Biometrika (2021), **1**, 1, *pp*. 1–18
Printed in Great Britain
Advance Access publication on 19 March 2021

Scalar-on-function local linear regression and beyond

BY F. FERRATY
University of Toulouse, Toulouse Mathematics Institute, 118 route de Narbonne, 31062 Toulouse cedex, France
ferraty@math.univ-toulouse.fr

AND S. NAGY
Charles University, Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Prague, Czech Republic
nagy@karlin.mff.cuni.cz

SUMMARY

It is common to want to regress a scalar response on a random function. This paper presents results that advocate local linear regression based on a projection as a nonparametric approach to this problem. Our asymptotic results demonstrate that functional local linear regression outperforms its functional local constant counterpart. Beyond the estimation of the regression operator itself, local linear regression is also a useful tool for predicting the functional derivative of the regression operator, a promising mathematical object on its own. The local linear estimator of the functional derivative is shown to be consistent. For both the estimator of the regression functional and the estimator of its derivative, theoretical properties are detailed. On simulated datasets we illustrate good finite sample properties of the proposed methods. On a real data example of a single-functional index model we indicate how the functional derivative of the regression operator provides an original, fast, and widely applicable estimation method.

Some key words: Asymptotics; Functional data; Functional derivative of regression operator; Functional index model; Nearest neighbours method; Local linear regression; Scalar-on-function regression.

1. INTRODUCTION

Functional data analysis is a toolbox of statistical techniques for dealing with datasets of random functions (Ramsay & Silverman, 2002, 2005; Hsing & Eubank, 2015; Kokoszka & Reimherr, 2017). Regressing a scalar response Y on an explanatory random function X using the model $Y = m(X) + error$, with m unknown, is commonly termed scalar-on-function regression. In this setting, linear modelling (Cardot et al., 1999; Cai & Hall, 2006; Crambes et al., 2009) or its generalized versions (James, 2002; Müller & Stadtmüller, 2005) have been intensively studied in the literature. Although scalar-on-function linear regression is a powerful tool, its lack of flexibility when nonlinearities occur led the statistical community to develop a nonparametric approach. For an overview on the functional version of the Nadaraya-Watson kernel estimator, known also as the local constant regressor, see Ferraty & Vieu (2006). Since the development of local linear estimation for multivariate data in the 1990s (see for instance Fan, 1992; Fan & Gijbels, 1992; Fan, 1993; Ruppert & Wand, 1994; Cheng et al., 1997; Hall & Marron, 1997, and the monograph of Fan & Gijbels, 1996), it is well known that local linear regression outperforms the usual Nadaraya-Watson kernel estimator. Therefore, it became

perhaps the most popular nonparametric regression technique. Surprisingly, in the framework of functional data, there are only two papers that focus on the estimation of the regression operator in the scalar-on-function local linear regression model. In Baíllo & Grané (2009), a projection approach to the problem similar to the study presented here is proposed, but the asymptotics derived in that paper appears to suffer from a lack of rigour. The theoretical work in Berlinet et al. (2011) proposes an alternative estimating procedure by regularizing a non-bounded linear operator. The latter method does not directly relate to the approach taken here, as it does not rely on the use of projections. All in all, the theory and practice of scalar-on-function local linear regression is severely underdeveloped, and thus the method is far from being as popular as it is in the multivariate case.

