Journal of Econometrics 253 (2026) 106168

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Multivariate kernel regression in vector and product metric spaces

Marcia Schafgans [a,*], Victoria Zinde-Walsh [b]

[a] Economics Department, London School of Economics, Houghton Street, London, WC2A 2AE, UK
[b] Economics Department, McGill University and CIREQ, 855 Sherbrooke St. W., Montreal, Quebec, H3A 2T7, Canada

ARTICLE INFO

Keywords:
Nadaraya–Watson estimator
Singular distribution
Multivariate functional regression
Small cube probability

ABSTRACT

This paper derives limit properties of nonparametric kernel regression estimators without requiring existence of density for regressors in $\mathbb{R}^q$. In functional regression limit properties are established for multivariate functional regression. The rate and asymptotic normality for the Nadaraya–Watson (NW) estimator is established for distributions of regressors in $\mathbb{R}^q$ that allow for mass points, factor structure, multicollinearity and nonlinear dependence, as well as fractal distribution; when bounded density exists we provide statistical guarantees for the standard rate and the asymptotic normality without requiring smoothness. We demonstrate faster convergence associated with dimension reducing types of singularity, such as a fractal distribution or a factor structure in the regressors. The paper extends asymptotic normality 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 $\mathbb{R}^q$. For functional regression the simulations underline the importance of accounting for multiple functional regressors. We demonstrate the applicability and advantages of the NW estimator in our empirical study, which reexamines the job training program evaluation based on the LaLonde data .

1. Introduction

This paper extends nonparametric kernel regression to more general regressor settings than those considered in the literature. The general regression model

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

is free from the difficulty of choosing a parametric specification. We focus on the Nadaraya–Watson (NW) estimator, introduced by Nadaraya (1965) and Watson (1964), which recognizes that a continuous regression function can be estimated pointwise by a weighted average that attaches higher weights to close-by observations.

Here we emphasize the fact that the regressor X can be a vector in $\mathbb{R}^q$, or alternatively X may belong to a function space, a more general metric space, or comprise of components from several such metric spaces. In fact, the components of X do not necessarily have to belong to spaces of vectors or functions but could be intervals, graphs, or networks, as long as a metric (or even a semi-metric) can be defined for each space. Y represents a scalar dependent variable, u denotes an unobserved error, and the conditional mean function m satisfies some smoothness assumptions.

The NW estimator has been used extensively with $X \in \mathbb{R}^q$ (see e.g. the textbook Li and Racine, 2007, for discussion and examples) and has recently been introduced to functional regression by Ferraty and Vieu (2004). Well known limit distributional results for

* Corresponding author.
E-mail addresses: m.schafgans@lse.ac.uk (M. Schafgans), victoria.zinde-walsh@mcgill.ca (V. Zinde-Walsh).

https://doi.org/10.1016/j.jeconom.2025.106168
Received 3 May 2024; Received in revised form 4 December 2025; Accepted 6 December 2025

the NW estimator were derived in $\mathbb{R}^q$ under restrictions requiring existence and smoothness of the density. For functional regression (where there is no density) the limit distributional results were derived in a univariate context only.

In this paper we establish asymptotic normality of the NW estimator for regression in the presence of a general regressor X that could have a multivariate singular distribution in $\mathbb{R}^q$ or is comprised of any number of functional and vector regressors. A singular distribution does not admit a density function that integrates to it.

In settings where data has both discrete and continuous components (Li and Racine, 2007) obtained asymptotic normality of the NW estimator without having to deal with the singularity by treating the discrete and (absolutely) continuous components separately. However, sometimes the distinction between discrete and continuous variables is not straightforward; continuous variables could be discretized with different levels of discretization. When data with both discrete and continuous components is viewed as a vector in a Euclidean space, $X \in \mathbb{R}^q$, the distribution of X is singular. In our simulations we demonstrate that there may be no gain from avoiding the singularity by considering the discrete regressors separately.

The presence of latent factors in the continuous regressors, common in macroeconomic and finance models (e.g., Bai et al., 2006, for portfolio, stock returns and macroeconomic data) could also imply a singular distribution. For example, if the regressor is a $q \times 1$ vector $X \sim N(0, \Sigma)$ with Σ a singular matrix of rank $r < q$, the distribution is singular. Similarly, non-linear common factors, such as in Hotelling (1929) spatial model of horizontal differentiation which assumes that each consumer has an 'ideal' variety identified by his location on the unit circle (see also Desmet and Parente, 2010, imply a smaller effective dimension for the regressor space, resulting in a singular distribution over $\mathbb{R}^q$. This also is true when there exists a functional relation between the regressors (e.g., with exact collinearity that can arise in production functions, Ackerberg et al., 2015).

Singularity also originates from a fractal structure in the data; examples in economics include the daily prices in the cotton market (Mandelbrot, 1997), financial markets, and networks (see Takayasu and Takaysu, 2009). Fractals are common to many geographic features, including coastlines, river networks and landforms, and have been used in urban growth studies (e.g., Shen, 2002) and spatial econometrics in general. Furthermore, singularities also result when continuously distributed variables exhibit mass points (e.g. Arulampalam et al., 2017, for neonatal mortality and Olson, 1998, for weekly hours worked).

We demonstrate the benefit of extending the NW estimator to regression with singular data by applying it to the data from a randomized experiment in a job training program evaluation study by Lalonde (1986). Following the work by Rosenbaum and Rubin (1983), Dehejia and Wahba (1999, 2002) applied propensity score methods to the LaLonde data for estimation of causal treatment effects in an attempt to generalize the experimental results to nonexperimental data. The propensity score matching was used instead of a multivariate nonparametric model with matching on individual characteristics which was deemed impractical because of the high dimensionality of the regressors. The benefits of kernel regression were analyzed in Heckman et al. (1997, 1998) (without allowing for singularity). The LaLonde data and methodologies were discussed by Angrist and Pischke (2009) and in a recent review by Imbens and Xu (2024). As shown here the discreteness of most of the regressors implies reduced dimension of the support of the joint distribution; for continuous variables existence of continuous density is not imposed and mass at zero in income is accounted for. Our asymptotic results provide the validity of the NW estimator for this singular distribution. The kernel estimators we employ give new insights into the heterogeneous effects of the program, based on a variety of individual characteristics and compare quite well with random forest estimates of the conditional average treatment effect on the treated (CATT) (Wager and Athey, 2018).

Our results also extend to functional regression (see, e.g. Ramsay and Silverman, 2005) where estimation and inference techniques have been developed by Ferraty and Vieu (2004) and pointwise asymptotic normality was established in regression for a Banach or metric space by Ferraty and Vieu (2006), Ferraty et al. (2007) and Geenens (2015) in the i.i.d. case. Masry (2005) derived the limit distribution for a strongly mixing process. Recently Kurisu et al. (2025) made a case for extending the univariate set-up of functional regression by considering jointly a random vector and a function to obtain an estimate for the propensity score used in evaluating the average treatment effect. We establish asymptotic normality in multivariate functional regression with regressors in any number of heterogeneous metric spaces. 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] Our simulations show that using multivariate rather than univariate functional regression can improve the fit of the kernel estimator.

We derive asymptotic normality results for a random regressor X supported on some domain in a vector space, $\mathbb{R}^q$, or metric, semi-metric space, $\Xi^{[1]}$, or a product of such spaces $\Xi^{[q]} \equiv \Xi_1^{[1]} \times \ldots \times \Xi_q^{[1]}$. The metrics on $\Xi_l^{[1]}$, $\|.\|_l$, may differ for each of the q components of function spaces, thus as in Kurisu et al. (2025) one may be the $\mathbb{R}^1$ space and the other one a function space. A key ingredient in our technical derivations is small cube probability, which characterizes local properties of X in the general multivariate case in place of the density. We introduce this concept here.

