW estimator, we obtain 50 out of sample observations drawn from the DGP (constant across simulations). As discussed in the supplemental appendix, at the cross-validated bandwidth not all out-of-sample observations are sufficiently close to the random sample in each simulation. These out-of-sample observations will be reflective of settings where the small ball probability (as represented by observed neighbors given cross-validated bandwidth) is too small. As can be seen from Table 4, the overall RMSE performance on these 50 out-of-sample observations is comparable to the in-sample RMSE.

The pointwise RMSE is large at out-of-sample observations where the small ball probability (as represented by the number of neighbors) is small. This can be seen from Figure 5 which presents for the cross-validation bandwidth the average number of neighbors by pointwise RMSE.

Include Figure 5

6 Conclusions

The paper established the rate and asymptotic normality for the Nadaraya-Watson estimator for arbitrary conditioning distributions in R^q including cases of singularity that could arise from fractal distribution of regressors, factor structure, mass points or discreteness in some of the regressors. In fact, it is demonstrated that the convergence rate is faster than the standard rate when the regressors exhibit dimension reducing types of singularity (e.g. a fractal distribution or a factor structure). Finite sample evidence confirms the rate

improvement due to singularity in regression over R^q. For the case of an absolutely continuous distribution statistical guarantees for the standard rate and the asymptotic normality are provided without requiring further smoothness.

The paper extends the limit theory of kernel functional regression to multivariate regression over a product of any number of metric spaces, which include functional and vector spaces. The theoretical results provide the pointwise asymptotic normality for such multivariate regressions. Simulations highlight the importance of accounting for additional (functional) regressors in the presence of a functional regressor.

A The general technical results

In this appendix, we provide derivations for moments and bounds on the moments that are used in establishing the limit properties of the estimators.

We provide the expectation of multivariate functions $\tilde{\psi}(W)$ for $W = \left(W^1, \cdots, W^{\tilde{q}}\right)$, defined locally relative to the measure $P_{\tilde{X}}$. We distinguish two cases depending on whether $\tilde{X} \in \Xi^{[q]}$ or $\tilde{X} \in \Xi^{[2q]} = \Xi^{[q]} \times \Xi^{[q]}$ (product space).

Case 1. If $\tilde{X} \in \Xi^{[q]}$ then $\tilde{q}, \Xi^{[\tilde{q}]}, \tilde{X}$, $\tilde{x}, \tilde{h}$ coincide with $q, \Xi^{[q]}, X, x, h$.
The function $\tilde{\psi}(W)$ is

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

when $\Xi^{[q]} = R^q$; correspondingly, on general $\Xi^{[q]}$:

$$\tilde{\psi}(W) = \psi_+\left(\frac{\|x - X\|}{h}\right) = \psi_+\left(\frac{\left\|x^1 - X^1\right\|_1}{h^1}, \cdots, \frac{\|x^q - X^q\|_q}{h^q}\right). \tag{A.2}$$

Case 2. If $\tilde{X} \in \Xi^{[2q]}$ then $\tilde{q}, \Xi^{[\tilde{q}]}, \tilde{X}$, $\tilde{x}, \tilde{h}$ coincide with $2q$, $\Xi^{[q]} \times \Xi^{[q]}$, $(X_t, X_s) = \left(X_t^1, \cdots, X_t^q, X_s^1, \cdots, X_s^q\right)$, and $(x, x), (h, h)$.
The function $\tilde{\psi}(W)$ for the corresponding product spaces $\Xi^{[\tilde{q}]} = \Xi^{[q]} \times \Xi^{[q]}$ is

$$\begin{aligned}
\tilde{\psi}(W) &= \psi\left(\frac{x - X_t}{h}\right)\psi\left(\frac{x - X_s}{h}\right), \text{ or} \\
\tilde{\psi}(W) &= \psi_+\left(\frac{\|x - X_t\|}{h}\right)\psi_+\left(\frac{\|x - X_s\|}{h}\right).
\end{aligned}$$

Note that Case 1 is the general case. Case 2 is specific to evaluation of covariances arising from any Case 1.

We consider the set of the indices $\{1, 2, \cdots, \tilde{q}\}$; there are $2^{\tilde{q}}$ subsets of this set including the empty set, $\varnothing$. Denote each subset by I_ξ; $\xi = 0, 1, \cdots, 2^{\tilde{q}} - 1$ 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. $q(\xi)$ denotes the cardinality of the subset I_ξ. The complement of the subset

I_ξ is denoted by I_ξ^c. Thus, for instance, for $\tilde{q} = q = 3$, there are 8 subsets: $I_0 = \varnothing; I_1 = \{1\}, I_2 = \{2\}, I_3 = \{3\}, I_4 = \{1,2\}, I_5 = \{1,3\}; I_6 = \{2,3\}, I_7 = \{1,2,3\}$.

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

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

Using the delta-function operator, $\delta\left(z^j = a\right)$, that applied to $g\left(z^1, \cdots, z^j, \cdots, z^{\tilde{q}}\right)$ sets the jth component to the 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 often $a = 1$ denoting a point on the boundary).

We deal with three situations for $\tilde{\psi}(W)$. First, the following Lemma A.1 provides the expectation of $\tilde{\psi}(W) = \tilde{\psi}\left(\frac{\tilde{x}-\tilde{X}}{\tilde{h}}\right)$ with respect to the probability measure $P_{\tilde{X}}$ for a sufficiently differentiable random function $\tilde{\psi}$ on $R^{\tilde{q}}$ with support $[-1,1]^{\tilde{q}}$. The probability measure $P_{\tilde{X}}$ could be P_X or a measure on the product space with P_X marginals. Second, Corollary A.1 to Lemma A.1 details this result for a symmetric function on $[-1,1]^{\tilde{q}}$. Third, Lemma A.2 provides the expectation of a sufficiently differentiable function $\tilde{\psi}_+\left(\frac{\|\tilde{x}-\tilde{X}\|}{h}\right)$ given on $[0,1]^{\tilde{q}}$ relative to a measure on $\Xi^{[\tilde{q}]}$. This is followed by examining moments for the product $g(\tilde{X})\tilde{\psi}\left(\frac{\tilde{x}-\tilde{X}}{\tilde{h}}\right)$.

The moments for $\tilde{\psi}(W)$ are expressed via sums of integrals for every ξ over some $S^{q(\xi)} \in R^{q(\xi)}$ where usually $S^{q(\xi)} = [-1,1]^{q(\xi)}$ or $[0,1]^{q(\xi)}$. The integrals involve probability measures of some sets, $\widetilde{C}$, $P_{\tilde{X}}(\widetilde{C})$, where $\widetilde{C}$ is inside the cube $\tilde{C}(x,h)$ in $\Xi^{[\tilde{q}]}$. Given a local function $\tilde{\psi}(.)$ and the operator $\Delta_{a,\xi}$ we define

$$\mathcal{I}_\xi\left(\tilde{x}; S^{q(\xi)}, P_{\tilde{X}}(\widetilde{C}), \Delta_{a,\xi}\tilde{\psi}\right) = \int_{S^{q(\xi)}} P_{\tilde{X}}(\widetilde{C})\left(\Delta_{a,\xi}\tilde{\psi}(v)\right) dv(\xi). \tag{A.4}$$

Lemma A.1 *Suppose that the function $\tilde{\psi}(W)$ is sufficiently differentiable with support on $[-1,1]^{\tilde{q}}$ and $P_{\tilde{X}}$ is the measure associated with $\tilde{X} \in R^{\tilde{q}}$ given by a distribution function $F_{\tilde{X}}(\cdot)$ on the space $R^{\tilde{q}}$. Then with $P_{\tilde{X}}(\widetilde{C})$ expressed as $F_{\tilde{X}}(\tilde{x} - \lambda_\xi \circ \tilde{h}, \tilde{x} + \tilde{h})$ where λ_ξ is a vector with components:*

$$\{\lambda_\xi\}^j = v^j \text{ if } j \in I_\xi, \text{ otherwise } \{\lambda_\xi\}^j = 1, \tag{A.5}$$

the expectation is

$$\begin{aligned} & E\left(\tilde{\psi}\left(\frac{\tilde{x}-\tilde{X}}{\tilde{h}}\right)\right) \\ = & \sum_{\xi=0}^{2^{\tilde{q}}-1}\int_{[-1,1]^{q(\xi)}} F_{\tilde{X}}(\tilde{x}-\lambda_\xi\circ\tilde{h},\tilde{x}+\tilde{h})\left(\Delta_{1,\xi}\tilde{\psi}\left(v^1,\cdots,v^{\tilde{q}}\right)\right)dv(\xi) \\ = & \sum_{\xi=0}^{2^{\tilde{q}}-1}\mathcal{I}_\xi\left(\tilde{x};[-1,1]^{q(\xi)},F_{\tilde{X}}(\tilde{x}-\lambda_\xi\circ\tilde{h},\tilde{x}+\tilde{h}),\Delta_{1,\xi}\tilde{\psi}\right) \end{aligned} \tag{A.6}$$

The notation $F_{\tilde{X}}(\tilde{x}-\lambda_\xi\circ\tilde{h},\tilde{x}+\tilde{h})$ where $\tilde{X}\in R^q$, equals

$$F_{\tilde{X}}(\tilde{x}-\lambda_\xi\circ\tilde{h},\tilde{x}+\tilde{h})=\int_{\tilde{C}}dF_{\tilde{X}}$$

with $\tilde{C}=\prod_{j\in I_\xi}\left[\tilde{x}^j-v^j\tilde{h}^j,\tilde{x}^j+\tilde{h}^j\right]\prod_j\int_\xi^c\left[\tilde{x}^j-\tilde{h}^j,\tilde{x}^j+\tilde{h}^j\right]$, where $F_{\tilde{X}}$ is the distribution function of $\tilde{X}$.

Corollary A.1 *Suppose that the function $\tilde{\psi}(W)$ is sufficiently differentiable with support on $[-1,1]^{\tilde{q}}$. If in addition to the conditions of Lemma A.1 the function $\tilde{\psi}(\cdot)$ is symmetric around zero in every argument, then for $\xi\in I_\xi$ the corresponding $\widetilde{C}=C(\tilde{x},\lambda_\xi\circ\tilde{h})$ where λ_ξ a vector with components:*

$$\{\lambda_\xi\}^j=v^j \text{ if } j\in I_\xi, \text{ otherwise } \{\lambda_\xi\}^j=1, \tag{A.7}$$

and

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

We see that symmetry simplifies the moment expression.

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

Lemma A.2 *Suppose that the function* $\tilde{\psi}(W) = \tilde{\psi}_+\left(\frac{\left\|\tilde{x}^1-\tilde{X}^1\right\|_1}{\tilde{h}^1},\cdots,\frac{\left\|\tilde{x}^q-\tilde{X}^{\tilde{q}}\right\|_q}{h^{\tilde{q}}}\right)$ *where* $\tilde{\psi}_+(\cdot)$ *defined on* $[0,1]^{\tilde{q}}$ *is sufficiently differentiable and* $P_{\tilde{X}}$ *is a probability measure defined on* $\Xi^{[\tilde{q}]}$ *and* $\widetilde{C} = C(\tilde{x}, \lambda_\xi \circ \tilde{h})$ *with* λ_ξ *defined in (A.7) then*

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

A simplified expression for functions that take zero value on the boundary is provided in the next corollary. We say that a function $\tilde{\psi}(\cdot)$ in Lemmas A.1, A.2 and Corollary A.1 is zero on the boundary if for any w with at least one w^l with $\left|w^l\right| = 1$ the value $\tilde{\psi}(w)$ is zero. In this case the only non-zero term in the sums in (A.6,A.8, A.9) corresponds to $\xi = 2^{\tilde{q}} - 1$.