This paper introduces local linear regression as a useful tool in the setting of scalar-on-function nonparametric regression. It turns out that *functional local linear regression*, that is, local linear regression when the regressor is a random function, is not only a convenient method of estimating the regression operator. As an exciting by-product we obtain an easy and fast method for estimating the functional derivative m'_x of the regression operator m at any function x. The functional derivative is a linear functional that represents a local linear approximation to the regression operator m around x; for a precise statement see (H1) below. In what follows, we use the Riesz representation theorem, and identify the functional derivative m'_x with its unique representing function. What makes the estimation of functional derivatives of such great interest? A first motivation is given in the pioneering works Hall et al. (2009) and Müller & Yao (2010) where estimating procedures are developed without considering the local linear regression setting. There, it was convincingly demonstrated that the concept of the functional derivative greatly facilitates the interpretation of results. As a further step in the pursuit for understanding how one can use the functional derivative in a natural way, let us consider the functional Taylor expansion of the regression operator. For a small positive real η and a direction u, that is a function u such that $\|u\| = 1$, Taylor's expansion and the Riesz representation theorem allow us to write $m(x + \eta\, u) - m(x) = \eta \langle m'_x, u\rangle + O(\eta^2)$. A first order approximation of the magnitude of the difference $m(x + \eta\, u) - m(x)$ is therefore the interval $[-\eta\|m'_x\|, \eta\|m'_x\|]$ — the smaller $\|m'_x\|$ is, the less sensitive to small perturbations at x is m. In a sense, $\|m'_x\|$ can be seen as a measure of reliability for the prediction of m at x. For another example where the functional derivative appears as a successful tool in interesting statistical problems, consider, for instance, the single-functional index model (Amato et al., 2006; Ait-Saïdi et al., 2008; Chen et al., 2011; Jiang & Wang, 2011). That model takes the form $m(x) = \mu + g\left(\langle \beta, x\rangle\right)$, where μ is an unknown scalar and the scalar response interacts with the functional covariate only through an unknown functional direction β combined with an unknown real-valued link function g. Extending the *average derivative estimation* method introduced in Härdle & Stoker (1989) to the functional setting, it is easy to show that $\mathrm{E}\left(m'_X\right)$ is proportional to the functional direction β. Thus, given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, as soon as one is able to obtain estimates $\hat{m}'_{X_1}, \ldots, \hat{m}'_{X_n}$ of the functional derivatives $m'_{X_1}, \ldots, m'_{X_n}$, one can compute $\hat{\mathrm{E}}\left(m'_X\right) \coloneqq n^{-1} \sum_i \hat{m}'_{X_i}$. The quantity $\hat{\mathrm{E}}\left(m'_X\right) / \|\hat{\mathrm{E}}\left(m'_X\right)\|$ is a reasonable estimator of the functional index β.

All our examples emphasize the major role that the functional derivative of the regression operator plays in important aspects of statistics: interpretation, reliability and methodology. This is why we propose to revisit functional local linear regression by focusing not only on the regression operator, but also on its functional derivative. In this work, a projection approach to functional local linear estimation is adopted. Theoretical properties of the estimator of the regression operator m are stated. Simultaneously, a local linear estimator of the functional derivative m'_x at x is proposed. For both estimators of the regression operator and its functional derivative, original tech-

nical tools are developed in order to study their theoretical behaviour, encompassing the nearest neighbours method for bandwidth selection. The nonparametric model combined with the functional setting make the derivation of asymptotic results challenging. The curse of dimensionality phenomenon discussed in Section 3 results in rates of convergence usually slower than the standard, polynomial rate in the sample size n. However, this drawback is mitigated because (1) this is the first time that rates of convergence are obtained when estimating the functional derivative of the regression operator, (2) polynomial rates are reachable in some useful function spaces, (3) nice finite sample properties of our local linear estimators are highlighted in a simulation study. Whilst the implementation of the estimator of the regression operator is straightforward, selection of the smoothing parameter for estimating its functional derivative poses a major challenge. This is why an ad hoc bootstrap procedure is introduced to pilot the bandwidth choice. The whole procedure, including the choice of the parameters for estimating the regression operator and its functional derivative, is fast and fully automated, and its implementation is user-friendly for practitioners. The paper is concluded with a benchmark real dataset analysis that is used to illustrate the important role of the functional derivatives in functional data analysis. The reader will find, in association with this document, a comprehensive online supplementary material, which includes efficient `R` (R Core Team, 2020) implementation of the newly proposed methods available at `https://bitbucket.org/StanislavNagy/fllr` or at request from the authors, additional simulations, hypotheses with discussion, proofs of the theorems, and some further theoretical considerations.