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 $B(x,h) = \{X : \|x - X\| \le h\}$ centered at x 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 $\mathbb{R}^q$ or, in general, in product metric spaces $\Xi^{[q]}$.

Kankanala and Zinde-Walsh (2024) introduced small cube probability for a cuboid. A cuboid $C(x,h)$ centered around $x = (x^1, \ldots, x^q) \in \mathbb{R}^q$ for a vector $h = (h^1, \ldots h^q)'$ with positive finite components is defined as the set $C(x,h) = \{X \in \mathbb{R}^q : |X^l - x^l| \le h^l, l = 1, \ldots, q\}$. With the distribution function of X given by F_X the corresponding probability measure is

[1] E.g., Caldeira et al. (2020) compares the model forecasting aggregate stock market excess return on a function representing the history of returns with regression models based on traditional predictors. Ferraty and Nagy (2022) compare the performance of separate models for predicting adult height with functional regressors (one being growth velocity profiles from ages 1–10 and the other for 5–8).

$P_X(C(x,h)) = \int_{C(x,h)} dF_X$. The small cube probability permits us to extend the regression on univariate metric spaces to $\Xi^{[q]}$ where the probability measure P_X is defined. For the cuboid

$$C(x,h) = \left\{ X : \left\| X^l - x^l \right\|_l \le h^l, l = 1, \dots, q \right\} = \left\{ X : X^l \in B^l\left(x^l, h^l\right), l = 1, \dots, q \right\}. \tag{2}$$

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

One of our contributions is the derivation of auxiliary technical results that express moments for the multivariate kernels and related functions in terms of the small cube probabilities without appealing to differentiability on which previous multivariate derivations relied. The moments and moment bounds are derived for general multivariate local functions under arbitrary distributions over $\mathbb{R}^q$ or probability measures over $\Xi^{[q]}$. Bounds on a moment functional expressed via power of small cube probability pinpoint the rate of growth of the functional. These results generalize the derivations for the univariate kernel used in functional regression to the multivariate setting. The full details of these auxiliary technical results are presented in the supplemental material (Appendix A). The moment expressions could find use in other contexts, for instance for local linear and local polynomial estimation in $\mathbb{R}^q$ or in products of suitable metric spaces, $\Xi^{[q]}$, kernel estimation of distribution functions and conditional distributions in $\mathbb{R}^q$ as well as to kernel regression of objects in metric spaces on objects in products of spaces.

Implementation of the NW estimator relies on a tuning bandwidth parameter. We show that in $\mathbb{R}^q$ a popular cross-validation method of choosing a bandwidth with properties that were worked out for the absolutely continuous (a.c.) case, has similar properties in some empirically relevant classes of singular distributions with dimension-reducing singularity. We also examine adaptive bandwidth selection for regressor distributions that are represented by a mixture of a continuous distribution with some mass points.

We provide simulation evidence on some important features of the behavior of the NW estimator under possible singularity of the distribution of regressors, F_X, in $\mathbb{R}^q$, in particular on the pointwise rate of convergence and specific impact of mass points. We examine the behavior of the NW estimator for models with dependence on both a functional object in $\Xi^{[1]}$ and a random variable.

The structure of the paper is as follows. Section 2 provides the set-up suitable for the multivariate vector and functional regression highlighting the probability measure for the regressor. Section 3 gives the asymptotic normality results under the most general distributional assumptions. Section 4 discusses implementation, in particular, bandwidth selection. Section 5 provides a sketch of the simulation results and Section 6 is devoted to the empirical study. The supplementary material collects various auxiliary results and the proofs as well as the details of the Monte Carlo simulations and the empirical study.

2. The set-up and assumptions

This section provides the formula for the Nadaraya–Watson (NW) kernel estimator over $\Xi^{[q]}$, introduces some useful notation and gives formal assumptions. The distributional assumptions are very general in that they do not restrict the distribution over $\mathbb{R}^q$ to have absolutely continuous components, and apply to the probability measure over the multivariate metric space $\Xi^{[q]}$ for an arbitrary q.

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

In the product space $\Xi^{[q]}$ we define a vector w as $\left(w^1, \dots, w^q\right)^T$ where each component is in the corresponding space, thus for $\Xi = \mathbb{R}^q$, w is a q -dimensional vector of reals, in $\Xi = \Xi^{[q]}$ each $w^l \in \Xi_l^{[1]}$, $l = 1, \dots, q$. The bandwidth vector is $h = \left(h^1, \dots, h^q\right) \in \mathbb{R}^q$ with $0 < \underline{h} = \min\left\{h^1, \dots, h^q\right\} > 0$ and $\bar{h} = \max\left\{h^1, \dots, h^q\right\}$. We use the same notation $\|\cdot\|$ for the absolute value of a scalar in $\mathbb{R}^1$, the Euclidean norm for a vector in $\mathbb{R}^q$ 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 Nadaraya–Watson (NW) estimator

The NW estimator, $\widehat{m}(x)$, for a sample $\left\{\left(Y_i, X_i\right)\right\}_{i=1}^n$ generated by (1) is defined below. Generically the argument of the kernel function is

$$W_X(x) = \begin{cases} h^{-1}(x - X) & = \left(\left(h^1\right)^{-1}(x^1 - X^1), \dots, (h^q)^{-1}(x^q - X^q)\right) \quad \text{on } \Xi = \mathbb{R}^q \\ h^{-1}\|x - X\| & = \left(\left(h^1\right)^{-1}\left\|x^1 - X^1\right\|_1, \dots, (h^q)^{-1}\|x^q - X^q\|_q\right) \quad \text{on } \Xi = \Xi^{[q]}. \end{cases} \tag{3}$$

The NW estimator is given by

$$\widehat{m}(x) = B_n^{-1}(x) A_n(x), \text{ with} \tag{4}$$

$$B_n(x) = \frac{1}{n}\sum_{i=1}^n K\left(W_i(x)\right); \; A_n(x) = \frac{1}{n}\sum_{i=1}^n K\left(W_i(x)\right) Y_i. \tag{5}$$

where $K(W_i(x)) = K(W_{X_i}(x))$ is a multivariate (non-negative) kernel function and h usually depends on n; x such that at least for some i we have that $K\left(W_i(x)\right) > 0$. The kernel function K and bandwidth vector h determine the properties for the NW estimator. In the metric space $\Xi = \Xi^{[1]}$ the kernel function K is defined for a univariate non-negative argument; in the case $\Xi = \Xi^{[q]}$ with $q > 1$ different bandwidths could appear for the different components $W_X^l(x), l = 1, \dots, q$. With a symmetric kernel on $\mathbb{R}^q$ we can just write $W_X(x) = h^{-1}\|x - X\|$ for any Ξ.

2.2. The kernel

We restrict the multivariate kernel functions on $\mathbb{R}^q$ to have bounded support and be suitably differentiable in the interior.

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

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

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 and is continuous at any point on the interior of its support.

The following assumption is made on the kernel function.

Assumption 2 (Kernel)**.**

(a) The kernel function $K(w) = K(w^1,\dots,w^q)$ is a sufficiently differentiable density function.
(b) $K(w)$ is non-negative; $K(w)$ is non-increasing for $w : w^j \geq 0,\ j = 1,\dots,q$.
(c) $K(w)$ is either symmetric (with respect to zero) with support on $[-1,1]^q$ or $K(w)$ is supported on $[0,1]^q$.
(d) $K(w)$ satisfies $K(\iota) > 0$ where $\iota = (1,\dots,1)'$.