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

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

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*

$$\left|E\left(\tilde{\psi}\left(\frac{\tilde{x}-\tilde{X}}{\tilde{h}}\right)\right)\right| \le P_{\tilde{X}}(C(\tilde{x},\tilde{h}))M_{E\psi}(\tilde{x}),$$

where

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

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.

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

$$\frac{P_{\tilde{X}}(C(\tilde{x},\tilde{h}))}{P_{\tilde{X}}(C(\tilde{x},\varepsilon\tilde{h}))} < C_{\varepsilon} < \infty,$$

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

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

where

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

(b) If ψ satisfies

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

then there is the lower bound:

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

where

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

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

$$\Omega_{\tilde{g}}(\widetilde{C}) = \int_{\widetilde{C}} \tilde{g}(z)dP_{\tilde{X}}(z) \tag{A.11}$$

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 $\tilde{\psi}(.)$, for a bounded continuous function $\tilde{g}(.)$ the moment*

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

with $\Omega_{\tilde{g}}\left(C(\tilde{x},\lambda_{\xi}\circ\tilde{h})\right)$ given by (A.11) with $\widetilde{C} = C(\tilde{x},\lambda_{\xi}\circ\tilde{h})$

(b) The moment is bounded

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

with

$$M_{Eg\psi} = \sup_{x \in C(\tilde{x},\tilde{h})} |\tilde{g}(x)| \, M_{E\psi};$$

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

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

with

$$L_{Eg\psi} = \max\left\{ 0, \left(\inf_{x \in C(\tilde{x},\tilde{h})} \tilde{g}(x) \right) L_{E\psi} \right\}$$

where for (a)

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

under (b)

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

When $\inf_{x \in C(\tilde{x},\tilde{h})} \tilde{g}(x) > 0$, the lower bound is positive.

B Proofs of general technical results

Proof of Lemma A.1.
Remark that for a univariate function

$$\tilde{\psi}(1) = \tilde{\psi}(w) + \int_w^1 (\partial_v) \tilde{\psi}(v) \, dv,$$

thus

$$\tilde{\psi}(w) = \tilde{\psi}(1) - \int_w^1 (\partial_v) \tilde{\psi}(v) \, dv.$$

Generalizing to the multivariate case the function $\tilde{\psi}(w^1, \cdots, w^{\tilde{q}})$ (over $R^{\tilde{q}}$) we obtain

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

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

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

To each term of (B.1) with $v^j = \frac{\tilde{x}^j - t^j}{\tilde{h}^j}$, we next apply a change of variables: $t^j = \tilde{x}^j - v^j \tilde{h}^j$ for all $j \in I_\xi$. Here

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

We replace the integration limits w^j in $S_v^{q(\xi)}$ with $W^j \equiv \frac{\tilde{x}^j - \tilde{X}^j}{\tilde{h}^j}$, where $\tilde{X}^j$ denotes the value of a random element in $R^{\tilde{q}}$. The limits of integrals change with $1 \to \tilde{x}^j - \tilde{h}^j; W^j \equiv \frac{\tilde{x}^j - \tilde{X}^j}{\tilde{h}^j} \to \tilde{X}^j$. Since the function $\tilde{\psi}(\frac{\tilde{x} - \tilde{X}}{\tilde{h}})$ is zero outside of the set

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

$\tilde{X}^j \geq \tilde{x}^j - b^j \tilde{h}^j$. We obtain

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

where $S_t^{q(\xi)} = \prod_{j \in I_\xi} [\tilde{x}^j - \tilde{h}^j, \tilde{X}^j]$. The $(-1)^{q(\xi)}$ term arising from the change of variables vanishes due to the reversal in the limits integral.

The expectation is

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

where the first equality uses linearity of the expectation operator and the second equality 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 $j \in I_\xi$

$$t^j \leq \tilde{X}^j \leq \tilde{x}^j + \tilde{h}^j$$

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

$$\begin{aligned}\int_{S_X^{\tilde{q}}(t(\xi))} dP_{\tilde{X}}(\tilde{X}) &= P_{\tilde{X}}(\prod_{j\in I_\xi}[t^j, \tilde{x}^j + \tilde{h}^j] \prod_{j\in I_\xi^c}\left[\tilde{x}^j - \tilde{h}^j, \tilde{x}^j + \tilde{h}^j\right]) \\ &= F_{\tilde{X}}(\tilde{x} - \lambda_\xi(t) \circ \tilde{h}, \tilde{x} + \tilde{h})\end{aligned}$$

where

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

For the limits of the integral with respect to $t(\xi)$, we use $\widetilde{S}_t^{q(\xi)} = \prod_{j\in I_\xi}[\tilde{x}^j - \tilde{h}^j, \tilde{x}^j + \tilde{h}^j]$. The last displayed expression (B.2) then becomes

$$\begin{aligned}&\sum_{\xi=0}^{2^{\tilde{q}}-1} \int_{\widetilde{S}_t^{q(\xi)}} F_{\tilde{X}}(\tilde{x} - \lambda_\xi(t) \circ \tilde{h}, \tilde{x} + \tilde{h}) \left(\Delta_{1,\xi}\tilde{\psi}\left(\frac{\tilde{x}^1 - t^1}{\tilde{h}^1}, \cdots, \frac{\tilde{x}^{\tilde{q}} - t^{\tilde{q}}}{\tilde{h}^{\tilde{q}}}\right)\right) \\ &\times \prod_{j\in I_\xi}\left(\tilde{h}^j\right)^{-1} dt(\xi)\end{aligned}$$

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

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

where

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

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

Therefore,

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

■

Proof of Corollary A.1.

From Lemma A.1

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

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

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

Write

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

Consider some ξ and the corresponding multivariate integral. Next, we integrate in such an integral with respect to v^{j_1} (meanwhile holding $v^{j_2}, \cdots, v^{j_{q(\xi)}}$ constant)

$$\begin{aligned} & \int_{-1}^{1} F_{\tilde{X}}(\tilde{x}-\lambda_\xi \circ \tilde{h}, \tilde{x}+\tilde{h}) \prod_{j\in I_\xi^c} \delta\left(v^j=1\right) \prod_{j\in I_\xi} \left(-\partial_j\right) \tilde{\psi}\left(v\right) dv^{j_1} \\ = & \int_{-1}^{0} F_{\tilde{X}}(\tilde{x}-\lambda_\xi \circ \tilde{h}, \tilde{x}+\tilde{h}) \prod_{j\in I_\xi^c} \delta\left(v^j=1\right) \prod_{j\in I_\xi} \left(-\partial_j\right) \tilde{\psi}\left(v\right) dv^{j_1} \quad \text{(B.3)} \\ & + \int_{0}^{1} F_{\tilde{X}}(\tilde{x}-\lambda_\xi \circ \tilde{h}, \tilde{x}+\tilde{h}) \prod_{j\in I_\xi^c} \delta\left(v^j=1\right) \prod_{j\in I_\xi} \left(-\partial_j\right) \tilde{\psi}\left(v\right) dv^{j_1}. \end{aligned}$$

In the first integral, on the rhs of (B.3) apply a change of variable $v^{j_1} = -z^{j_1}$; define $\lambda_\xi\left(j_1\right)$ 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

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

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

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

A simple change of notation in the second integral on the rhs of (B.3) (replacing v^{j_1} by z^{j_1} in v and denoting the resulting vector by $v\left(z^{j_1}\right)$) allows us to express the sum of the two integrals in (B.3) as

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

note that λ_ξ also has v_1 replaced with z_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

$$\int_{-1}^{1} \left\{ \int_{0}^{1} \left\{ F_{\tilde{X}}(\tilde{x} - \lambda_\xi \circ \tilde{h}, \tilde{x} + \tilde{h}) - F_{\tilde{X}}(\tilde{x} - \lambda_\xi (j_1) \circ \tilde{h}, \tilde{x} + \tilde{h}) \right\} \right.$$
$$\left. \prod_{j \in I_\xi^c} (\delta \left(v^j = 1 \right)) \prod_{j \in I_\xi} (-\partial_j) \tilde{\psi} \left(v \left(z^{j_1} \right) \right) dz^{j_1} \right\} dv^{j_2}.$$

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

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

note that λ_ξ also has v_2 also replaced with z_2.

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

$$\int_0^1 \cdots \int_0^1 \left\{ \underbrace{\int_{\tilde{x}_1 - \tilde{h}_1}^{\tilde{x}_1 + \tilde{h}_1} \int_{\tilde{x}_2 - \tilde{h}_2}^{\tilde{x}_2 + \tilde{h}_2} \cdots}_{j \in I_\xi^c} \underbrace{\int_{\tilde{x}_{j_1} - z_{j_1} \tilde{h}_{j_1}}^{\tilde{x}_{j_1} + z_{j_1} \tilde{h}_{j_1}} \cdots \int_{\tilde{x}_{j_{q(\xi)}} - z_{j_{q(\xi)}} \tilde{h}_{j_{q(\xi)}}}^{\tilde{x}_{j_{q(\xi)}} + z_{j_{q(\xi)}} \tilde{h}_{j_{q(\xi)}}}}_{j \in I_\xi} dF_{\tilde{X}} \right\}$$
$$\times \prod_{j \in I_\xi^c} \delta \left(v^j = 1 \right) \prod_{j \in I_\xi} (-\partial_j) \tilde{\psi} \left(v(z^{j_1}, z^{j_2}, \cdots, z^{j_{q(\xi)}}) \right) dz(\xi)$$

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 for every ξ

$$\int_0^1\cdots\int_0^1\left\{\underbrace{\int_{\tilde{x}_1-\tilde{h}_1}^{\tilde{x}_1+\tilde{h}_1}\int_{\tilde{x}_2-\tilde{h}_2}^{\tilde{x}_2+\tilde{h}_2}\cdots}_{j\in I_\xi^c}\underbrace{\int_{\tilde{x}_{j_1}-z_{j_1}\tilde{h}_{j_1}}^{\tilde{x}_{j_1}+z_{j_1}\tilde{h}_{j_1}}\cdots\int_{\tilde{x}_{j_{q(\xi)}}-z_{j_{q(\xi)}}\tilde{h}_{j_{q(\xi)}}}^{\tilde{x}_{j_{q(\xi)}}+z_{j_{q(\xi)}}\tilde{h}_{j_{q(\xi)}}}}_{j\in I_\xi}dF_{\tilde{X}}\right\}$$

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

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

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

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

This concludes the proof. ■

Proof of Lemma A.2.
Consider $\Xi^{[\tilde{q}]}$. When $\Xi^{[\tilde{q}]}=\Xi^{[q]}$, the vector $\tilde{x}=x$, but when $\Xi^{[\tilde{q}]}=\Xi^{[q]}\times\Xi^{[q]}$, $\tilde{x}=(x;x)$. Given $\tilde{x}$ the probability measure $P_{\tilde{X}}$ on $\Xi^{[\tilde{q}]}$ defines a distribution $F_{\tilde{Z}_+}$ in $R^{\tilde{q}}$ with support on $R_+^{\tilde{q}}$, a non-negative multivariate quadrant, given by the measurable mapping $\Xi^{[\tilde{q}]}\to R_+^{\tilde{q}}$ with $\tilde{X}=\left(X^1,\cdots,X^{\tilde{q}}\right)$ mapped into a random vector

$$\tilde{Z}_+=\left(\tilde{Z}_+^1,\cdots,\tilde{Z}_+^{\tilde{q}}\right);\ \tilde{Z}_+^i=\left\|\tilde{x}^i-\tilde{X}^i\right\|_i.$$

Then $\tilde{X}=\tilde{x}$ transforms into $\tilde{Z}_+=0$. The $F_{\tilde{Z}}$ measure of the cube $C(0,r)$, with $r\in R_+^{\tilde{q}}$, is concentrated in the non-negative quadrant and is given by

$$F_{\tilde{Z}_+}\left(C(0,r)\right)=P_{\tilde{X}}(C\left(\tilde{x},r\right))=P_{\tilde{X}}\left(\prod_{i=1}^{\tilde{q}}B(\left\|\tilde{x}^i-\tilde{X}^i\right\|_i\leq r^i)\right).$$