2. Functional local linear estimation

Let X be an H-valued random function where H is the separable Hilbert space of square integrable functions defined on $[0,1]$ equipped with the inner product $\langle\cdot,\cdot\rangle$, and let $\|\cdot\|$ be the associated norm. This work focuses on the relationship between X and a scalar response Y by considering the nonparametric regression model $Y = m(X) + \varepsilon$ where $\mathrm{E}(\varepsilon|X) = 0$. The regression operator m mapping H into $\mathbb{R}$ is unknown, and assumed smooth enough in a neighbourhood $\mathcal{N}_x$ of a given $x \in H$.

(H1) For any u in a neighbourhood of $0 \in H$, there exists $\zeta = x + t\,u$ with $t \in (0,1)$ such that

$$m(x+u) \;=\; m(x) + \langle m'_x, u\rangle + \frac{1}{2}\langle m''_\zeta u, u\rangle,$$

where $m'_x \in H$, m''_ζ is a Hilbert-Schmidt linear operator mapping H into H, and $v \mapsto m''_v$ is Lipschitz in $v \in \mathcal{N}_x$.

In other words, one focuses on those regression operators m for which the second order Taylor expansion is valid. Note that condition (H1) is the functional counterpart of what is standardly required in the finite-dimensional local linear regression setting. In particular, (H1) guarantees the continuity of the functional m at $x \in H$.

Based on an n-sample $(X_i,\, Y_i)_{i=1,\dots,n}$ of independent and identically distributed copies of $(X,\, Y)$, our main task is to estimate the regression operator m, as well as the functional derivative m'_x. To this end, one extends the sum of weighted squared errors to the functional setting

$$SWSE(a;\,\beta) \;\coloneqq\; \sum_{i=1}^{n} (Y_i - a - \langle\beta, X_i - x\rangle)^2\, K\left(h^{-1}\|X_i - x\|\right),$$

where $K(\cdot)$ is a kernel function defined on $[0,1]$ and h is a bandwidth, being a positive smoothing parameter. The principle of the local linear criterion is to linearise the regression operator in

a neighbourhood of x. In other words, for any X_i close to x, one considers that $\mathrm{E}(Y_i|X_i) = a + \langle \beta, X_i - x \rangle$. The real a and the square integrable function β can be interpreted as the value of the regression operator and its functional derivative at x, respectively. Both a and β depend on x; yet, to keep the notation simple this will not be accentuated in the text. The estimation of the functional derivative m'_x will be based on the estimation of the function β. Instead of focusing on β itself, it is computationally advantageous to consider its projection $\sum_{j=1}^{J} \langle \phi_j, \beta \rangle \phi_j$ onto the J-dimensional subspace $\mathcal{S}_J$ of H spanned by the orthonormal sequence $\phi_1, \ldots, \phi_J$ that is completed by $\phi_{J+1}, \phi_{J+2}, \ldots$ in order to get an orthonormal basis of H. In this notation the sum of weighted squared errors criterion is approximated by

$$SWSE_J(a;\, b_1, \ldots, b_J) \coloneqq \sum_{i=1}^{n} \left(Y_i - a - \sum_{j=1}^{J} b_j \langle \phi_j,\, X_i - x \rangle \right)^2 K\left(h^{-1} \| X_i - x \| \right),$$

where for any j, $b_j \coloneqq \langle \phi_j, \beta \rangle$. Then, $\hat{m}(x) \coloneqq \hat{a}$ and $\hat{m}'_x \coloneqq \sum_{j=1}^{J} \hat{b}_j \phi_j$ where $(\hat{a};\, \hat{b}_1, \ldots, \hat{b}_J) \coloneqq \arg\inf_{(a;\, b_1, \ldots, b_J)} SWSE_J(a;\, b_1, \ldots, b_J)$. By using vector and matrix notations, one is able to express the local linear estimators. Let $\mathrm{K} \coloneqq \mathrm{diag}\left\{ K\left(h^{-1} \| X_1 - x \| \right), \ldots, K\left(h^{-1} \| X_n - x \| \right) \right\}$ be a diagonal $n \times n$ matrix, $\mathrm{Y} \coloneqq [Y_1, \ldots, Y_n]^{\mathsf{T}}$, and Φ the following $n \times (J+1)$ matrix

$$\Phi \coloneqq \begin{bmatrix} 1 & \langle \phi_1,\, X_1 - x \rangle & \cdots & \langle \phi_J,\, X_1 - x \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \langle \phi_1,\, X_n - x \rangle & \cdots & \langle \phi_J,\, X_n - x \rangle \end{bmatrix}.$$