Assumption 2(a–c) are satisfied by the commonly employed product kernels of Epanechnikov or quartic kernels. Assumption 2(d) is not usual for kernel regression on $\mathbb{R}^q$; in the context of univariate functional regression it is satisfied by a Type I kernel defined in Ferraty and Vieu (2006) as $K : C_1 I_{[0,1]} \leq K \leq C_2 I_{[0,1]}$ with some $0 < C_1 \leq C_2 < \infty$. Condition (d) in conjunction with (a–c) provides the same type of univariate kernel. Extended to a multivariate setting it can be said that a kernel that satisfies Assumption 2(a–d) is a type I kernel. The uniform kernel is an example. The functional regression literature demonstrates that with kernels of type I asymptotic normality can be established in more general univariate settings. As commonly used in $\mathbb{R}^q$ kernels are not of type I, the asymptotic normality results are given separately to apply under Assumption 2(a–c) and under the full Assumption 2.

2.3. Additional assumptions

Consider the process $\{(X_i,Y_i)\}_{i\in\mathbb{N}}$. An i.i.d sequence would provide the simplest characterization, but strong mixing 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 (Data Generating Process and Moments)**.**

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

$$\alpha(l) < Cl^{-\kappa};\ \ \kappa > \frac{2(2+\zeta)}{\zeta}.$$

(b) $E(u|X=x) = 0$; $\mu_2(x) = E(u^2|X=x)$ satisfies $0 < L_{\mu_2} < \mu_2(x) < M_{\mu_2} < \infty$, $\mu_2(x)$ is continuous in the neighborhood of x.
(c) $E|Y_i|^{2+\zeta} < \infty$ and $E(|u|^{2+\zeta}|X=x) < \infty$.
(d) For $x\in\Xi^{[q]}$ and $i\neq j$ the bivariate function

$$\mu(x_1,x_2) = E\left(|u_iu_j||X_i = x_1, X_j = x_2\right)$$

is continuous in a neighborhood of the point $(x,x)\in\Xi^{[q]}\times\Xi^{[q]}$.
(e) The conditional expectation $E\left(|Y_iY_j||X_i,X_j\right) \leq C < \infty$ for all i,j.

The assumption requires a polynomial bound on the rate of decline of the mixing coefficient with a link to the moment of Y; it is similar to those in Masry (2005) and Hong and Linton (2020).

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

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

Assumption 4 requires Holder continuity of $m(x)$; it would follow from differentiability or Lipschitz continuity in $\mathbb{R}^q$ with $\delta = 1$. In the above assumptions, and below, L and 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.

2.4. The probability measures

For the probability measure P_X on a generic space Ξ, that could coincide with $\mathbb{R}^q$, $\Xi^{[1]}$, or $\Xi^{[q]}$, any point $x \in \Xi$ is a point of support if for $\underline{h} > 0$ the measure $P_X\left(C\left(x, \underline{h}\right)\right) > 0$.

2.4.1. Measures on $\mathbb{R}^q$

By the Lebesgue decomposition, the distribution F_X on $\mathbb{R}^q$ 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 :

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

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, but with some discrete components or mass points in some of the continuous components the distribution can no longer be absolutely continuous and is singular. In many applications at least one of the variables is assumed continuous and in a semiparametric regression often an index model is assumed (single index in Ichimura, 1993; multiple index in Donkers and Schafgans, 2008) to avoid singularity as well as to 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, 2024) that will have a similar beneficial effect on the convergence of the kernel estimator.

2.4.2. Measures on metric spaces and products

The discussion in this section applies to the space $\mathbb{R}^q$ as a special case. Particular classes of probability measures considered in univariate functional regression (e.g. Ferraty and Vieu, 2006; Ferraty et al., 2007) have a small ball probability centered at a point x of support either with a polynomial (fractal) rate of decline $P_X(B(x,h)) \sim C(x)h^{\tau} > 0$ ($\tau > 0$), or with an exponential type rate of decline $P_X(B(x,h)) \sim C(x)\exp(-h^{-\tau_1}\log h^{-\tau_2})$ ($\tau_1 > 0, \tau_2 > 0$) as $h \to 0$. This characterization can be applied to the multivariate setting by replacing $B(x,h)$ with the cuboid and the univariate bandwidth in the rate with $\bar{h}$. The exponential rate of decay of the small ball probability requires a type I kernel and leads 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 (corresponds to observation of smooth curves with common shape) as noted by Ferraty and Nagy (2022).

Kernels of type I play an important role in establishing pointwise asymptotic normality in the absence of any restrictions on the decline of the small cube measure. For kernels that may not be of type I sufficient conditions on the shrinkage of the probability measure as $h \to 0$ were proposed in Assumption H_3 in Ferraty et al. (2007), Ferraty and Vieu (2006) and were referred to in various subsequent papers on functional regression, e.g. Hong and Linton (2020). The assumption below generalizes these conditions to apply to $C(x,h)$ on $\Xi^{[q]}$; the assumption is both necessary for the conditions to hold (see the supplementary material, Appendix B) and at the same time sufficient for the convergence results.

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

$$\frac{P_X(C(x,h))}{P_X(C(x,\varepsilon h))} < C_\varepsilon < \infty. \tag{6}$$

Definition 1. $\mathcal{D}$ is the class of probability measures that satisfies (6).[2]

A polynomial decay condition places a measure into class $\mathcal{D}$. Indeed if the small cube probability satisfies

$$0 < L_P(x)(2\underline{h})^{s(x)q} \leq P_X(C(x,h)) \leq M_P(x)\left(2\bar{h}\right)^{s(x)q} < \infty \tag{7}$$

where for some c, $H(x)$, $1 \leq c < \infty$, $0 < H(x) < \infty$ and $\bar{h} = c\underline{h} < H(x)$, $0 \leq s(x) \leq 1$ and $M_F(x)/L_F(x) < B < \infty$ at all points of support x, (6) holds with $C_\varepsilon = B(c/\varepsilon)^q$.

Condition (7) applies quite widely and holds for many distributions of regressors used in econometric models. In $\mathbb{R}^q$ it is satisfied by any absolutely continuous distribution with a positive bounded density function $f_X(x)$ where $M_P(x) \geq \sup_{\tilde{x} \in C(x,H)} f_X(\tilde{x})$; $L_F(x) = \inf_{\tilde{x} \in C(x,H/c)} f_X(\tilde{x})$ and $s(x) = 1$. If x is an isolated mass point then (7) applies with $s = 0$. If X has a linear structure with r common factors, the probability measure is singular and satisfies (7) with $s(x) = s = \frac{q}{r}$. For a fractal distribution that is singular with constant s, $0 < s < 1$, the bounds also apply.

Condition (7) is satisfied by the general class of Ahlfors (1966) regular (A-r) distributions common in statistics, where for this class $s(x) = s$, and $L_P(x) = L$ and $M_P(x) = M$ are constants, as well as by a finite mixture of such distributions (as proved in the

[2] Condition (6) is equivalent to the doubling property (e.g. Vol'berg and Konyagin, 1988) that states that (6) 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 (6) holds for $C_\varepsilon = C_{1/2}^{\kappa_1}$; if (6) holds, then the constant for doubling is $C_{1/2} = C_\varepsilon^{\kappa_2}$. We introduce the form (6) in case there is a preference for some ε.

supplementary material, Appendix B). Thus an absolutely continuous distribution ($s = 1$) or, more generally, a measure given by a continuous possibly singular distribution function that satisfies (7) contaminated with some mass points ($s = 0$) is in D ; this applies to the empirical example examined here, ensuring the pointwise asymptotic normality of the NW estimator with standard kernels.

2.4.3. Joint measure

Consider 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). The joint measure $P_{s,t}(C(x,h) \times C(x,h))$, defined as $\Pr\left(X_t \in C(x,h), X_s \in C(x,h)\right)$, is a product of the measures of the cuboid in the case of independency. With dependence an additional assumption is made on how the joint measure relates to the small cuboid measure. We provide the same assumption as in e.g. Masry (2005) and Hong and Linton (2020) for the small cube probability.