We can consider the symmetric function $\tilde{\psi}\left(\cdot\right)$ on $R^{\tilde{q}}$ with support in $[-1,1]^{\tilde{q}}$, based on the given $\tilde{\psi}_+\left(\cdot\right)$ defined as $\tilde{\psi}\left(v^1,\cdots,v^{\tilde{q}}\right)=\tilde{\psi}_+\left(\left\|v^1\right\|,\cdots,\left\|v^{\tilde{q}}\right\|\right)$. Thus the symmetric function $\tilde{\psi}\left(\frac{\tilde{x}-\tilde{X}}{\tilde{h}}\right)$ of Corollary A.1 can be written as $\tilde{\psi}\left(\frac{\tilde{Z}}{\tilde{h}}\right)$ and its values are given by $\tilde{\psi}\left(\frac{\tilde{Z}}{\tilde{h}}\right)=\tilde{\psi}_+\left(\frac{\tilde{Z}_+}{\tilde{h}}\right)$. The result from Corollary A.1 then written with $\tilde{x}=0$ and the probability measure corresponding to the distribution $F_{\tilde{Z}_+}$

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

becomes

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

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

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

as required. ■

Proof of Corollary A.2.
(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 $\lambda_{\xi}\circ\tilde{h}$ are between zero and components of $\tilde{h}$,

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

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

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

this can be bounded by

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

Recall that $F_{\tilde{X}}(\tilde{x}-\tilde{h},\tilde{x}+\tilde{h})=P_{\tilde{X}}(C(\tilde{x},\tilde{h}))$. Thus under Lemma A.1 or Lemma A.2

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

where

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

■

Proof of Lemma A.4. (a) The lower bound follows after substituting the lower bound, $C_\varepsilon P_{\tilde{X}}(C(\tilde{x},\tilde{h}))$ for $P_{\tilde{X}}(C(\tilde{x},\varepsilon\tilde{h}))$ in (A.10).
(b) The term $\tilde{\psi}(1,..,1) > 0$ appears in the moment expression for the expectation which appears for $\xi = 0$ with $I_0 = \emptyset$. 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 $dF_{\tilde{X}}$ or $dP_{\tilde{X}}$ is replaced by $d\Omega_{\tilde{X}}$ in all the derivations providing the result.
(b) The upper bound follows from the boundedness of the function $g(\tilde{X})$ around $\tilde{x}$.
(c) The lower bound in the case of positive $g(\tilde{X})$ follows; when $g(\tilde{X})$ can take non-positive values, the lower bound on the absolute value of the moment is zero.

■

C Proofs of lemmas and theorems

C.1 Some preliminaries: moments and moment bounds

The proofs of the lemmas and theorems apply the results in Appendix A.

The next lemma and its corollary provide bounds used in the proofs below, which implement the results from Appendix A.

Lemma C.1 *Given Assumption 1, 2(a-c) and 3, the moment for* $\tilde{K}\left(\frac{\tilde{x}-\tilde{X}}{\tilde{h}}\right)$, *or* $\tilde{K}\left(\frac{\|\tilde{x}-\tilde{X}\|}{\tilde{h}}\right)$ *(with* $\tilde{K}(\cdot)$ *given by either* $K(\cdot)$, *or* $K(\cdot)K(\cdot)$*) satisfies*

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

with

(a)

$$M_{EK^m} = 2^{\tilde{q}} \max_{0\leq\xi\leq 2^{\tilde{q}}-1} \int_{[0,1]^{q(\psi)}} \left|\Delta_{1,\xi}\tilde{K}^m(v)\right| dv(\xi);$$

(b) when Assumption 5 holds, for some $0 < \varepsilon < 1$

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

(c) when instead, the kernel is type I (add Assumption 2(d))

$$L_{EK^m} = \tilde{K}^m(\iota) > 0.$$

with $\iota = (1, \cdots, 1)^T$.

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.1.
(a) The upper bound is obtained by substituting $\tilde{K}^m(.)$ for $\tilde{\psi}(.)$ in Lemma A.3.
(b) The bound follows from Lemma A.4 (a).
(c) The bound follows from Lemma A.4 (b). ■

Recall, for some bounded continuous function $\tilde{g}(\tilde{X})$, we defined

$$\Omega_{\tilde{g}}(C(\tilde{x}, \tilde{h})) = \int_{C(\tilde{x}, \tilde{h})} \tilde{g}(z) dP_{\tilde{X}}(z). \tag{C.1}$$

Corollary C.1 *Under the conditions of Lemma C.1 for a bounded function* $\tilde{g}(\cdot)$ *the moment*

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

with $\Omega_{\tilde{g}}\left(C(\tilde{x}, \tilde{h} \circ v)\right)$ *given by (A.11) is bounded as*

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

where

$$L_{EK^m g} = \max\{0, \inf_{x \in C(\tilde{x}, \tilde{h})} \tilde{g}(x) L_{EK^m}\};\ M_{EK^m} = \sup_{x \in C(\tilde{x}, \tilde{h})} |\tilde{g}(x) M_{EK^m}|.$$

The corollary follows directly from Lemma C.1. Note that here the lower bound does not have to be positive.

Next, we provide preliminary results on moments that demonstrate for a strongly mixing process the leading terms of the variance of $B(x)$ and $A(x)$ are the same as under independence.

Lemma C.2 *(Preliminary for moments) Under either of the following sets of conditions (i) Assumptions 1-4 and 6 or (ii) Assumptions 1, 2(a-c), 3-6*

$$\begin{aligned} L_{EK}P_X\left(C\left(x,h\right)\right) &\leq EB\left(x\right) \leq M_{EK}P_X\left(C\left(x,h\right)\right); \qquad \text{(C.2)} \\ varB\left(x\right) &= \frac{1}{n}E\left[K^2\left(\frac{x-X}{h}\right)\right]\left(1+o\left(1\right)\right) \\ &\leq \frac{1}{n}M_{varK}P_X\left(C\left(x,h\right)\right) \qquad \text{(C.3)} \end{aligned}$$

and

$$\begin{aligned} EA\left(x\right) &= E\left[m\left(X\right)K\left(\frac{x-X}{h}\right)\right] \leq M_{EmK}P_X\left(C\left(x,h\right)\right); \qquad \text{(C.4)} \\ varA\left(x\right) &= \frac{1}{n}E\left[K^2\left(\frac{x-X}{h}\right)\left(\mu_2\left(X\right)\right)+m\left(X\right)^2\right)\right]\left(1+o\left(1\right)\right) \\ &\leq \frac{1}{n}M_{varA}P_X\left(C\left(x,h\right)\right). \qquad \text{(C.5)} \end{aligned}$$

Proof of Lemma C.2

The results for the first moments follow immediately from Lemma C.1. and Corollary C.1. To establish the results for the variances, consider a generic random variable Q_i, that could be either $Q_i\left(1\right) = K\left(\frac{x-X_i}{h}\right)$, or $Q_i\left(2\right) = K\left(\frac{x-X_i}{h}\right)Y_i$. Our assumptions imply that $\{Q_i\left(\cdot\right)\}$ is a stationary mixing sequence with the mixing coefficient $\alpha\left(s\right)$. We show

$$E\left|Q\left(\cdot\right)\right|^{2+\zeta} \leq M_{|Q|^{2+\zeta}}P_X\left(C\left(x,h\right)\right). \qquad \text{(C.6)}$$

Indeed for $Q_i\left(2\right)$

$$E\left|Q_i\left(2\right)\right|^{2+\zeta} = E\left|K\left(\frac{x-X_i}{h}\right)Y_i\right|^{2+\zeta} \leq E\left(\left|K\left(\frac{x-X_i}{h}\right)\right|^{2+\zeta}\right)E\left(\left|Y_i\right|^{2+\zeta}\right),$$

where by Assumption 3(c) $E\left(\left|Y_i\right|^{2+\zeta}\right) < \infty$. For $Q_i\left(1\right)$ $(C.6)$ follows immediately.

Write

$$\begin{aligned} var\left(\frac{1}{n}\sum_{i=1}^{n}Q_i\right) &= \frac{1}{n^2}\sum_{i=1}^{n}varQ_i + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{s=1}^{n-i}cov(Q_i,Q_{i+s}) \\ &= \frac{1}{n}varQ_1 + \frac{2}{n^2}\sum_{s=1}^{n-1}\left(n-s\right)cov(Q_1,Q_{1+s}). \end{aligned}$$

By the results from Lemma C.1. and Corollary C.1 we have that

$$0 < \frac{1}{n}L_{varQ}P_X\left(C\left(x,h\right)\right) < \frac{1}{n}varQ_1 < \frac{1}{n}M_{varQ}P_X\left(C\left(x,h\right)\right),$$

where the positivity of the lower bound follows from the assumptions (including $\mu_2\left(x\right) > 0$). Thus $\frac{1}{n}varQ_1 \approx O\left(\frac{1}{n}P_X\left(C\left(x,h\right)\right)\right)$.

It remains to show that the expression with the covariances goes to zero faster. Following the usual approach (e.g., as in Masry, 2005) consider some integer u_n between 1 and n and use the partitioned sum

$$\frac{1}{n^2}\left(\sum_{s=1}^{u_n-1}(n-s)\,cov(Q_1,Q_{1+s})+\sum_{s=u_n}^{n-1}(n-s)\,cov(Q_1,Q_{1+s})\right)=E_1+E_2$$

Using Assumption 6, E_1 can be bounded

$$E_1\leq\frac{u_n}{n}M_{varQ}P_{\tilde{X}}\left(C\left(x,h\right)\times C\left(x,h\right)\right)\leq\frac{u_n}{n}M_{varQ}M_{FF}P_X(C(x,h))^2,$$

which by appropriate selection of u_n specified below is of smaller order. To bound E_2 we utilize Davydov's Lemma (1968) which provides

$$|cov(Q_1,Q_{1+s})|\leq 8\left(E\left|Q\right|^{2+\zeta}\right)^{\frac{2}{2+\zeta}}s^{-\kappa\frac{\zeta}{2+\zeta}}.$$

As E_2 can then be bounded by a geometric progression, is is dominated by $|n\;cov(Q_1,Q_{1+u_n})|$ and hence

$$\begin{aligned}E_2 &\leq nu_n^{-k\frac{\zeta}{2+\zeta}}8\left(E\left|Q\right|^{2+\zeta}\right)^{\frac{2}{2+\zeta}}\\ &\leq M_{covQ}nu_n^{-k\frac{\zeta}{2+\zeta}}P_X\left(C\left(x,h\right)\right)^{\frac{2}{2+\zeta}}.\end{aligned}$$

By setting $u_n=\left[P_X\left(C\left(x,h\right)\right)\right]^{-\frac{2(2+\zeta)}{\kappa\zeta}}$ we can bound the expression with the covariances by

$$\begin{aligned}&\frac{2}{n^2}\sum_{s=1}^{n-1}(n-s)\,cov(Q_1,Q_{1+s})\\ \leq\;& M\left\{\frac{1}{n}\left[P_X\left(C\left(x,h\right)\right)\right]^{2-\frac{2(2+\zeta)}{\kappa\zeta}}+\frac{1}{n}P_X\left(C\left(x,h\right)\right)^{\frac{2}{2+\zeta}+2}\right\}\end{aligned}$$

and since $2-\frac{2(2+\zeta)}{\kappa\zeta}>1$ the first component is $o\left(\frac{1}{n}P_X\left(C\left(x,h\right)\right)\right)$; the second component is of smaller order. ■

C.2 Proof of Lemma from Section 2

Proof of Lemma 1.
(a) Given (9) with $\tau_0\left(\varepsilon\right)=C\left(\varepsilon\right)>0$ we have that for small enough h

$$P_X\left(B\left(x,\varepsilon h\right)\right)>C_\varepsilon P_X\left(B\left(x,h\right)\right),$$

where $\tilde{C}=C\left(\varepsilon\right)+\delta$, $\delta>0$ and (11) holds.