Define $\hat{\mathrm{b}} \coloneqq \left[\hat{b}_1, \ldots, \hat{b}_J \right]^{\mathsf{T}}$, $\Phi \coloneqq [\phi_1, \ldots, \phi_J]^{\mathsf{T}}$, 0 the $J \times 1$ null vector, $\boldsymbol{e}$ the $(J+1)$-dimensional vector $\left[1, 0^{\mathsf{T}}\right]^{\mathsf{T}}$ and I the $J \times J$ identity matrix. Then, it is easy to see that $\left[\hat{a} \,|\, \hat{\mathrm{b}}^{\mathsf{T}} \right]^{\mathsf{T}} = \left(\Phi^{\mathsf{T}} \mathrm{K} \Phi \right)^{-1} \Phi^{\mathsf{T}} \mathrm{KY}$, $\hat{m}(x) = \mathrm{e}^{\mathsf{T}} \left(\Phi^{\mathsf{T}} \mathrm{K} \Phi \right)^{-1} \Phi^{\mathsf{T}} \mathrm{KY}$ and $\hat{m}'_x = \Phi^{\mathsf{T}} \, [0|\mathrm{I}] \left(\Phi^{\mathsf{T}} \mathrm{K} \Phi \right)^{-1} \Phi^{\mathsf{T}} \mathrm{KY}$. The local linear approach has the nice property that both $\hat{m}(x)$ and $\hat{m}'_x$ are based on the common terms $\left(\Phi^{\mathsf{T}} \mathrm{K} \Phi \right)^{-1} \Phi^{\mathsf{T}} \mathrm{KY}$ and this is why the raw computational cost of $\hat{m}'_x$ is not much higher than the one for $\hat{m}(x)$.

3. ASYMPTOTIC STUDY

Assumptions (H2)–(H6) and related discussion are gathered in Section S.2·1 in the online supplement.

Let X stand for the sample $X_1, \ldots, X_n$ and write E_{X} and $\mathrm{Var}_{\mathrm{X}}$ for the conditional expectation and the conditional variance with respect to X, respectively. The asymptotic conditional bias and variance of $\hat{m}(x)$ are provided in the following theorem.

THEOREM 1. *Under conditions (*H1*)–(*H6*),*

(i) $\mathrm{E}_X\{\hat{m}(x)\} = m(x) + O_P\left(\| \mathcal{P}_{\mathcal{S}_J^{\perp}} m'_x \| \, h \right) + O_P\left(h^2 \right)$,

(ii) $\mathrm{Var}_X\{\hat{m}(x)\} = O_P\left[\{ n \, \pi_x(h) \}^{-1} \right]$,

where $\mathcal{S}_J^{\perp}$ is the orthogonal complement to the space $\mathcal{S}_J$ in H, $\mathcal{P}_{\mathcal{S}_J^{\perp}} : H \to \mathcal{S}_J^{\perp}$ is the orthogonal projection onto $\mathcal{S}_J^{\perp}$, and $\pi_x(h) \coloneqq P\left(\| X_i - x \| \leq h \right)$.

In the same situation, the conditional bias of the estimated functional local constant regression is of order h with the same conditional variance, see for instance Ferraty & Vieu (2006). The functional setting requires that the dimension J tends to infinity with the sample size n so that, the quantity $\|\mathcal{P}_{\mathcal{S}_J^\perp} m'_x\|$ tends to zero when n grows to infinity since m'_x is a square integrable function. Therefore $\hat{m}(x)$ outperforms the asymptotic behaviour of the kernel estimator in the functional local constant regression. Note that the conditional variance in THEOREM 1 involves the small ball probability $\pi_x(h)$, which is standard in the functional nonparametric setting (Ferraty & Vieu, 2006).