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

$$\sup_{t \neq s} P_{s,t}(C(x,h) \times C(x,h)) \leq M_{FF}\left(P_X(C(x,h))\right)^2. \tag{8}$$

3. Asymptotic normality of the NW estimator

Consider the NW estimator as given by (4), (5). As the sample size increases the bandwidths are assumed go to zero. For $\Xi = \mathbb{R}^q$ 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 univariate metric space was established in Györfi et al. (2002), the limit distribution in Masry (2005), Ferraty et al. (2007) and Geenens (2015).

The key to the asymptotic normality result is the derivation of the moments for multivariate functions of the form $g(X)K^m\left(h^{-1}\|x - X\|\right)$ for general probability measures and establishing lower and upper bounds (derivations in the supplementary material, Appendix B). The bounds provide expressions in terms of the small cube probability:

$$L_{EgK^m}(x)P_X(C(x,h)) \leq \left|E\left[g(X)K^m\left(h^{-1}\|x - X\|\right)\right]\right| \leq M_{EgK^m}(x)P_X(C(x,h)) \tag{9}$$

with constants $L_{EgK^m}(x)$ and $M_{EgK^m}(x)$ at x. Most important, (9) provides a lower bound on $EB_n(x) = EK\left(K^m\left(h^{-1}\|x - X\|\right)\right)$, given by $L_{EK}P_X(C(x,h))$, with appropriate conditions for L_{EK} to be strictly positive to ensure that the denominator of the NW estimator is such that it exists and the limit does not blow up. Type I kernel automatically entails that $L_{EK} > 0$, but for kernels such as Epanechnikov the bound requires Assumption 5. With $g(X)$ that is continuous at x

$$E\left[g(X)K^m\left(h^{-1}\|x - X\|\right)\right] = g(x)E\left[K^m\left(h^{-1}\|x - X\|\right)\right](1 + o(1)).$$

These moment expressions for distributions over $\mathbb{R}^q$ hold under the standard assumptions of existence and continuity of (bounded) density f_X, and the function g, where

$$E\left[g(X)K^m\left(h^{-1}\|x - X\|\right)\right] = \prod_{i=1}^{q}\left(-h^i\right)g(x)f_X(x)\int K^m(v)dv(1 + o(1)), \tag{10}$$

with more details about the $o(1)$ term under smoothness of f_X (see, e.g. derivations in Li and Racine, 2007). Once the moments and the bounds are derived, the proofs of asymptotic normality proceed along similar lines to those in Masry (2005).

The point-wise limit normality is provided in the next theorem under two alternative types of conditions: (i) with type I kernel without imposing further constraints on F_X, and (ii) not imposing the type I kernel but with the distributional Assumption 5. Denote the bias of the estimator given x, $E\left(\widehat{m}(x)\right) - m(x)$, by $bias\left(\widehat{m}(x)\right)$. The difference $\widehat{m}(x) - m(x)$ is delivered by $\frac{A_n^c(x)}{B_n(x)}$ with the "centered" $A_n^c(x) = A_n(x) - m(x)B_n(x)$.

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 $nP_X(C(x,h)) \to \infty$*

(a)

$$\frac{\sqrt{n}E\left[K\left(h^{-1}\|x - X\|\right)\right]}{\sqrt{\mu_2(x)E\left[K^2\left(h^{-1}\|x - X\|\right)\right]}}\left(\widehat{m}(x) - m(x) - bias(\widehat{m}(x)\right) \to_d Z \sim N(0,1);$$

(b) the rates are

$$bias(\widehat{m}(x) = O(\bar{h}^{\delta}) + O\left(nP_X(C(x,h))\right)^{-1};$$
$$\frac{\sqrt{n}E\left[K\left(h^{-1}\|x - X\|\right)\right]}{\sqrt{\mu_2(x)E\left[K^2\left(h^{-1}\|x - X\|\right)\right]}} \simeq O\left((nP_X(C(x,h))^{1/2}\right).$$

(c) for h such that $\bar{h}^{2\delta}\left(nP_X(C(x,h))\right) \to 0$

$$\frac{\sqrt{n}E\left[K\left(h^{-1}\|x - X\|\right)\right]}{\sqrt{\mu_2(x)E\left[K^2\left(h^{-1}\|x - X\|\right)\right]}}\left(\widehat{m}(x) - m(x)\right) \to_d Z \sim N(0,1).$$

Remarks.

1. A sequence of bandwidths at x that satisfy the conditions of the theorem always exists. Indeed, whatever the rate of monotonic decline in $P_X(C(x,h))$ as $h \to 0$ for $n \to \infty$ a sequence of h that depends on n such that $nP_X(C(x,h)) \to \infty$ always exists. The rate for the bias of $\widehat{m}(x)$ in $\Xi^{[q]}$ is established in the theorem as $O(\bar{h}^{\delta}) + O\left(\left(nP_X(C(x,h))\right)^{-1}\right)$. For the bias (squared) to disappear in the limit $\bar{h}^{2\delta}P_X(C(x,h))n$ needs to go to zero. If $P_X(C(x,h)) \to 0$ a bandwidth sequence that simultaneously satisfies $nP_X(C(x,h)) \to \infty$ and $\bar{h}^{2\delta}P_X(C(x,h))n \to 0$ can always be found; when x is a mass point $P_X(C(x,h))$ will be bounded from below, but selecting $h = o\left(n^{-1/2\delta}\right)$ for such a point makes the bias term go to zero.
2. The assumptions of Theorem 1 and the moment computations in the supplementary material (Appendix B) imply that $E\left[K\left(h^{-1}\|x - X\|\right)\right]$ has the same rate as $P_X(C(x,h))$ while $varA_n^c(x)$ declines at the rate $P_X(C(x,h))/n$. The rate for the asymptotic variance for $\widehat{m}(x)$ equals $\left(nP_X(C(x,h))\right)^{-1}$ (this goes to zero).
3. The limit result shows that when density exists for a distribution on $\mathbb{R}^q$, the standard convergence rate $n^{1/2}h^{q/2}$ applies since then $P_X(C(x,h)) = O(h^q)$. This rate holds even when the density is discontinuous. Without the usual smoothness assumptions made in the literature, statistical guarantees for the rate and for asymptotic normality are thus shown to hold.
4. If there is singularity at the point x that satisfies (7) with $s < 1$, then the rate 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 at that point the parametric rate $n^{1/2}$ holds.
5. Under continuous differentiability the rate of the bias can be reduced by employing a local linear estimator (see, e.g. the standard derivations in Li and Racine, 2007, and for univariate functional regression in Ferraty and Nagy, 2022). Establishing the distributional properties of the local linear estimator with arbitrary probability distributions in $\mathbb{R}^q$ and multivariate probability measures in a metric space can proceed similarly, but requires stronger assumptions.

The convergence rate in (c) of Theorem 1 is $O\left((nP_X(C(x,h))^{-1/2}\right)$.[3] Existence of a limit variance $\sigma^2_{\widehat{m}(x)}$ requires that $\left(nP_X(C(x,h))\right)\frac{\left[EK\left(h^{-1}\|x-X\|\right)\right]^2}{\mu_2(x)E\left[K^2\left(h^{-1}\|x-X\|\right)\right]}$ converges. Without additional assumptions it is possible that the ratio does not converge; see example in the supplementary material (Appendix B) that provides a case when convergence does not hold; this happens when the small cube probability declines very rapidly and the kernel is not uniform. Suitable additional 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 restrictions on the probability measure on $\Xi^{[1]}$ that are sufficient for the convergence. Generally, one needs to ensure that the limits given below on the expectation of the kernel function and its square hold.[4]

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

$$\left(P_X(C(x,h))\right)^{-1}E\left[K^s\left(h^{-1}\|x - X\|\right)\right] \to \bar{B}_s(x); s = 1, 2.$$

This assumption holds quite widely. From the moment expressions it can easily be shown that it holds for the uniform kernel without any additional distributional assumptions. In the case of continuous density it holds by virtue of (10) with

$$\bar{B}_1(x) = f_X(x)\int K(v)dv, \quad \bar{B}_2(x) = f_X(x)\int K^2(v)dv. \tag{11}$$

Suppose that singularity arises, because of combining discrete and continuous variables in $\mathbb{R}^q$ or functional dependence between the regressors, that restrict the support of the distribution to be in some subspace of dimension $r < q$, $V(r) \subset \mathbb{R}^q$. If the distribution on $V(r)$ is absolutely continuous with a continuous density, then derivations provide similar limits to (11) with integration over $V(r)$ and density restricted to $V(r)$.