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

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

where $\tilde{s}<1$. From (10) it follows that $P_X\left(B\left(x,\tilde{s}v\right)\right)v>CvP_X\left(B\left(x,v\right)\right)$, thus for $\varepsilon=\tilde{s}$ (11) holds. ■

C.3 Proofs of the main results

Here we prove the asymptotic normality results of the Nadaraya-Watson (NW) estimator, $\widehat{m}(x)$ defined as:

$$\begin{aligned}\widehat{m}(x) &= B_n^{-1}(x) A_n(x), \\ B_n(x) &= \frac{1}{n}\sum_{i=1}^{n} K(W_i(x)); \; A_n(x) = \frac{1}{n}\sum_{i=1}^{n} K(W_i(x)) Y_i.\end{aligned}$$

The proofs 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 form of two lemmas and a proposition together with their proofs.

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

Lemma C.3 *Under either of the following conditions (i) Assumptions 1-4 or (ii) 1, 2(a-c), 3-5*
(a)

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

(b)

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

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

Proof of Lemma C.3. (a) follows immediately by making use of the lower and upper bounds in $(C.2)$ and the upper bound from $(C.3)$ which provide

$$\frac{B_n(x) - EB_n(x)}{EB_n(x)} = O_p\left(\left(\frac{1}{nP_X(C(x,h))}\right)^{1/2}\right)$$

which is $o(1)$.
(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 $o_p(1)$. ■

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. Applying the definition in (C.1), we define $\Omega_m(C(x,h))$ for $g(x) = m(x)$ and $\Omega_{m^2+\mu_2}(C(x,h))$ for $g(x) = m^2(x)+\mu_2(x)$. The next lemma then expresses the moments for the numerator of the estimator at x.

Lemma C.4 *Under the conditions of Lemma C.3*

(a)

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

(b)

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

Proof of Lemma C.4. The moments are established similarly to $(C.4), (C.5)$. (a) Using the expressions from Corollary C.1

$$\begin{aligned} E\zeta_{in} &= E(E(\zeta_{in}|X_i)) \qquad (C.7)\\ &= E\left[K\left(\frac{x-X_i}{h}\right)(m(X_i) + E(u_i|X_i))\right] \\ &= \int K\left(\frac{x-X}{h}\right) m(X)\,dF_X \qquad (C.8)\\ &= \int_{[0,1]^q} \Omega_m(C(x, h\circ v))(-1)^q \partial K(v)\,dv, \end{aligned}$$

where we use $E(u|X) = 0$ and the definition of $\Omega_m(\cdot,\cdot)$.

(b) Similarly to (a)

$$\begin{aligned} E\zeta_{in}^2 &= E\left(E\left[K^2\left(\frac{x-X_i}{h}\right)(m(X_i)+u_i)^2|X_i\right]\right) \qquad (C.9)\\ &= \int K^2\left(\frac{x-X}{h}\right)\left(m^2(X)+\mu_2(X)\right)dF_X \\ &= \int_{[0,1]^q} \Omega_{m^2+\mu_2}(C(x, h\circ v))(-1)^q \partial K^2(v)\,dv. \end{aligned}$$

■

Next, define

$$\widetilde{\zeta}_{in} = \frac{1}{\sqrt{n}}\frac{\zeta_{in} - E\zeta_{in}}{\sqrt{var\zeta_{in}}}. \qquad (C.10)$$

Proposition C.1 *Under the conditions of Lemma C.4 and Assumption 6* $n \to \infty$, $h \to 0$

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

Proof of Proposition C.1.

Our proof proceeds identically to the Proof of Theorem 4 in Masry (2005, pp.172-176). For convenience of the reader we list the main steps in the notation of our paper: we replace the $\tilde{Z}_{ni}$ of that paper by scaled $\sqrt{n}\widetilde{\zeta}_{in}$ from (C.10) and denote $\sqrt{n}\widetilde{\zeta}_{in}$ by V_{in}. We consider $S_n = \sum_{i=1}^{n} V_{in}$ in place of their $\sum_{i=1}^{n} \tilde{Z}_{in}$; our standardization implies that for us $\sigma^2(x) = 1$. Thus we need to show

$$\frac{1}{\sqrt{n}} S_n \to^L N(0,1).$$

We follow Masry (2005) in partitioning $\{1, \cdots, n\}$ into alternating big blocks of size u_n and small blocks of size v_n, respectively. The block sizes satisfy the following conditions as $n \to \infty$:

(a) $\{v_n\} : v_n \to \infty$, but $v_n = o\left(\left(nP_X(x,h)\right)^{1/2}\right); \left(\frac{n}{P_X(x,h)}\right)^{1/2} \alpha(v_n) \to 0$. This implies that there exists a sequence of positive integers $\{q_n\}$ such that $q_n \to \infty$ but $q_n v_n = o\left(\left(nP_X(x,h)\right)^{1/2}\right)$ and $q_n \left(\frac{n}{P_X(x,h)}\right)^{1/2} \alpha(v_n) \to 0$.

(b) $\{u_n\} : u_n \to \infty$, $u_n = \lfloor \left(nP_X(x,h)\right)^{1/2} / q_n \rfloor$, where $\lfloor . \rfloor$ stands for integer value here and below.

Next, denote by η_j the jth sum over a large block:

$$\eta_j = \sum_{i=j(u+v)+1}^{j(u+v)+u} V_{in};$$

by ξ_j the jth sum over a small block:

$$\xi_j = \sum_{i=j(u+v)+u+1}^{(j+1)(u+v)} V_{in}.$$

Here j varies from 0 to $k-1$ with $k = \lfloor \frac{n}{u_n + v_n} \rfloor$. Finally, after k large blocks, followed by a small block each, there may be a remainder part in the total sum:

$$\zeta_k = \sum_{i=k(u+v)+1}^{n} V_{in}.$$

The sum S_n is represented as

$$\begin{aligned} S_n &= S_n^I + S_n^{II} + S_n^{III} \\ S_n^I &= \sum_{j=0}^{k-1} \eta_j;\ S_n^{II} = \sum_{j=0}^{k-1} \xi_j;\ S_n^{III} = \zeta_k. \end{aligned}$$

Identically to Masry we can show that

$$\frac{1}{n}E\left(S_n^{II}\right)^2 \to 0;$$
$$\frac{1}{n}E\left(S_n^{III}\right)^2 \to 0;$$
$$E\left(\exp(itn^{-1/2}S_n^I\right) - \prod_{j=0}^{k-1} E\left(\exp(itn^{-1/2}\eta_j\right) \to 0;$$
$$\frac{1}{n}\sum_{j=0}^{k-1} E\eta_j^2 \to 1 \quad ; \frac{1}{n}\sum_{j=0}^{k-1} E\left[\eta_j^2 I\left\{\left|\eta_j\right| > \varepsilon\sqrt{n}\right\}\right] \to 0$$

for any $\varepsilon > 0$. The results require similar evaluation of second moments, using the conditions on the relative sizes of the blocks in the sums and the strong mixing condition; for the last limit Masry (2005) employs a truncation argument to deal with the fact that the response variable Y_i is not necessarily bounded. The limits show that only S_n^I matters in the limit and that it satisfies the Lindeberg-Feller theorem. ■

Proof of Theorem 1.

The asymptotic normality result in Proposition C.1 provides the 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. ■

Proof of Theorem 2.

The result follows on combining Theorem 1 with the evaluation of the bias $\widehat{m}(x)$, $E\left(\frac{A_n(x)}{B_n(x)}\right) - m(x)$. In the product space $\Xi = \Xi^{[q]}$ the following lemma applies. To show part (a) we prove the following Lemma after which part (b) follows.

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

Proof of Lemma C.5.

We make use of the following expansion of $\frac{1}{B}$ around $\frac{1}{EB}$:

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

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

$$\begin{aligned}\text{Bias}(\widehat{m}(x)) &= E\left(\frac{A_n(x)}{B_n(x)} - m(x)\right) \\ &= E\frac{A_n}{EB_n}\left(1 - \frac{(B_n(x) - EB_n(x))}{EB_n(x)} + \cdots\right) - m(x) \\ &= T_1 + T_2;\end{aligned}$$

with

$$\begin{aligned}T_1(x) &= \frac{1}{E(B_n(x))}\left(EA_n(x) - m(x)E(B_n(x))\right); \\ T_2(x) &= \frac{E\left[A_n(x)\left(1 - \frac{(B_n(x)-EB_n(x))}{EB_n(x)} + \cdots\right)\right]}{EB_n}.\end{aligned}$$

Then substituting the expectations

$$T_1(x) = \frac{\int K\left(\frac{x-X}{h}\right)(m(X) - m(x))dF_X}{\int K\left(\frac{x-X}{h}\right)dF_X}$$

and by Assumption 4 and moment bounds

$$|T_1(x)| \leq M_{\Delta m}\bar{h}^{\delta}\frac{M_{EK}}{L_{EK}} = O\left(\bar{h}^{\delta}\right).$$

For $T_2(x)$, since $E\left(\frac{B_n(x)-EB_n(x)}{EB_n(x)}\right)^2 = o(1)$ and using the geometric series expansion, we have that

$$\begin{aligned}|T_2(x)| &\leq \frac{\left(EA_n^2(x)E\left(1 - \frac{(B_n(x)-EB_n(x))}{EB_n(x)} + \cdots\right)^2\right)^{1/2}}{EB_n(x)} \\ &\leq \frac{\left(EA_n^2(x)\right)^{1/2}}{E(B_n(x))}(1 + o(1)).\end{aligned}$$

Using the preliminary Lemma C.2 we have that $E\left(A_n^2(x)\right) = varA_n(x)(1 + O(1))$ and is thus bounded by $\frac{1}{n}M_{varA}P(C(x,h))$, thus the overall bound for $|T_2|$ is

$$|T_2(x)| \leq \frac{M_{varA}}{L_{EB}}\frac{1}{\sqrt{nP(C(x,h))}}(1 + o(1)) = o(1),$$

since by the conditions of the Theorem $nP(C(x,h)) \to \infty$. Combining these results shows the bias is $O\left(\bar{h}^{\delta}\right)$. ■

Proof of Theorem 3
The result follows by substituting the limits in Assumption 7 into the result of Theorem 2. ■

Acknowledgements: The authors would like to thank the participants at the Econometric Study Group conference in Bristol and the Canadian Econometric Study Group conference for their comments. We thank Jeffrey Racine for his discussion and valuable suggestions at the CESG 2023 and Sid Kankanala for insightful comments on earlier versions of the paper.

Funding: This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant 253139.

References

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

[2] Ackerberg, D.A., K. Caves, and G. Frazer (2015) "Identification properties of recent production function estimators," *Econometrics*, **83**, 2411–2451.

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

[4] Caldeira J.F., R. Gupta, H.S. Torrent (2020) "Forecasting U.S. Aggregate Stock Market Excess Return: Do Functional Data Analysis Add Economic Value?" *Mathematics*. 2020; **8(11)**:2042. https://doi.org/10.3390/math8112042 .

[5] 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.

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

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

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

[9] Davydov, Y.A. (1968) "Convergence of distributions generated by stationary stochastic processes," *Theory of Probability & its Applications*, **13** (4), 691–696.

[10] Delsol, L. (2009) "Advances on asymptotic normality in non-parametric functional time series analysis," *Statistics*, **43**, 13–33.