The functional setting involves implicitly the dimension J of the approximating subspace $\mathcal{S}_J$ in the rate of convergence. Indeed, the asymptotic behaviour of the conditional bias depends on the quantity $\|\mathcal{P}_{\mathcal{S}_J^\perp} m'_x\|$, which assesses the approximation error of the functional derivative m'_x in $\mathcal{S}_J$. From a theoretical point of view, $\hat{m}(x)$ involves $J \times J$ matrices and J-dimensional vectors with J converging to infinity. This makes the asymptotic study much harder in comparison with the finite-dimensional setting. The issue of infinite dimension is overcome by deriving accurately, in an element-wise sense, the asymptotic behaviour of the matrices and vectors involved in both local linear estimators.

Once the theoretical properties of $\hat{m}(x)$ have been given, a natural and interesting issue concerns the asymptotic behaviour of the functional derivative $\hat{m}'_x$ of m at x. The next result details the conditional bias and variance of $\hat{m}'_x$. Set $\gamma_{j,k}^{1,1}(t) = \mathrm{E}\left(\langle \phi_j, X_1 - x\rangle \langle \phi_k, X_1 - x\rangle | \|X_1 - x\|^2 = t\right)$; let $\gamma_{j,k}^{1,1'}$ be its derivative, and set λ_J the smallest eigenvalue of the $J \times J$ matrix Γ, whose (j,k)-th element is defined by $[\Gamma]_{jk} \coloneqq \gamma_{j,k}^{1,1'}(0)$.

THEOREM 2. *As soon as (*H1*)–(*H6*) are fulfilled, conditionally on $X_1, \ldots, X_n$,*

$$\|\hat{m}'_x - m'_x\| = O_P\left(\lambda_J^{-1} \|\mathcal{P}_{\mathcal{S}_J^\perp} m'_x\|\right) + O_P\left(\lambda_J^{-1} h\right) + O_P\left[h^{-1}\{\lambda_J^2\, n\, \pi_x(h)\}^{-1/2}\right] + O_P\left[h^{-1}\{\lambda_J\, n\, \pi_x(h)\}^{-1/2}\sqrt{J}\right].$$

The first two summands in the formula above correspond to the conditional bias of $\hat{m}'_x$, while the remaining terms come from its conditional variance. As n tends to infinity, h tends to zero, the dimension J goes to infinity, and the smallest eigenvalue λ_J to zero; see the comments on (H3) in Section S.2·1. Therefore, the rate of convergence of $\hat{m}'_x$ is slower than that of $\hat{m}(x)$. Compared to THEOREM 1, one h is removed in the conditional bias, and h^{-1} is added to the terms that relate to the conditional variance, which corresponds to the standard degradation of the convergence rate observed in the multivariate case. But the functional setting adds specific terms like J and λ_J which deteriorate the asymptotic behaviour. Nevertheless, as pointed out below, the finite sample properties of $\hat{m}'_x$ are surprisingly good.

The curse of dimensionality is a well known drawback in nonparametric multivariate regression. The optimal global rate of convergence of the regression estimator is polynomial in n, and deteriorates drastically with an increasing dimension of the space where the predictor takes its values, see for instance Stone (1982). In the general functional setting, when considering X as an infinite-dimensional predictor, the small ball probability $\pi_x(h)$ tends to zero faster than any polynomial in h. Therefore, the rate of convergence of the functional kernel estimator in local constant scalar-on-function nonparametric regression is slower than the inverse of any polynomial in n (see Ferraty & Vieu, 2006, Chap. 13, and Hall et al., 2009). This is still true for the

scalar-on-function local linear regression framework although the introduction of the quantities J, λ_J and $\|\mathcal{P}_{\mathcal{S}_J^\perp} m'_x\|$ in the asymptotic behaviour of the estimators makes the reading of the rates of convergence not easy. Nevertheless, some particular situations lead to usual rates of convergence when balancing the leading bias and variance terms.