Define now

$$\alpha(n,h) = nP_X(C(x,h)); \quad \sigma^2_{\widehat{m}(x)} = \mu_2(x)\bar{B}_2(x)/(\bar{B}_1(x))^2.$$

Theorem 2. *Under the conditions of Theorem 1 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 $\mathbb{R}^q$ under smoothness assumptions on the distribution F_X. For functional regression our assumptions are comparable to those of Ferraty et al. (2007), Masry (2005), and subsequent papers while they make the extension to multivariate functional regression possible.

[3] This convergence rate obtains under $h \to 0$. In the presence of an irrelevant regressor, say $x^{(2)}$, such that $m(x) = m\left(x^{(1)}\right)$ for all $x = \left(x^{(1)}, x^{(2)}\right)$, this requirement can be restricted to the function $m\left(x^{(1)}\right)$ with the irrelevant $x^{(2)}$ eliminated. For the estimator this elimination can be achieved by setting the bandwidth on components of $x^{(2)}$ to be larger than the range of those variables, possibly infinite.

[4] This implies that the extra condition is also required for the Corollary 1 of Hong and Linton (2020).

4. Implementation and bandwidth selection

Estimation of $m(x)$ requires a selection of the kernel, K, and bandwidth, h. As may be 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. Aside from the estimator of the 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 $\mathbb{R}^q$, the leading term of the limit variance is expressed via the density function, often in the actual implementation the corresponding estimators do not make use of plug-in expressions, instead estimating the variance directly from the data and possibly with bootstrap (see Hall and Horowitz, 2013).

Cross-validation procedures in popular statistical packages (such as R) provide a single bandwidth (vector) that was shown to be consistent for the "optimal" bandwidth: minimizer of weighted integrated mean squared error, WIMSE, (e.g. Li and Racine, 2007). The proofs of consistency relied on absolute continuity of the regressors. The consistency results extend to some classes of singular distributions.

WIMSE is defined for an absolutely continuous distribution with density function $f_X(x)$ as

$$\int E\left(\hat{m}(x)-m(x)\right)^2 M(x) f_X(x) dx$$

with some weighting function $M(x)$ chosen to mitigate boundary effects. The expression can be written with dF_X replacing $f_X(x)dx$ (valid in the case of singularity):

$$\int E\left(\hat{m}(x)-m(x)\right)^2 M(x) dF_X = \int \left[var\left(\hat{m}(x)\right) + bias^2\left(\hat{m}(x)\right)\right] M(x) dF_X. \tag{12}$$

This function depends on the bandwidth vector h used in the estimator (see the review of bandwidth selection methods, including cross-validation and plug-in in Köhler et al., 2014). The "optimal" bandwidth vector h^0 is a minimizer of the WIMSE criterion function based on a trade-off between the variance and bias of the NW estimator.

In the cross-validation procedure the finite sample analogue of WIMSE replaces the expectation by

$$CV = n^{-1} \sum_{i=1}^{n} \left(Y_i - \hat{m}_{-i}\left(X_i\right)\right)^2 M\left(X_i\right)$$

employing the leave-one-out kernel estimator, $\hat{m}_{-i}$, and provides the bandwidth vector h_{cv} by minimizing the CV criterion.

Hall et al. (2007) gave a general result about consistency of the cross-validated bandwidth for regression over $\mathbb{R}^q$ with discrete and continuous regressors, with some of the regressors possibly being irrelevant. Their general result in Theorem 2.1 was obtained under a set of assumptions that required independent identically distributed observations, restrictions on the support of the probability measure, two continuous derivatives for density, the regression function, and the conditional variance of the error; in addition, for the d continuous relevant regressors $h^o = n^{-\frac{1}{4+rd}} a^o$ holds with the vector a^o having unique, positive and finite components. This result was extended to weakly dependent data by Li et al. (2009) under assumptions that replaced the i.i.d. assumption by requiring strict stationarity and β-mixing in the process for $\{x, y\}$ and martingale difference error, with suitable restrictions on the mixing parameters.

The result on the cross-validated bandwidth applies more widely. For instance, consider a singular distribution of $X \in \mathbb{R}^q$, where there is a functional dependence among the continuous variables in the presence of possibly some discrete covariates such that the support of the distribution is restricted to a subspace $V(r) \subset \mathbb{R}^q$ of dimension $r < q$ represented by a union of affine subspaces. If, restricted to $V(r)$, the distribution function is such that the conditions of Theorem 2.1 of Hall et al. (2007) or Theorem 1 of Li et al. (2009) are satisfied (Assumption CV) then the conclusions of those theorems are valid and the consistency of the bandwidth and automatic dimension reduction by smoothing out irrelevant regressors hold for this singular distribution. More details are provided in the supplementary material (Appendix B).

Importantly, no knowledge of $V(r)$ or r is required. This implies that for functionally dependent continuous regressors the knowledge of the number of factors is not required for the consistency of the cross-validated bandwidth or the automatic dimension reduction. We conjecture that in many other cases with possible singularity the cross-validation procedure will facilitate dimension reduction by smoothing out irrelevant variables.

Bandwidth selection could benefit from adaptation to different types of singularity. The treatment of adaptive bandwidth selection in the literature (Fan and Gijbels, 1996; Sain, 1994; Demir and Toktamis, 2010) typically focuses on adjusting the smoothing parameter to accommodate the varying data density, but not dealing with singularity or mass points. Adaptive bandwidths can provide a better fit of the criterion function by increasing the number of observations used to estimate the function at a point of sparsity.[5] Such bandwidths can similarly be constructed for cases of singular distributions. But these adaptation procedures still need to be investigated in the case of general mixtures of singular distributions. However, singularity adaptation simplifies considerably for the empirically important case of a mixture of an absolutely continuous distribution with mass points, where the two levels of singularity can be separated. The approach is detailed in the supplementary material (Appendix B).

[5] Given some initial bandwidth $\tilde{h}$ and density estimate at this bandwidth, $\hat{f}_X$, an adaptive bandwidth is defined for each point as $h\left(X_i\right) = \tilde{h}\left(\frac{\hat{f}_X(X_i)}{G}\right)^{-\alpha}$ where $G = \left(\prod \hat{f}_X\left(X_j\right)\right)^{1/n}$ is the geometric mean of the densities and α is typically selected to be $1/2$. One could construct $\hat{f}_X(x)$ with a uniform kernel in which case it is identical to an estimate of $P(C(x,\tilde{h}))$ by the proportion of observations in the $\tilde{h}$ cuboid around x.

Table 1
Empirical rate of convergence (i.e., $-\alpha_1$ for $O(n^{-\alpha_1})$) in the mass point and high derivative setting.

$F_X(x) = 0.2F^d(x) + 0.8\Phi(x)$			$F_X(x) =$ trinormal (x)	
X	NW	NW_a	X	NW
0.00	−0.455	−0.515	0.00	−0.449
0.10	−0.186	−0.443	0.50	−0.381
0.20	−0.411	−0.438	0.75	−0.416
0.30	−0.465	−0.431	1.00	−0.413
0.40	−0.461	−0.425		

Note: The column labeled NW_a contains the results implementing the adaptive bandwidth selection procedure in the presence of masspoints.

5. Simulations

This section provides the highlights of various simulations that show features of the finite sample performance of the NW estimator under singularity. Additional details and features are in the supplementary material (Appendix C).