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

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

[13] Ferraty, F., A. Laksaci, A. Tadj, and P. Vieu (2010) "Rate of uniform consistency for nonparametric estimates with functional variables," *Journal of Statistical Planning and Inference*, **140**, 335–352.

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

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

[16] Ferraty, F., I. van Keilegom, and P. Vieu (2010) "On the validity of the bootstrap in non-parametric functional regression," *Scandinavian Journal of Statistics*, **37**, 286–306.

[17] Ferraty, F. and P. Vieu (2004) "Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination," *Journal of Nonparametric Statistics*, **16**, 111–125.

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

[19] Gasser, T., P. Hall, and B. Presnell (1998) "Nonparametric estimation of the mode of a distribution of random curves," *Journal of Royal Statistical Society, Series B*, **60**, 681–691.

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

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

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

[23] Hall, P., Q. Li, and J.S. Racine (2007) "Nonparametric estimation of regression functions in the presence of irrelevant regressors," *The Review of Economics and Statistics*, **89**, 784–789.

[24] Hall, P. and J.S. Racine (2015) "Infinite order cross-validated local polynomial regression," *Journal of Econometrics*, **185**, 510-525.

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

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

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

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

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

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

[31] Jun, B and H. Song (2019) "Tests for Detercting Probability Mass Points," *The Korean Economic Review*, **35**, 205–248.

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

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

[34] Kotlyarova, Y, M. 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.

[35] Kurisu, D., T. Otsu, and M. Xu (2023) "Nonparametric causal inference with functional covariate," Sticerd Discussion paper.

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

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

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

[39] Mandlebrot, B. (1997) Fractals and Scaling in Finance, Discontinuity, Concentration, Risk. Selecta Volume E Springer.

[40] Masry, E. (2005) Nonparametric regression estimation for dependent functional data: asymptotic normality, Stochastic Processes and their Applications,Volume 115, Issue 1, Pages 155–177,

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

[42] Olson, C.A. (1998) "A comparison of parametric and semiparametric estimates of the effect of spousal health insurance coverage on weekly hours worked by wifes," *Journal of Applied Econometrics*, **13**, 543–565.

[43] Pollard, D. (2001) "A User's Guide to Measure Theoretic Probability," Cambridge series in Statistical and Probabilistic Mathematics.

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

[45] Ramsay, J. O. and B.W. Silverman (2005) "Functional Data Analysis," Springer, New York.

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

[47] Vol'berg, A.L. and S.V. Konyagin (1988) "On measures with the doubling condition," *Mathematics of the USSR Izvestiya,* **30**, 629–638.

[48] Wang, M.-C. and J. van Ryzin (1981) "A class of smooth estimators for discrete distributions," *Biometrika*, **68**, 301–309.

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

ANNEX with Figures and Tables

Figure 1: Distribution of conditioning variables (point mass and high derivatives).

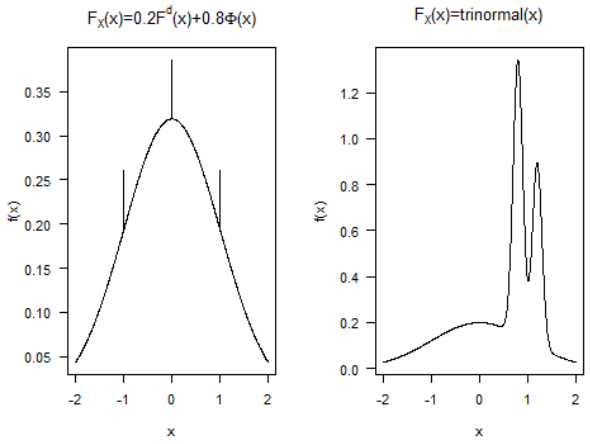


Figure 2: Local RMSE NW estimator by distribution of conditioning variables (point mass and high derivatives), n=1,000.

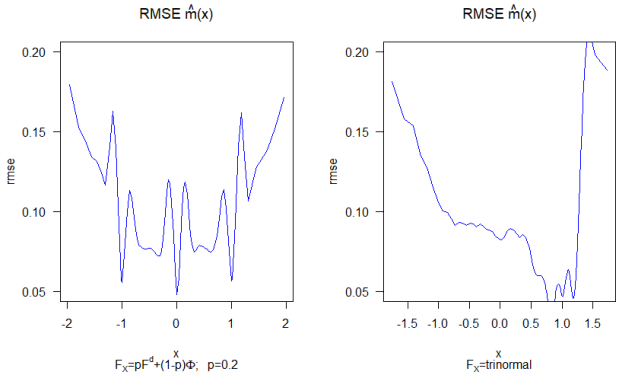


Table 1: Empirical rate of convergence (point mass and high derivatives).

$F_X(x)=0.2F^d(x)+0.8\Phi(x)$

X	Oracle	NW	Ratio
0.00	-0.498	-0.470	0.944
0.10	-0.498	-0.203	0.408
0.20	-0.498	-0.384	0.770
0.40	-0.499	-0.475	0.953

$F_X(x)$=trinormal(x)

X	Oracle	NW	Ratio
0.00	-0.511	-0.449	0.878
0.50	-0.492	-0.381	0.774
0.75	-0.492	-0.416	0.844
1.00	-0.497	-0.413	0.831

Figure 3: Local RMSE NW and SI estimator.

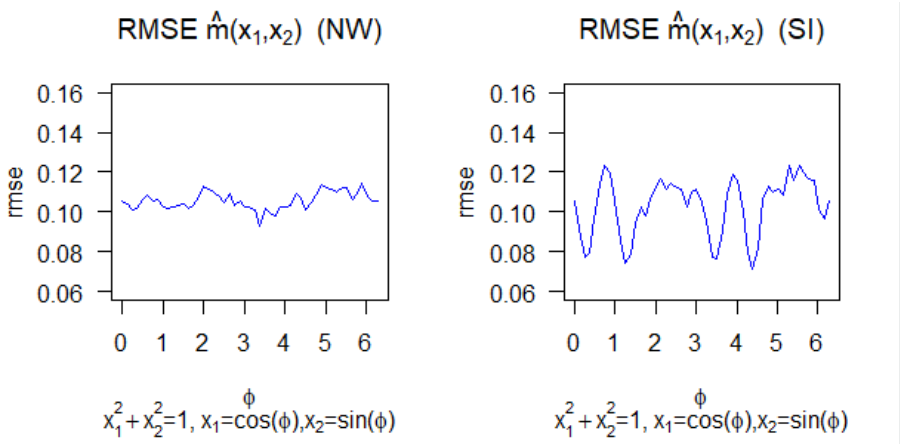


Table 2 Comparison of the rate of the NW and SI estimator to the "Oracle" (True model) estimator.

phi	Oracle	NW	Ratio	SI	Ratio
0.00	-0.509	-0.444	0.872	-0.383	0.752
0.25	-0.505	-0.448	0.887	-0.362	0.716
0.50	-0.510	-0.448	0.879	-0.365	0.715
1.00	-0.515	-0.451	0.875	-0.393	0.764
1.50	-0.499	-0.465	0.932	-0.375	0.752
1.75	-0.507	-0.454	0.897	-0.446	0.880

Table 3 Comparison of the rate of the NW.c, NW.d, and SI estimator to the "Oracle" estimator.

X_1	X_2	d(k)	Oracle	NW.c	Ratio	NW.d	Ratio	SI	Ratio
1.5	2.5	4	-0.531	-0.451	0.849	-0.422	0.795	-0.398	0.750
2.0	2.0	4	-0.517	-0.442	0.856	-0.429	0.840	-0.344	0.665
2.5	1.5	4	-0.501	-0.437	0.872	-0.406	0.809	-0.394	0.786
1.5	4.5	6	-0.515	-0.458	0.890	-0.410	0.797	-0.407	0.791
2.0	4.0	6	-0.510	-0.395	0.776	-0.395	0.776	-0.381	0.747
2.5	3.5	6	-0.503	-0.452	0.898	-0.430	0.854	-0.451	0.896
1.5	5.5	7	-0.511	-0.447	0.875	-0.392	0.767	-0.427	0.836
2.0	5.0	7	-0.510	-0.517	1.012	-0.412	0.808	-0.389	0.763
2.5	4.5	7	-0.508	-0.474	0.931	-0.431	0.848	-0.518	1.019

Table 4 RMSE of NW estimator in the presence of functional regressor X_1 at cross validated bandwidth, n=250.

	X_2 =N(0,1)	$X_2=m_1(Z)$	
		ρ =0.0	**ρ=0.8**
In-sample			
RMSE	0.746	0.915	1.058
Misspecification:			
$RMSE_1$	1.140	1.918	2.099
$RMSE_2$	1.833	1.854	1.746
Out-of-sample			
RMSE	0.915(4)	1.026(19)	1.210(18)

Note: $RMSE_1$ stands for the RMSE obtained when ignoring the X_2 regressor and $RMSE_2$ stands for the RMSE achieved when ignoring the X_1 regressor. The number in brackets indicates the number of simulations (out of 500) where at the cross-validation bandwidth no neighbour to the out-of-sample observation exists.

Figure 4: Pointwise RMSE of NW estimator in presence of functional regressor at cross-validated bandwidth, n=250 (50 out-of-sample observations) ; the true value of m(x) of out-of-sample observations on the horizontal axis

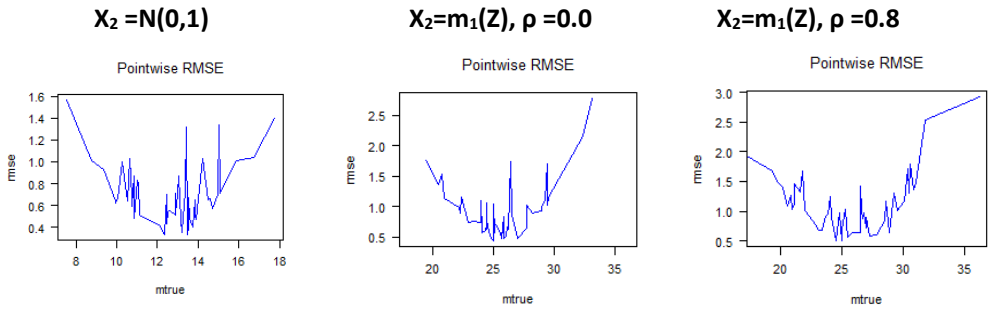
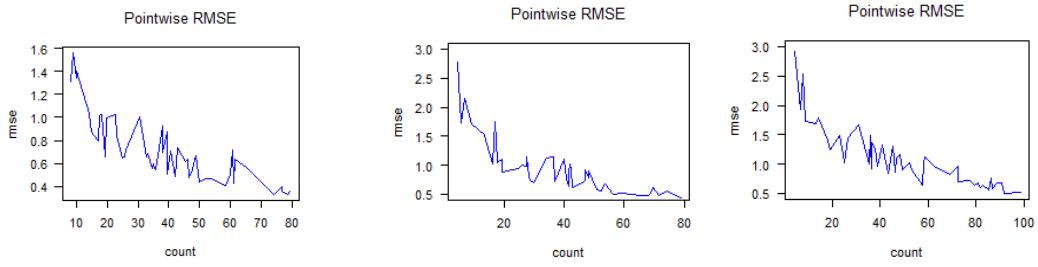


Figure 5: RMSE at cross-validated bandwidth as a function of the number of neighbors, n=250 (out-of-sample).



D Simulations Supplement

In this supplement we provide more details of the simulation and insights they provide. Simulations were conducted in R, and make use of the NP package in R (Hayfield and Racine, 2008) where suitable.

D.1 Univariate (Point mass example)

The conditional mean function under consideration in Section 5.1 is displayed in Figure D.1.