This is the case when X is an $\mathcal{S}_D$-valued random function with $D \geq 1$ a fixed integer, i.e. $D = J$ does not depend on n. In practice, the case when the functional data can be accurately approximated in a finite-dimensional space is not rare. It may correspond to observation of smooth curves with common shapes; for instance, spectrometric profiles, or growth curves and their derivatives as the studied benchmark Berkeley growth data in Section 4·3 below. In this situation, $\hat{m}(x) - m(x) = O_P\left\{n^{-2/(D+4)}\right\}$, and $\|\hat{m}'_x - m'_x\| = O_P\left\{n^{-1/(D+4)}\right\}$. Considering X as a random function valued in a finite-dimensional subspace, i.e. $X = \sum_{j=1}^{D} \langle X,\, \phi_j\rangle\, \phi_j$, allows to recover the rates of convergence derived from those obtained in the standard J-multivariate local linear regression setting (see, for instance, Ruppert & Wand, 1994).

Let us now consider the recently introduced concept of the mixture inner product space (MIPS) and the notion of a MIPS-valued random variable (Lin et al., 2018). It provides an elegant framework to deal with the density in infinite-dimensional function spaces. Then, $\hat{m}(x) - m(x) = O_P\left\{n^{-2/(D+4)}\right\}$, and if λ_J is asymptotically equivalent to $J^{-\mu}$ with $1/2 < \mu < 1$, $\|\hat{m}'_x - m'_x\| = O_P\left\{n^{(-1+\mu)/(D+5-\mu)}\right\}$. This infinite-dimensional setting is equivalent to the finite-dimensional setting for $\hat{m}(x)$ and produces an acceptable rate of convergence for $\hat{m}'_x$. It confirms that MIPS is a relevant framework for functional data that can bypass the curse of dimensionality issue.

The reader will find more details with respect to both previous situations in the online supplement, see Section S.2·7, COROLLARIES S.2 and S.3.

An interesting alternative to global bandwidth is the one based on k-nearest neighbours method. From the precursor works of Stone (1977a), Devroye (1978), Devroye (1981) and Mack (1981), see also the book of Györfi et al. (2006), it is well known that the k-nearest neighbours (kNN) method is a relevant tool to build local bandwidths. It provides flexible estimators with easy implementation. This is why the kNN bandwidth, i.e., the local linear estimator at x involving only its k-nearest neighbours, is proposed in the implementation below. Formally, when focusing on the estimation at x, the kNN bandwidth H_k is the smallest bandwidth h such that the number of functional regressors X_i belonging to the ball $B(x,h)$ of radius h centred at x is equal to k. Although the consistency of the kNN kernel estimator for scalar-on-function nonparametric regression reproduces arguments similar to those used in the standard multivariate setting (see Burba et al., 2009), for scalar-on-function local linear regression one has to propose new asymptotic developments. First, a simple trick is used. Instead of using H_k, we introduce, for any i from 1 to n, the kNN bandwidth $H_{k,-i}$ obtained without considering the ith functional regressor X_i, that is $H_{k,-i} := \inf\left\{h\colon \sum_{\ell=1,\ell\neq i}^{n} 1_{\{X_\ell \in B(x,h)\}} = k\right\}$. The kNN estimators of $m(x)$ and m'_x are those obtained when h is replaced with $H_{k,-i}$, i.e. $\hat{m}_{kNN}(x) = \mathrm{e}^{\mathsf{T}}\left(\Phi^{\mathsf{T}}\mathrm{K}_{kNN}\Phi\right)^{-1}\Phi^{\mathsf{T}}\mathrm{K}_{kNN}\mathrm{Y}$ and $\hat{m}'_{x,\,kNN} = \Phi^{\mathsf{T}}\,[0|\mathrm{I}]\left(\Phi^{\mathsf{T}}\mathrm{K}_{kNN}\Phi\right)^{-1}\Phi^{\mathsf{T}}\mathrm{K}_{kNN}\mathrm{Y}$ with $\mathrm{K}_{kNN} := \mathrm{diag}\left\{K\left(H_{k,-1}^{-1}\|X_1 - x\|\right), \ldots, K\left(H_{k,-n}^{-1}\|X_n - x\|\right)\right\}$.