5.1. Univariate (point mass example)

In this example we consider the regression distribution with mass points. Alongside we examine the trinormal mixture considered in Kotlyarova et al. (2016), an a.c. distribution which represents features (high density derivatives) that makes it comparable to a singular distribution.

The distribution with mass points, following Jun and Song (2019), is given by

$$F_X(x) = pF^d(x) + (1-p)\Phi(x) \quad \text{with } p = .2,$$

where F^d is the discrete uniform distribution function with $D = \{-1, 0, 1\}$ the set of mass points; Φ is the standard Gaussian distribution function.

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

$$Y_i = \sin(2.5X_i) + \sigma\varepsilon_i,$$

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 $K(u) = \frac{3}{4}(1-u^2)1(u^2 \le 1)$ and obtain the leave-one-out cross-validated bandwidth.

We analyze the pointwise RMSE 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 regress $\log(RMSE)$ 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 should be close to -0.4.[6]

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

For the 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). The empirical rate at mass points is close to -0.5 when the bandwidth is set equal to zero. The empirical convergence rate is slow for points close to the mass points (within the small ball probability measure under cross validated bandwidth) due to the boundary weight associated with mass in the neighborhood. Bandwidth adaptive to masspoints improves the rate. The convergence rates for the trinormal distribution, are reflective of usual smooth nonparametric regression although are somewhat faster at points with high derivatives.

5.2. Bivariate (with effective dimension 1)

We consider a model where $m(X) = \log(X_1) + \log(X_2)$ with regressors X_1 and X_2 satisfying $X_1 + X_2 = d(k)$, with fixed $d(k)$ corresponding to $k = 1, 2, 3$.[7] This is equivalent to a model with one continuous and one discrete regressor $m(X) = \log(X_1) + \log(D - X_1)$, with $D = d(k)$.

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

$$Y_i = \log(X_{1i}) + \log(X_{2i}) + \sigma\varepsilon_i$$

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

[7] An example could be where X_1 and X_2 represent earnings of the husband and wife and, for tax purposes, their combined income is set at some $d(k)$.

Table 2
Empirical Rates of the NW.c and NW.d estimators.

X_1	X_2	$d(k)$	NW.c (X_1, X_2)	NW.d $(X_1, d(k))$	
				ordered	unordered
1.5	2.5	4	−0.451	−0.422	−0.419
2.0	2.0	4	−0.436	−0.429	−0.426
2.5	1.5	4	−0.429	−0.406	−0.404
1.5	4.5	6	−0.457	−0.410	−0.422
1.5	5.5	7	−0.445	−0.392	−0.410

Note: The column labeled "ordered" contains the NW.d estimator where the discrete kernel is used for the discrete regressor; the column labeled "unordered" uses the Epanechnikov kernel.

Table 3
RMSE of the NW estimator in the presence of functional regressor X_1 at cross validated bandwidth, $n = 250$.

	$X_2 = N(0,1)$	$X_2 = m_1(Z)$	
		$\rho = 0.0$	$\rho = 0.8$
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 where the X_2 regressor is excluded and $RMSE_2$ stands for the RMSE when ignoring the functional regressor. The number in brackets indicates the number of simulations (out of 500) where at the cross-validation bandwidth no neighbor to the out-of-sample observation exists.

for different sample sizes with the additive error chosen as in the previous simulation. 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 first using X_1 and X_2 as regressors (NW.c) and second using X_1 and D as regressors (NW.d) and obtain the leave-one-out cross-validated bandwidths. For the discrete regressor D we use special discrete kernel weights proposed by Wang and Ryzin (1981) in accordance with Racine and Li (2004).

In Table 2 we provide illustrative results comparing the empirical rate of convergence of the NW.c and NW.d at a grid of points.

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 ($q = 2$). The estimate of the empirical rate for NW.c is slightly faster than NW.d, moreover, indicating that there is no gain from separate treatment of discrete regressors. With the reduced dimension structure here therefore one gets the rate corresponding to the Hausdorf dimension of the regressor space automatically without the need to recognize that it is possible to transform the regressors to one discrete, and one continuous variable.

5.3. *Bivariate (in the presence of a functional regressor)*

Here we examine a functional regressor in a multivariate setting. Consider a bivariate conditional mean function $m(X) = m(X_1, X_2)$, where X_1 is a functional regressor and $X_2 \in \mathbb{R}$ may be correlated with some $m_1(X_1)$. Let

$$Y_i = m_1(X_{1i}) + X_{2i} + \sigma\varepsilon_i.$$

Following (Ferraty et al., 2007), the functional regressor is defined as

$$X_{1i}(t) = \sin(w_i t) + (a_i + 2\pi)t + b_i, \quad t \in (-1, 1)$$

with a_i and b_i drawn from $U(-1, 1)$, w_i drawn from $U(-\pi, \pi)$ and

$$m_1(X_{1i}) = \int_{-1}^{1} |X'_{1i}(t)|(1 - \cos(\pi t))dt.$$

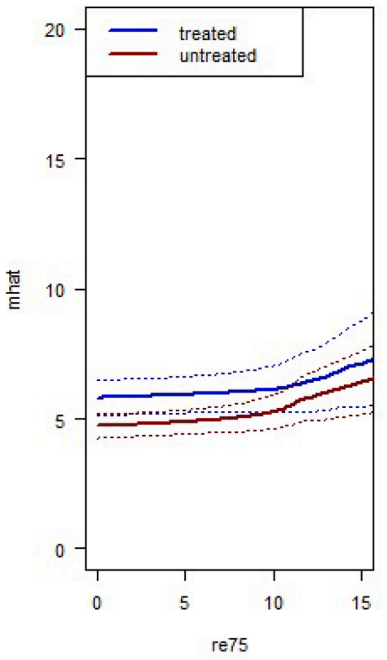

Fig. 1. Nonparametric fit of the conditional expectation by pre-treatment earnings and treatment status (cross validated bandwidth, Epanechnikov kernel).
Note: All graphs related to the empirical application are rescaled with all numbers denoted in '000$s.

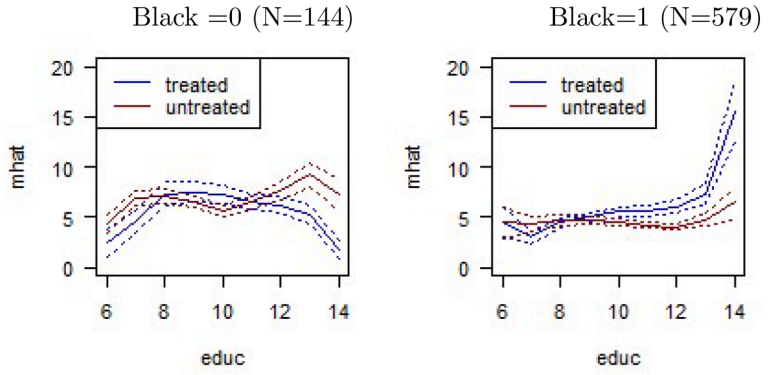

Fig. 2. Nonparametric fit of the conditional expectation by years of education, race, and treatment status with median pre-treatment earnings and age) (cross validated bandwidth, Epanechnikov kernel).
Note: The median pre-treatment earnings equals $936 and the median age is 23. The estimates are rescaled and are denoted in '000$.

For X_2 we consider two possibilities: (a) a N(0,1) random variable independent of X_1; (b) $X_2 = m_1(Z)$ where $Z(t)$ is a functional regressor similar to $X_1(t)$ with (a_i, b_i, w_i) replaced by (a_i', b_i', w_i') where the correlation between (a_i', b_i', w_i') and (a_i, b_i, w_i) is given by ρ (and set equal to either 0 or 0.8).