Include Figure D.1

Various results are provided here for the point mass distribution for different values of p, in particular $p = 0, 0.01$, and $p = 0.1$, to allow us to see the impact of changing the probability of mass points.

From Table D.1, we observe that the mean cross-validated bandwidth across the 500 simulation show a decrease in the presence of mass points.

Include Table D.1

The bandwidth under the trinormal conditioning distribution is comparable to that of the distribution with mass points with $p = 0.2$, which is why we focussed on this setting in the main text.

The overall performance indicators of the NW estimator (mean absolute error (MAE) and root mean squared error (RMSE)) reveal an improved aggregate performance arising from the presence of mass points, as does an increased sample size and larger signal to noise ratio. Details hereof are provided in Table D.2.

Include Table D.2

The MAE typically is smaller than the RMSE which is a reflection of the variability associated with the NW estimator across simulations.

In Figure 2, we evaluated the NW nonparametric fit locally at the cross validated bandwidth when $n = 1,000$. In order to assess the sensitivity of the results to the selected bandwidth, we provide here the results for different bandwidth choices with the results that apply to the cross-validated bandwidth (h_{cv}) displayed in red. The RMSE results, together with the bias and standard deviation, are presented in Figure D.2.

Include Figure D.2

As expected, smaller bandwidths result in a smaller bias and larger variance at all points, where the standard deviation increases at points where the distribution is more sparse. The cross-validated bandwidth provides reasonable performance in that this bias is small.

The impact of increasing the probability of mass points is marked. While, as expected, the bias is close to zero at the mass points (where a parametric rate of convergence can be obtained), when evaluating points slightly above/below

these mass points its impact is non-negligible. The impact on the standard deviation around the point mass points also becomes more pronounced.

Interesting patterns are also observed when we consider the trinormal mixture. In regions where the derivative of the trinormal mixture is large, small standard deviations are observed as well as larger fluctuations in this bias; the bias and standard deviation are fairly stable when x takes values from [-1,0.5] where the distribution is not sparse and does not have large derivatives, while the standard deviation increases again where the distribution is more sparse.

Finally, in Table D.3 we provide a more extensive set of results regarding the pointwise convergence rates.

Include Table D.3

The results obtained for different probability of mass point is similar, the pointwise convergence rates estimates close to masspoints as expected improves when a smaller bandwidth is chosen.

D.2 Bivariate (with singularity: X_1 and X_2 on unit circle)

The conditional mean function under consideration in Section 5.2 is displayed in Figure D.3.

Include Figure D.3

The display on the left uses $\phi \in [0, 2\pi]$ on the horizontal axis with $(X_1, X_2) = (\cos(\phi), \sin(\phi))$; the display on the right has $X_1 \in [-1, 1]$ on the horizontal axis.

From Table D.4 we can see that the leave-one-out cross validated bandwidths for the NW local constant estimator are comparable for both arguments X_1 and X_2 given its symmetric formulation.

Include Table D.4

The estimate on X_2 in the single index, $b_1 X_1 + b_2 X_2$, with b_1 normalized to 1, on average is close to 1 as expected.

The overall performance indicators of the NW estimator (mean absolute error (MAE) and root mean squared error (RMSE)) are very similar for the NW kernel regression estimator and the single index (SI) estimator. Using the cross-validated bandwidth the MAE (RMSE) equals 0.801 (1.003) for the NW kernel estimator against 0.800 (1.002) for the single index estimator when $n = 1,000$. With a small sample size ($n = 100$) the single index model shows a slightly larger improvement vis-a-vis the NW kernel regression estimator. Nevertheless, the benefits of the reduced dimension achieved by the single index model is appropriately reflected in the NW kernel regression estimator. Both aggregate criteria improve when the signal to noise ratio is larger or the sample size increases. Details hereof are provided in Table D.6.

Include Table D.5

In Figure D.4 we present the average RMSE over the 500 simulations, together with the bias and standard deviation, for the NW and SI estimator when

$n = 1,000$. The graphs reveal the sensitivity to the bandwidth choice. The standard deviation of the estimates show a much larger 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.

Include Figure D.4

In Table D.6 we present the results of local empirical rate of convergence in the setting where we do not have the singularity. Here X_1 and X_2 are drawn independently on $U[-1, 1]$ instead.

Include Table D.6

D.3 Bivariate (with mass of (X_1, X_2) on 1-dimensional line)

The conditional mean function under consideration in Section 5.3 is displayed in Figure D.5.

Include Figure D.5

In Table D.7, the overall performance indicators (MAE and MSE) are given for the NW and SI estimators at different sample sizes. For the ordered discrete variable in NW.d, the geometric weights

$$W(h, i, j) = \begin{cases} \frac{1}{2}(1-h)h^{|i-j|} & |i-j| \geq 1 \\ 1-h & i = j, h \in [0, 1] \end{cases}$$

are used where i and j denote values taken by the ordered discrete variable (see also Wang et al., 1981). The overall performance of the NW.d estimator suggest a slight improvement over NW.c for small samples. Nevertheless, the results are comparable to those obtained when implementing the single index model that uses $(\log(X_1), \log(X_2))$ as the conditioning variables (oracle). At the cross validated bandwidth, $n = 500$, and $srn = 1$, the MAE on average equals 0.354, 0.351, and 0.352 for the NP.c, NP.d, and SI approach respectively; the RMSE respectively 0.443, 0.440, and 0.441.

Include Table D.7

In Figure D.6, we present the bias, standard deviation and RMSE of the NW.c/NW.d and SI estimators over a grid of values for x_1 separately for each sub-population. For the NW estimators, the standard deviation and RMSE reveal patterns across the range of X_1 which is not dissimilar for different values of $d(k)$ and reveal increases at the boundary values with X_1 taking values close to 1 or 3. For the SI estimator, the standard deviation and RMSE pattern over X_1 shows more dissimilarities depending on the value of $d(k)$.

Include Figure D.6

D.4 Bivariate (in the presence of a functional regressor)

In Figure D.7 (a) we display a random sample of 20 random curves of the functional regressor X_1 (and Z) used in Section 5.4 and in (b) we display the associated $m_1(X_1)$ realizations used in defining the conditional mean.

Include Figure D.7

The cross-validated bandwidths for $h_{1,cv}$ and $h_{2,cv}$ were obtained using a grid search on $[h_{1,low}, h_{1,up}] \times [h_{2,low}, h_{2,up}]$; we set the lower bound equal to the minimal distance ensuring each regressor has at least one neighboring observation and the upper bound equals half the maximum distance for each regressor, $j = 1, 2$.

In Table D.8 we present details of the cross-validated bandwidths obtained for the models considered in Section 5.4. The columns specify the different additional control X_2 considered, either standard normal, or $m_1(Z)$ where Z is a functional regressor like X_1 with correlation of its parameters (a, b, w) given by ρ. The attained residual sum of squares achieved is around 90% of the sum of squares residual of the Oracle OLS. Ignoring either the functional or scalar regressor when implementing the NW estimator deteriorates the residual sum of squares. When X_2 is given by an independently drawn N(0,1)regressor, the loss of ignoring X_2 is more severe than dropping the functional regressor (1.616 versus 1.170). When using $X_2 = m_1(Z)$ as additional regressor, the loss of dropping either regressor is stronger when $\rho = 0$ than $\rho = 0.8$ as expected, in particular the loss of dropping the additional control drop from 1.377 to 1.178 when ρ increases. The impact on the residual sum of squares of dropping either is similar in this setting.

Include Table D.8

Table D.9 presents the aggregate performance of the NW estimator in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Their performance is shown to be comparable to their out-of-sample performance. Here we consider 50 out of sample observations drawn from the same DGP (held constant across simulations). Unlike Ferraty et al. (2007), we apply cross-validated bandwidth $h_{1,cv}$ and $h_{2,cv}$ instead of defining bandwidths in terms of nearest neighbors. While the latter would ensure we can always obtain the NW estimator for given number of neighbors determined by cross-validation, here our cross validated bandwidth does result in a few settings where for at least one of the randomly drawn out-of-sample observations no neighbor exist in one or more of simulations. These out-of-sample observations will be reflective of settings where the density (as represented by observed neighbors given cross-validated bandwidth) is small. This is clear when looking at the pointwise performance of the NW estimator presented next.

Include Table D.9

In Figure D.8 we present for the 50 out-of-sample observations the pointwise bias, standard deviation and RMSE of the NW estimator by bandwidth. For

each point, simulations which did not have any neighbors to the out-of-sample observation were excluded. On the horizontal axis the true conditioning expectation $m_1(X_1) + X_2$ for the out-of-sample observations is shown. In blue the results are presented at the cross-validated bandwidth. The green and red lines apply 0.9hcv and 1.1hcv respectively. As expected, smaller bandwidths result in a smaller bias and larger variance at all points. The standard deviation and rmse is large where the density for a out-of-sample observation, as represented by the number of neighbors, is small.

Include Figure D.8

ANNEX with Figures and Tables

Figure D.1: Univariate conditional mean.

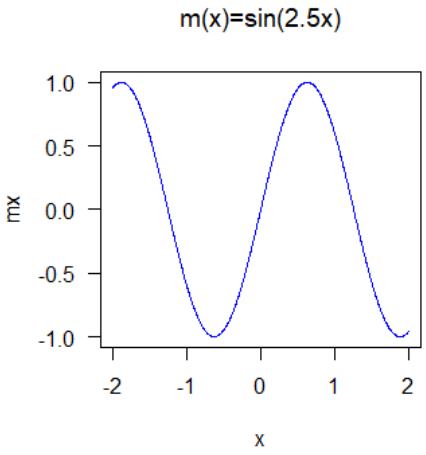


Table D.1: Some details of cross validated bandwidth for NW kernel estimator by conditioning distribution, sample size (n), and signal to noise ratio (snr).

	snr=1					
	n=100		**n= 500**		**n=1,000**	
	Mean	Std	Mean	Std	Mean	Std
F_{mass} (p=0)	0.393	0.123	0.305	0.069	0.267	0.057
F_{mass} (p=0.05)	0.393	0.121	0.291	0.068	0.257	0.054
F_{mass} (p=0.10)	0.380	0.118	0.278	0.070	0.243	0.050
F_{mass} (p=0.20)	0.261	0.115	0.250	0.067	0.222	0.050
$F_{mixtureN}$	0.348	0.112	0.256	0.065	0.216	0.054

	snr=2					
	n=100		**n= 500**		**n=1,000**	
	Mean	Std	Mean	Std	Mean	Std
F_{mass} (p=0)	0.351	0.101	0.265	0.061	0.236	0.046
F_{mass} (p=0.05)	0.341	0.099	0.252	0.060	0.224	0.044
F_{mass} (p=0.10)	0.330	0.093	0.237	0.059	0.212	0.044
F_{mass} (p=0.20)	0.310	0.088	0.211	0.054	0.189	0.041
$F_{mixtureN}$	0.304	0.092	0.218	0.054	0.178	0.044

Table D.2: Some details of aggregate performance (MAE and RMSE) by conditioning distribution, sample size (n), and signal to noise ratio (snr).