THEOREM 3. *Under conditions (*H1*)–(*H3*) and (*H5*)–(*H7*), conditionally on* $X_1, \ldots, X_n$,

(i) $\hat{m}_{kNN}(x) - m(x) = O_P\left\{\|\mathcal{P}_{\mathcal{S}_J^\perp} m'_x\|\, \pi_x^{-1}(k/n)\right\} + O_P\left\{\pi_x^{-1}(k/n)^2\right\} + O_P\left(1/\sqrt{k}\right)$,

(ii) $\|\hat{m}'_{x,\,kNN} - m'_x\| = O_P\left(\lambda_J^{-1}\|\mathcal{P}_{\mathcal{S}_J^{\perp}} m'_x\|\right) + O_P\left\{\lambda_J^{-1}\pi_x^{-1}(k/n)\right\}$
$+ O_P\left[\left\{\pi_x^{-1}(k/n)\lambda_J\sqrt{k}\right\}^{-1}\right] + O_P\left[\sqrt{J}\left\{\pi_x^{-1}(k/n)\sqrt{\lambda_J\, k}\right\}^{-1}\right],$

where $\pi_x^{-1}(u) \coloneqq \inf\{h\colon \pi_x(h) \geq u\}$ *is the generalized inverse function of* $\pi_x(\cdot)$.

According to the conditions on k and k/n, see (H7) in Section S.2·4 in the online supplement, and the definition of $\pi_x^{-1}(\cdot)$, the quantities $\pi_x^{-1}(k/n)$ and $1/\sqrt{k}$ go to 0 as n tends to infinity. When comparing the rates of convergence with those given in THEOREMS 1 and 2, $\pi_x^{-1}(k/n)$ is the local counterpart of the global bandwidth h, and k plays the role of $n\,\pi_x(h)$. This interpretation is not surprising. Thanks to LEMMA S.5 in the companion online supplement, $H_k \underset{a.s.}{\sim} \pi_x^{-1}(k/n)$ and $n\,\pi_x(H_k) \underset{a.s.}{\sim} k$. When going back to the particular situations involved in the discussion on the curse of dimensionality, i.e. the setting of COROLLARIES S.2 and S.3, $\pi_x^{-1}(k/n) = O\left\{(k/n)^{1/D}\right\}$. THEOREM 3-*(i)* then provides the rate of convergence obtained in the standard multivariate setting, see for instance Mack (1981)

The functional local linear estimator depends on the first basis elements $\phi_1, \ldots, \phi_J$ of the space H of square integrable real functions on $[0, 1]$. When considering the orthogonal B-spline basis (see de Boor, 1978, and Redd, 2012), as soon as $|m_x^{\prime(p)}(u) - m_x^{\prime(p)}(v)| \leq C|u - v|^{\nu}$ with $\nu \in [0, 1]$, we have $\|\mathcal{P}_{\mathcal{S}_J^{\perp}} m'_x\| = O\left(J^{-p-\nu}\right)$. In addition, if m'_x is a periodic function, the use of the Fourier basis leads to the same result and the bias of $\hat{m}(x)$ becomes in both cases $\mathrm{E}_{\mathrm{X}}\{\hat{m}(x)\} = m(x) + O_P(h\, J^{-p-\nu}) + O_P(h^2)$ whereas the variance does not change. The reader will find more details in Section S.2·6 in the supplementary material. Another useful example is the data driven basis derived from the functional principal components analysis (FPCA) of the random function X (see Karhunen, 1946; Loève, 1946; Rao, 1958; Dauxois et al., 1982 for precursor works and Bosq, 2000; Yao et al., 2005; Hall & Hosseini-Nasab, 2006; Hall et al., 2006 for more recent statistical developments). In this setting, functions ϕ_j are the eigenfunctions of the covariance operator of X and the eigenanalysis of the empirical covariance operator provides a data driven basis $\hat{\phi}_1, \hat{\phi}_2, \ldots$. By convention, we assume that $\langle \phi_j, \hat{\phi}_j \rangle > 0$. This results in the new estimator $\hat{\hat{m}}(x) \coloneqq \mathrm{e}^{\mathsf{T}}\left(\hat{\Phi}^{\mathsf{T}}\mathrm{K}\hat{\Phi}\right)^{-1}\hat{\Phi}^{\mathsf{T}}\mathrm{KY}$ where, for $i = 1, \ldots, n$, $[\hat{\Phi}]_{i1} = 1$ and for $j = 2, \ldots, J$, $[\hat{\Phi}]_{ij} = \langle \hat{\phi}_j,\, X_i - x \rangle$. THEOREM S.1 in the companion supplement gives the rate of convergence of $\hat{\hat{m}}(x)$.