For the functional regressor X_1 we use the same metric as in Ferraty et al. (2007), that is $\|x_1 - X_1\|_1 = \sqrt{\int_{-1}^{1} \left(x_1'(t) - X_1'(t)\right)^2 dt}$. 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.

Table 3 shows RMSE of the NW estimator at the cross-validated bandwidths as well as RMSE where either the functional or scalar regressor is dropped. The loss from misspecifying the functional regression as univariate can be substantial.

6. Empirical study

The causal inference literature has made extensive use of the Lalonde (1986) data on the National Supported Work Demonstration (NSW) program following the release of that data by Dehejia and Wahba (1999, 2002). Their finding, that propensity score-based

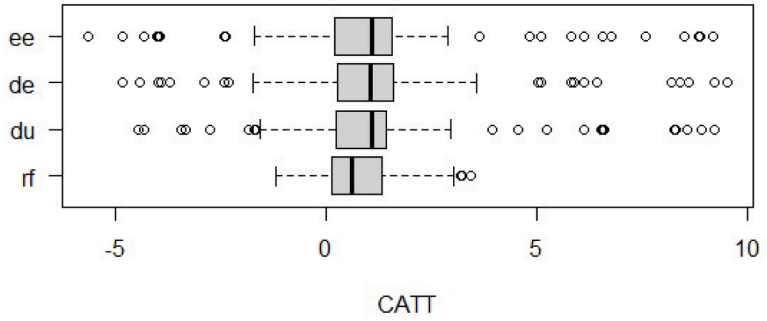

Fig. 3. Box-plots of the CATT estimates (NW and RF). *Note:* The estimates are rescaled and are denoted in '000$.

methods provide a way to generalize the experimental results on the impact of training to nonexperimental data, was influential and led to significant methodological advances and practical changes as discussed in the review by Imbens and Xu (2024). Here, we consider the experimental sample to analyze potential heterogeneous treatment effects using multivariate kernel estimation. Kernel-based matching on individual characteristics, advocated in Heckman et al. (1997, 1998), was not considered due to the claimed high dimensionality of the regressors. We show that it is both feasible and insightful for this data due to the dimension reduction implied by the presence of several discrete, discretized and categorical regressors. The mass point of the regressor on pre-treatment earnings at zero further contributes to the regressor singularity and we do not require continuity or indeed existence of density over positive values, thus kinks or mass at positive values are not excluded. The kernel estimator is applicable to such singular distributions.

We focus here on the full LaLonde NSW male sample which contains 297 treated individuals and 425 controls where the pre-intervention variables are well-matched.[8]

With Y denoting the post-treatment outcome, T the treatment and X the individual pre-treatment characteristic(s), we use the nonparametric regression model

$$m(x, j) = E(Y|X = x, T = j) \text{ for } j = 0, 1$$

to evaluate the heterogeneous effects of the treatment as

$$\tau(x) = m(x, 1) - m(x, 0).$$

The heterogeneous effect of treatment on the treated, also known as the conditional average treatment effect, CATT, is given by

$$\tau_T(x) = E(m(x, 1) - m(x, 0)|T = 1)$$

Focusing on the latter, we use the NW estimates to evaluate

$$\hat{\tau}_T(x_i) = \hat{m}(x_i, 1) - \hat{m}(x_i, 0) \quad i = 1, \ldots, n_T$$

for all treated individuals n_T (i.e., we use both the actual and the counterfactual treatment for our estimates).

First, we consider a bivariate kernel regression model where we only use the pre-treatment earnings (re75) as regressor X. Following that, we estimate the multivariate model with the full set of variables X, where in addition to the pre-treatment earnings we include years of education, high school "no degree" status, race, age, marital status, and pre-treatment unemployment status, u75. It is not unreasonable to attempt nonparametric estimation for this problem where the only truly continuous regressor is earnings (and possibly age and education) as singularity provides dimension reduction.

As with our simulations, we use the np package in R for the nonparametric estimation where we consider the Epanechnikov (e), Uniform (u) and discrete (d) kernel.[9] Bandwidth selection is based on cross validation and we consider the adaptive bandwidth selection approach that accounts for the masspoint. As was shown in our simulations the rate improvement associated with singularities does not require special attention to discrete variables to benefit from it.

Estimation results are reported in detail in the supplementary material (Appendix D). Below the main findings are summarized.

For the bivariate regression model, the cross validated bandwidths confirm that we should not smooth across treated and untreated observations and that local heterogeneous treatment effects as related to pre-treatment earnings are present. Fig. 1, displays estimates of the conditional expectation using the Epanechnikov kernel by treatment status and pre-treatment earnings together with the bootstrapped confidence bounds. It suggests that treatment for individuals at low levels of pre-treatment earnings, in particular, is beneficial.

The adaptive bandwidth results in a slightly better in-sample correlation between the post-treatment outcome, $re78$, and its fit (increasing from 0.2098 to 0.2116 (for OLS the correlation is 0.1697)); bandwidths obtained using non-masspoint-observations only are quite similar to those obtained when including the masspoints in this case. The NW estimates with the adaptive bandwidth

[8] In the sub-sample with 1974 earnings data in Dehejia and Wahba (1999) the distribution of 1975 earnings exhibits a significantly different mass at zero between the treated (68 %) and untreated (60 %).

[9] We use the discrete kernel proposed by Aitchison and Aitkin (1976), where $K((d - d_i)/h) = 1 - h$ if $d = d_i$, else h where $h \in [0, 1/2]$.

provide values of CATT that on average equal $920 (76), $920 (76), and $906 (80) (standard error in brackets) for the (e,e), (d,e), (d,u) kernels on (T, X), respectively.[10] For comparison, the local linear kernel based estimates on average equal $822 (49) with the (e,e) kernel, while the average of the CATT estimates based on random forest (RF) equal $848 (52). The CATT results of the kernel regression based approach are more variable than those obtained using the random forest approach. For observations at mass points, CATT estimates using adaptive bandwidth are closer to those obtained using the random forest based approach.

For the multivariate model the cross-validated bandwidths provide important insights. Firstly, even though pre-treatment earnings is still relevant, the bandwidth is much larger than in the baseline model for all kernels, suggesting a reduction of the heterogeneous impact with individual's pre-treatment earnings. The bandwidths selected for nodegree, hispanic and married are large, signaling that these variables are not relevant (these regressors are automatically smoothed out from the regression function). At the same time, the bandwidths for education and age imply a heterogeneous impact associated with those characteristics, although the size of the bandwidth for age is fairly large.

The inclusion of additional controls yields an improvement in the in-sample correlation between the post-treatment outcome and its fit. For the (e,e) kernel we see an increase in correlation from 0.210 in the bivariate model to 0.338 (for comparison, for OLS the correlation equals 0.209 when age squared is included as well); the results for the (d,e) and (d,u) kernel are comparable.

To highlight the heterogeneity of the treatment effect of education and its interplay with race, we display in Fig. 2 estimates of the conditional expectation by treatment status, years of education, and race for an individual with median age and pre-treatment earnings together with the bootstrapped confidence bounds.

The graph reflects a heterogeneity of the impact of treatment whereby the more educated individuals identified as black appear to benefit more from treatment than their nonblack counterparts. Gains of treatment arise where the confidence band around the estimated nonparametric fit $\widehat{m}(x,1)$ lies above that of $\widehat{m}(x,0)$; for non-black individuals this is at the middle range of education, for black individuals this starts around 10 years of education and is rising over that range. These results are further supported when evaluating the average CATT for black individuals across different levels of education (see supplemental material, Appendix D).

Box-plots of the CATT estimates for the multivariate model using the NW regression estimate and the RF estimates are presented in Fig. 3. The limit distributional results of Wager and Athey (2018) do not apply here as many components of X are not continuously distributed.