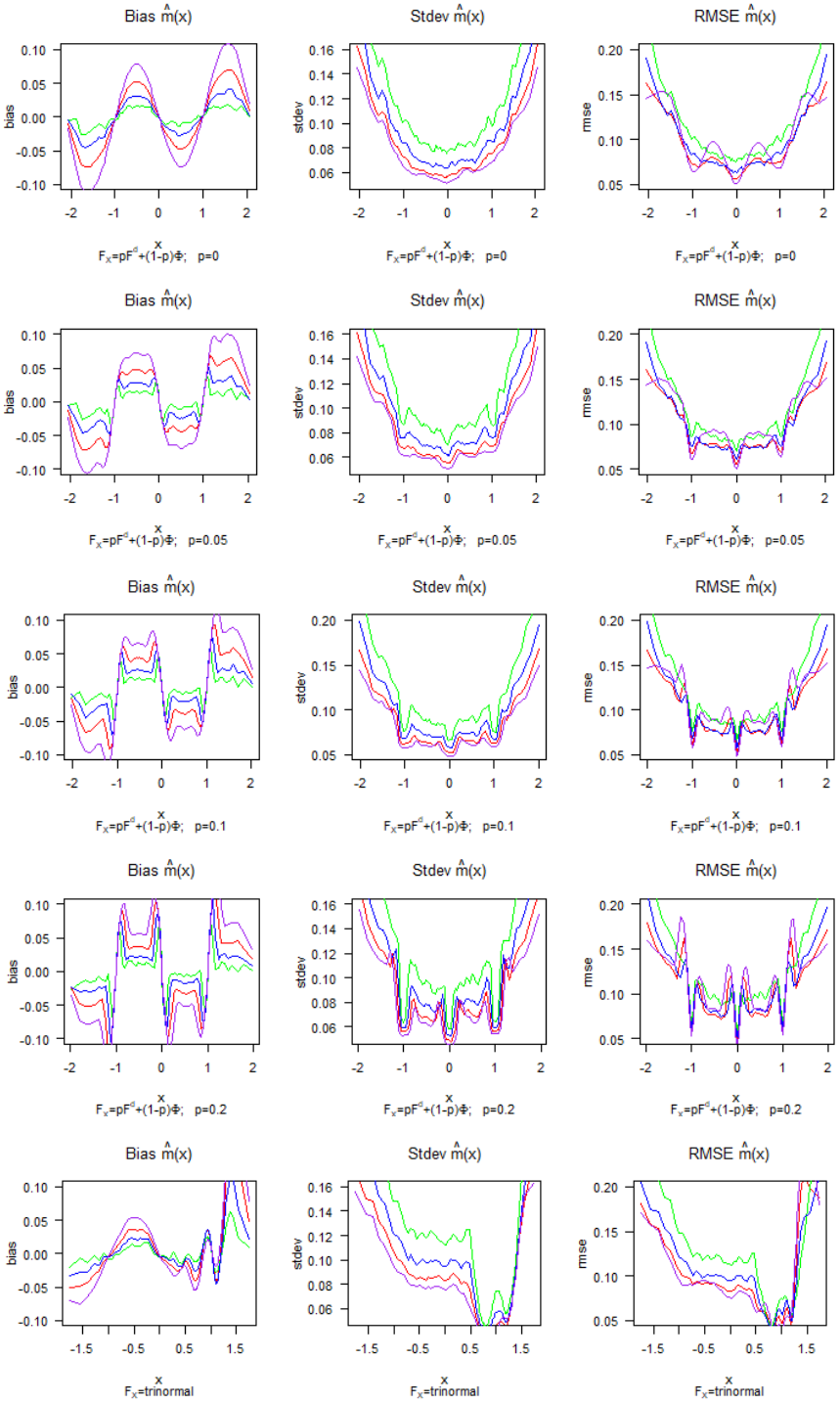
snr=1

	n=100		**n= 500**		**n=1,000**	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
$F_{mass}(p=0)$	0.591	0.739	0.572	0.717	0.569	0.713
$F_{mass}(p=0.01)$	0.584	0.731	0.565	0.708	0.562	0.704
$F_{mass}(p=0.1)$	0.577	0.722	0.558	0.700	0.555	0.695
$F_{mass}(p=0.2)$	0.562	0.703	0.543	0.681	0.540	0.677
$F_{mixtureN}$	0.541	0.677	0.525	0.658	0.521	0.653

snr=2

	n=100		**n= 500**		**n=1,000**	
$F_{mass}(p=0)$	0.423	0.530	0.406	0.509	0.403	0.506
$F_{mass}(p=0.01)$	0.418	0.524	0.401	0.503	0.398	0.499
$F_{mass}(p=0.1)$	0.414	0.518	0.396	0.497	0.393	0.493
$F_{mass}(p=0.2)$	0.404	0.506	0.386	0.484	0.383	0.481
$F_{mixtureN}$	0.389	0.487	0.373	0.467	0.369	0.463

Figure D.2: Bias, standard deviation and RMSE of NW kernel estimator by conditioning distribution and bandwidth (n=1,000)



Note: 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.

Table D.3: Empirical Rates.

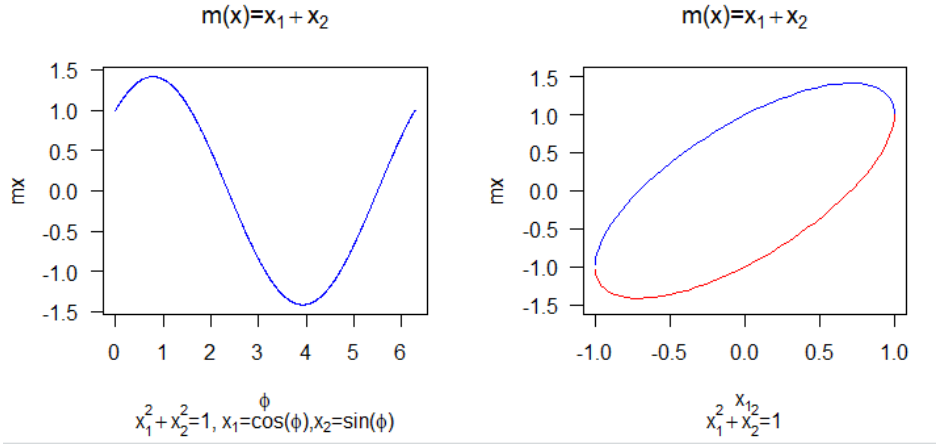
Point mass

		h=.75hcv		h=hcv		h=1.25hcv	
X	Oracle	NW	ratio	NW	ratio	NW	ratio
p=0.2							
0.0	-0.498	-0.479	0.961	-0.470	0.944	-0.462	0.928
0.1	-0.498	-0.252	0.507	-0.203	0.408	-0.194	0.390
0.2	-0.498	-0.435	0.874	-0.384	0.770	-0.272	0.545
0.4	-0.499	- 0.457	0.916	-0.475	0.953	-0.490	0.983
0.8	-0.499	-0.404	0.812	-0.367	0.735	-0.288	0.577
0.9	-0.499	-0.273	0.547	-0.232	0.466	-0.226	0.453
1.0	-0.499	-0.485	0.870	-0.479	0.961	-0.472	0.947
p=0.1							
0.0	-0.500	-0.461	0.921	-0.458	0.915	-0.458	0.914
0.1	-0.500	-0.239	0.477	-0.200	0.401	-0.190	0.381
0.2	-0.502	-0.449	0.894	-0.385	0.766	-0.267	0.532
0.4	-0.506	-0.445	0.879	-0.463	0.914	-0.479	0.946
0.8	-0.507	-0.402	0.794	-0.367	0.725	-0.293	0.577
0.9	-0.505	-0.270	0.532	-0.239	0.473	-0.237	0.469
1.0	-0.503	-0.476	0.945	-0.471	0.935	-0.465	0.924

Mixture of normal

		h=.75hcv		h=hcv		h=1.25hcv	
X	Oracle	NW	ratio	NW	ratio	NW	ratio
0.00	-0.511	-0.444	0.868	-0.449	0.878	-0.451	0.882
0.50	-0.492	-0.376	0.764	-0.381	0.775	-0.421	0.856
0.75	-0.492	-0.432	0.877	-0.416	0.845	-0.413	0.838
1.00	-0.497	-0.399	0.802	-0.413	0.832	-0.436	0.877
1.25	-0.511	-0.421	0.825	-0.397	0.778	-0.390	0.763
1.50	-0.509	-0.397	0.780	-0.356	0.699	-0.295	0.580

Figure D.3: Bivariate conditional mean with singularity.



(a) (b)

Notes: (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 D.4: Some details of cross validated bandwidth for NW/SI kernel estimator by sample size (n), and signal to noise ratio (snr)

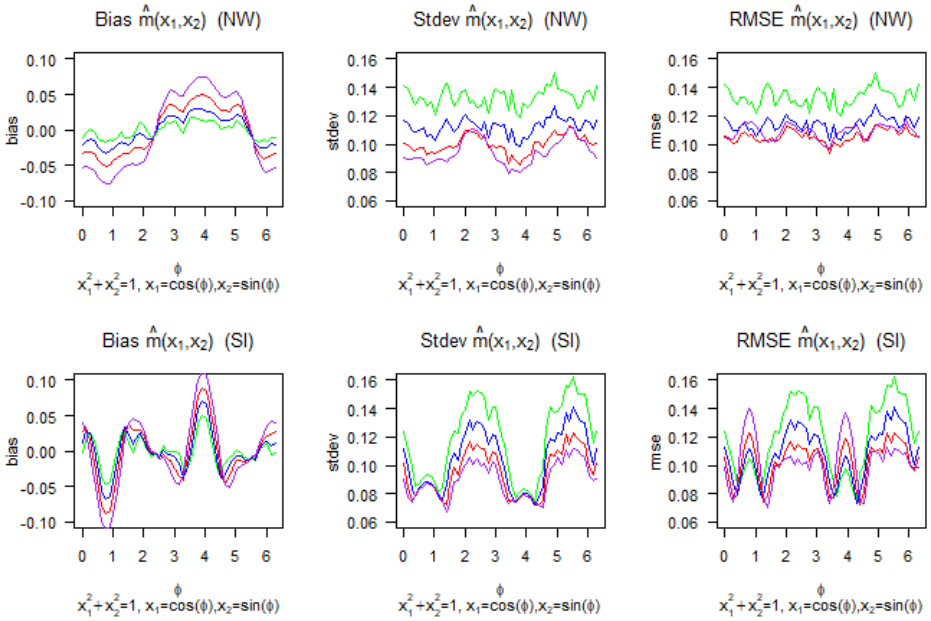
	snr=1			**snr=2**		
	n=100	n=500	n=1,000	n=100	n=500	n=1,000
NW Kernel estimator						
$h_{1,cv}$	0.861	0.604	0.523	0.735	0.529	0.456
	(.324)	(.182)	(.141)	(.267)	(.155)	(.112)
$h_{2,cv}$	0.879	0.611	0.528	0.748	0.530	0.459
	(.314)	(.180)	(.141)	(.261)	(.150)	(.114)
Single index						
b_2	1.050	1.006	1.008	1.028	1.003	1.005
	(.251)	(.108)	(.077)	(.175)	(.075)	(.053)
h_{cv}	0.795	0.478	0.386	0.653	0.388	0.311
	(.335)	(.187)	(.137)	(.277)	(.139)	(.099)

Table D.5: Some details of aggregate performance for NW/SI kernel estimator by sample size, and signal to noise ratio (snr).

		NW Kernel estimator			Single Index estimator		
		n=100	n=500	n=1,000	n=100	n=500	n=1,000
snr=1							
	MAE	0.810	0.804	0.801	0.810	0.801	0.800
	RMSE	1.014	1.007	1.003	1.003	1.004	1.002
snr=2							
	MAE	0.576	0.569	0.567	0.569	0.567	0.566
	RMSE	0.721	0.713	0.710	0.711	0.711	0.709

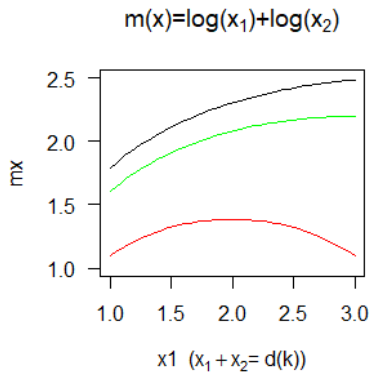
Note: The MAE is given by $\sum_{i=1}^{n}|y_i - \hat{m}(x_i)|\,/n$ *and the RMSE is given by* $\sqrt{\sum_{i=1}^{n}\left(y_i - \hat{m}(x_i)\right)^2\,/n}$ *where the leave-one-out estimator for m(x$_i$) is used.*

Figure D.4: Bias, standard deviation and RMSE of NW and SI kernel estimator by bandwidth (n=1,000).



Note: The top panel present the results for the Nadaraya-Watson (NW) local constant estimator and the bottom panels present the results for the Single Index (SI) estimator. The green and blue lines present results with 0.5hcv and 0.75hcv respectively, the purple line uses 1.25hcv, and the red line presents the cross-validated bandwidth.

Figure D.5: Bivariate conditional mean with reduced dimensionality



Note: The red line represents d(1)=3, the green line d(2)=6 and the black line d(3)=7.

Table D.6 Comparison of the rate of the NW and SI estimator to the "Oracle" (True model) estimator.

Non-singular (uniform [-1,1])

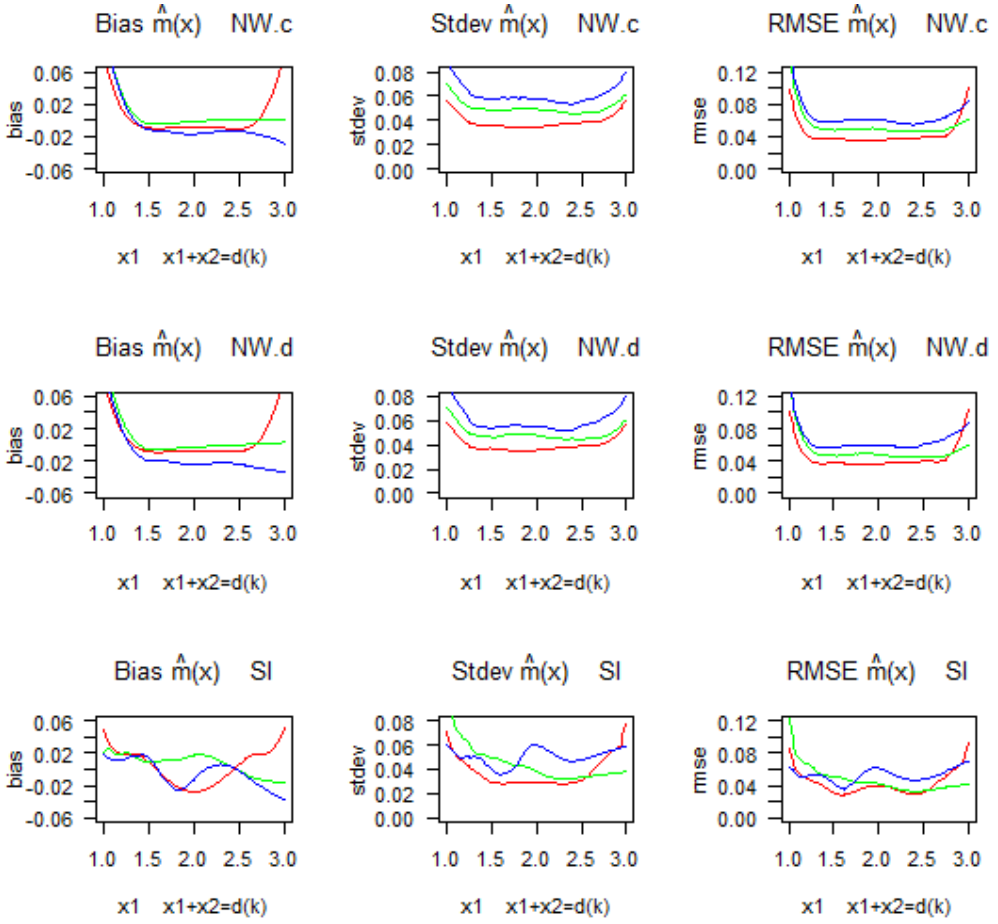
phi	Oracle	NW	Ratio	SI	Ratio
0.00	-0.511	-0.306	0.599	-0.441	0.864
0.25	-0.500	-0.493	0.993	-0.435	0.876
0.50	-0.502	-0.297	0.592	-0.447	0.891
1.00	-0.521	-0.291	0.559	-0.446	0.854
1.50	-0.503	-0.302	0.601	-0.444	0.882
1.75	-0.514	-0.458	0.891	-0.479	0.931

Table D.7 Aggregate performance of the NW and SI estimator with cross validated bandwidth by sample size and snr=1.

	NW estimator Based on X1 and X2		NW estimator Based on X1 and d(k)		SI estimator Based on logX1 and logX2	
N	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
100	0.360	0.450	0.350	0.438	0.354	0.442
500	0.354	0.443	0.351	0.440	0.352	0.441
1,000	0.352	0.442	0.351	0.440	0.351	0.440

Note: The MAE is given by $\sum_{i=1}^{n}|y_i - \widehat{m}(x_i)|$ *and the RMSE is given by* $\sqrt{\sum_{i=1}^{n}\left(y_i - \widehat{m}(x_i)\right)^2}/n$ *where the leave-one-out estimator for* $m(x_i)$ *is used.*

Figure D.6: Bias, Standard deviation and RMSE of NW and SI kernel estimator at cross-validated bandwidth, n=1,000 (separately d(k), k=1,2,3).



Note: Red lines red are for d(1)=4, blue lines for d(2)=6, and green lines for d(3)=7

Figure D.7 A sample of 20 random curves and associated function used for conditional mean $m(X)=m_1(X_1)+X_2$.

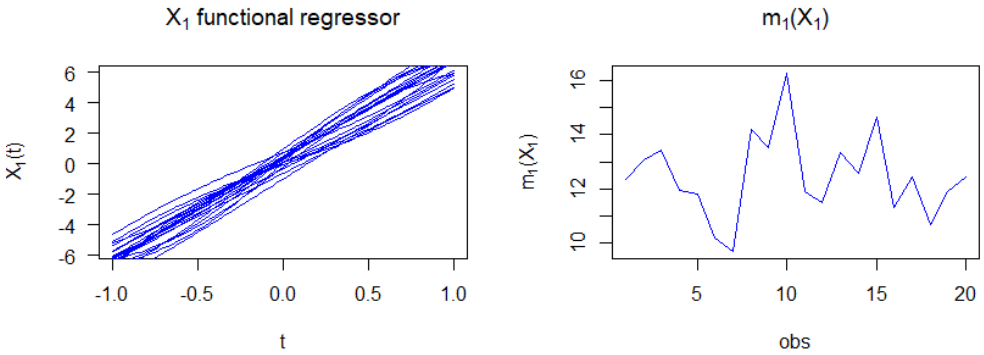


Table D.8 Some details of the cross validated bandwidths for NW in the presence of functional regressor X_1 and associated residual sum of squares.

	X_2 =N(0,1)	**$X_2=m_1(Z)$**	
		ρ =0.0	**ρ=0.8**
$h_{1,cv}$	1.369 (0.242)	1.583 (0.288)	1.746 (0.334)
$h_{2,cv}$	1.513 (0.310)	2.075 (0.414)	2.321 (0.546)
mean (CV)	4.670	7.189	12.020
Oracle:			
OLS/CV	0.890	0.891	0.916
Misspecification:			
CV_1/CV	1.170	1.409	1.280
CV_2/CV	1.616	1.377	1.178

Note: CV_1 stands for the mean (CV) achieved when ignoring the X_2 regressor and CV_2 stands for the mean (CV) achieved when ignoring the X_1 regressor

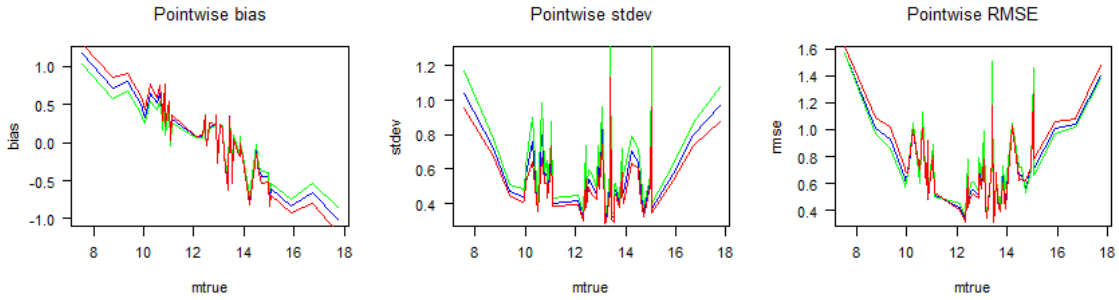
Table D.9 Aggregate performance (MAE and RMSE) and of the NW estimator in-sample and out-of-sample under the cross-validated bandwidth.

	X_2 =N(0,1)	$X_2=m_1(Z)$	
		ρ =0.0	ρ=0.8
In-sample			
MAE	0.570	0.697	0.804
RMSE	0.746	0.915	1.058
Out-of-sample			
MAE	0.577(4)	0.759(19)	0.920(18)
RMSE	0.758(4)	1.026(19)	1.210(18)

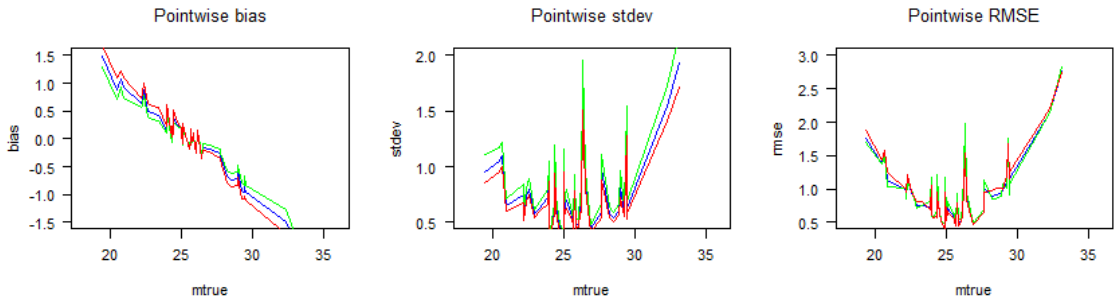
Note: The number in brackets indicate the number of simulations (out of 500) where at least one of the 50 out-of-sample observations did not have a neighbour with any observation in a given simulation using the cross validated bandwidth.

Figure D.8: Bias, Standard deviation and RMSE of NW by bandwidth, n=250 (out-of-sample).

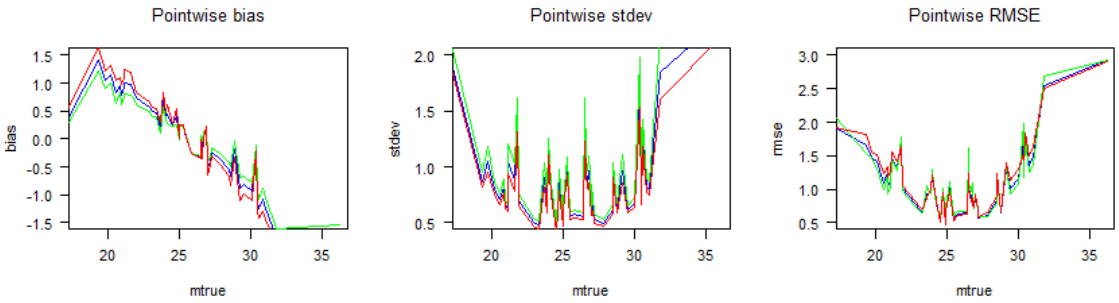
X_2 =N(0,1)



$X_2=m_1(Z)$, ρ =0.0



$X_2=m_1(Z)$, ρ =0.8



Note: The blue line presents the results using the cross-validated bandwidth, in green we consider 0.9hcv in red 1.1hcv.