The kernel based regression CATT results remain more variable than those provided by the random forest approach, but their interquartile range is comparable. The NW kernel based estimates of the CATT on average exceed the RF based estimates: $\$1,045$ (107), $1,018 (108), and $1,019 (104) for the (e,e), (d,e) and (d,u) kernel on (T, X) against $794 (54) based on the random forest.

The NW based results are stable across kernel, give interpretable insights and with cross-validation make it possible to detect irrelevant regressors.

Funding

Victoria Zinde-Walsh gratefully acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) grant 253139.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the participants at the Econometric Study Group conference in Bristol, the Canadian Econometric Study Group conference, and Saraswata Chaudhuri 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. We thank the Associate Editor and three anonymous referees for their careful reading of the paper and very helpful comments and suggestions.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jeconom.2025.106168.

References

Ackerberg, D.A., Caves, K., Frazer, G., 2015. Identification properties of recent production function estimators. Econometrica 83, 2411–2451.
Ahlfors, L., 1966. Lectures on Quasiconformal Mappings. Princeton University Press.
Aitchison, J., Aitkin, C. G.G., 1976. Multivariate binary discrimination by the kernel method. Biometrika 63, 413–420.

[10] As discussed in the supplemental material (Appendix D), we denote the kernel with two arguments: the first argument denotes the kernel applied to all binary regressors (treat, u75, nodegree, black, hispanic, and married) and the second argument denotes the kernel applied to the other regressors (re75, educ, and age).

Angrist, J.D., Pischke, J.-S., 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.
Arulampalam, W., Corradi, V., Gutknecht, D., 2017. Modeling heaped duration data: an application to neonatal mortality. J. Econom. 200, 363–377.
Bai, J.,Ng, S., 2006. Confidence intervals for diffusion index forecasts and inference with factor–augmented regressors. Econometrica 74, 1133–1150.
Caldeira, J.F., Gupta, R., Torrent, H.S., 2020. Forecasting u.s. aggregate stock market excess return: do functional data analysis add economic value? Mathematics 8. https://doi.org/10.3390/math8112042. https://doi.org/10.3390/math8112042
Dehejia, R.H., Wahba, S., 1999. Causal effect in nonexperimental studies: reevaluating the evaluation of training programs. J. Am. Stat. Assoc. 94, 1053–1062.
Dehejia, R.H., Wahba, S., 2002. Propensity score-matching methods for nonexperimental causal studies. Rev. Econ. Stat. 84, 151–161.
Demir, S., Toktamis, O., 2010. On the adaptive Nadaraya–Watson kernel regression estimators. Hacet. J. Math. Stat. 39, 429–437.
Desmet, K., Parente, S.L., 2010. Bigger is better: market size, demand elasticity, and innovation. Int. Econ. Rev. 51, 319–333.
Donkers, A.C., Schafgans, M. M.A., 2008. Estimation and specification of semiparametric index models. Econ. Theory 24, 1584–1606.
Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman and Hall.
Ferraty, F., Mas, A., Vieu, P., 2007. Nonparametric regression on functional data: inference and practical aspects. Aust. N. Z. J. Stat. 49, 267–286.
Ferraty, F., Nagy, S., 2022. Scalar-on-function local linear regression and beyond. Biometrika 109, 439–455.
Ferraty, F., Vieu, P., 2004. Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. J. Nonparametr. Stat. 16, 111–125.
Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer, New York.
Gasser, T., Hall, P., Presnell, B., 1998. Nonparametric estimation of the mode of a distribution of random curves. J. R. Stat. Soc., Ser. B 60, 681–691.
Geenens, G., 2015. Moments, errors, asymptotic normality and large deviation principle in nonparametric functional regression. Stat. Probab. Lett. 107, 369–377.
Györfi, L., Kohler, M., Krzyzak, A., Walk, H., 2002. A Distribution-Free Theory of Nonparametric Regression. Springer, New York.
Hall, P., Horowitz, J., 2013. A simple bootstrap method for constructing nonparametric confidence bands for functions. Ann. Stat. 41, 1892–1921.
Hall, P., Li, Q., Racine, J.S., 2007. Nonparametric estimation of regression functions in the presence of irrelevant regressors. Rev. Econ. Stat. 89, 784–789.
Hall, P., Racine, J.S., 2015. Infinite order cross-validated local polynomial regression. J. Econom. 185, 510–525.
Heckman, J.J., Ichimura, H., Todd, P.E., 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Rev. Econ. Stud. 64, 605–654.
Heckman, J.J., Ichimura, H., Todd, P.E., 1998. Matching as an econometric evaluation estimator. Rev. Econ. Stud. 65, 261–294.
Hong, S., Linton, O., 2020. Nonparametric estimation of infinite order regression and its application to the risk-return tradeoff. J. Econom. 219, 389–424.
Hotelling, H., 1929. Stability in competition. Econ. J. 39, 41–57.
Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single index models. J. Econom. 58, 71–120.
Imbens, G., Xu, Y., 2024. LaLonde (1986) after nearly four decades: Lessons learned. arXiv:2406.00827 (econ.EM).
Jun, B., Song, H., 2019. Tests for detecting probability mass points. Korean Econ. Rev. 35, 205–248.
Kankanala, S., Zinde-Walsh, V., 2024. Kernel-weighted specification testing under general distributions. Bernoulli 30, 1921–1944.
Köhler, M., Schindler, A., Sperlich, S., 2014. A review and comparison of bandwidth selection methods for kernel regression. Int. Stat. Rev. 82, 243–274.
Kotlyarova, Y., Schafgans, M., Zinde-Walsh, V., 2016. Smoothness: bias and efficiency of non-parametric kernel estimators. Adv. Econom. 36 , 561–589.
Kurisu, D., Otsu, T., Xu, M., 2025. Nonparametric causal inference with functional covariates. J. Bus. Econ. Stat. pp. 1–14. https://doi.org/10.1080/07350015.2025.2501563. https://doi.org/10.1080/07350015.2025.2501563
LaLonde, R., 1986. Evaluation the econometric evaluations of training programs with experimental data. Am. Econ. Rev. 76, 604–620.
Li, C., Ouyang, D., Racine, J.S., 2009. Nonparametric regression with weakly dependent data: the discrete and continuous regressor case. J. Nonparametr. Stat. 21, 697–711.
Li, Q., Racine, J.S., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press.
Mandelbrot, B., 1997. Fractals and Scaling in Finance, Discontinuity, Concentration, Risk. Vol. E Springer.
Masry, E., 2005. Nonparametric regression estimation for dependent functional data: asymptotic normality. Stoch. Process. Appl. 115, 155–177.
Nadaraya, E., 1965. On non-parametric estimates of density functions and regression curves. Theory Probab. Appl. 10, 186–190.
Olson, C.A., 1998. A comparison of parametric and semiparametric estimates of the effect of spousal health insurance coverage on weekly hours worked by wives. J. Appl. Econom. 13, 543–565.
Pollard, D., 2001. A user's guide to measure theoretic probability. In: Cambridge Series in Statistical and Probabilistic Mathematics.
Racine, J., Li, Q., 2004. Nonparametric estimation of regression function with both categorical and continuous data. J. Econom. 119, 99–130.
Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer, New York.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Sain, S.R., 1994. Adaptive kernel density estimation. Comput. Stat. Data Anal. 39, 165–186.
Shen, G., 2002. Fractal dimension and fractal growth of urbanized areas. Int. J. Geograph. Inf. Sci. 16, 419–437.
Takayasu, M., Takaysu, H., 2009. Encyclopedia of Complex Systems in Finance and Econometrics. Springer. Fractals and Economics.
Vol'berg, A.L., Konyagin, S.V., 1988. On measures with the doubling condition. Math. USSR-Izvestiya 30, 629–638.
Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. 113, 1228–1242.
Wang, M.-C., Ryzin, J.V., 1981. A class of smooth estimators for discrete distributions. Biometrika 68, 301–309.
Watson, G.S., 1964. Smooth regression analysis. Sankhya 26, 359–372.