4. IMPLEMENTATION

4·1. *Selection of tuning parameters*

In this section, we discuss the practical aspects and assess the finite sample performance of our local linear estimators for functional data. Beyond standard issues such as the choice of the tuning parameters in the estimation of the regression operator, a novel heuristic is developed for the selection of the bandwidth for the functional derivative. Later, in a comparative simulation study in Section 4·2, the finite sample performance of our estimator is contrasted with its competitors available in the literature. We conclude with a real data example and an application of the functional local linear estimator in Section 4·3. On a benchmark growth dataset we demonstrate a strong link between the considered scalar-on-function local linear regression, and the important single-functional index model, widely considered in the literature.

According to the definition of $\hat{m}(x)$ and $\hat{m}'_x$, two parameters have to be selected: the dimension J of the approximating subspace and the bandwidth h. Theoretical results of THEOREMS 1 and 2 emphasize different asymptotic behaviours for the estimators of the regression functional and its functional derivative. This is why the optimal parameters for estimating the regression operator do not match necessarily those designed for the functional derivative.

Choosing optimal parameters for the functional derivative is quite challenging because no quantity is directly available to compare with. The following ad hoc methodology is proposed. Firstly, optimal parameters h_{reg} and J_{reg} for the estimator of the regression operator $\hat{m}$ are selected. The `R` implementation of the estimating procedure that can be found in the online supplementary material allows for the use of two standard criteria — the (leave-one-out) cross-validation (see for instance Allen, 1974; Stone, 1974; Stone, 1977b) and an adaptation of the corrected Akaike information criterion (Akaike, 1973; Hurvich et al., 1998) to functional data. A general overview of these criteria can be found in Hastie et al. (2001). Secondly, the parameters h_{reg} and J_{reg} are used to build, for each random function X_i in the sample, a pilot estimator $\hat{m}'^{boot}_{X_i}$ of the functional derivative at X_i by means of a wild bootstrap procedure (Wu, 1986; Mammen, 1993; Ferraty et al., 2010). Then, the optimal parameters h_{deriv} and J_{deriv} are those minimizing the mean squared error

$$n^{-1}\sum_{i=1}^{n}\left\|\hat{m}'^{boot}_{X_i} - \hat{m}'_{X_i,-i}\right\|^2, \tag{1}$$

where $\hat{m}'_{X_i,-i}$ is the estimator of the functional derivative at X_i from a dataset with the ith observation removed. The finite-sample behaviour of the estimator $\hat{m}'_{X_i}$ observed with our simulated data indicates that h_{reg} underestimates h_{deriv}, see Figure 2(b) below. In other words, the estimator $\hat{m}'_{X_i}$ based on h_{reg} is not smooth enough. The introduced pilot estimator is the average of B bootstrapped estimators $\hat{m}'^{(1)}_{X_i}, \ldots, \hat{m}'^{(B)}_{X_i}$ based on h_{reg}. If each bootstrapped version $\hat{m}'^{(b)}_{X_i}$ of $\hat{m}'_{X_i}$ is artificially too rough, one can expect that their average, i.e. the pilot estimator $\hat{m}'^{boot}_{X_i}$, is smooth enough so that equation (1) provides a reasonable bandwidth h_{deriv}.

An important methodological point consists in translating the bandwidths into the number of nearest neighbours which results in the kNN estimators detailed in Section 3. An advantage of such a local approach is that it provides more flexible estimators while reducing a continuous set of candidates to a discrete one. In our implementation, all bandwidths are expressed in terms of nearest neighbours.

In order to assess the quality of the bandwidth choice for both estimators, a first model is simulated so that the regression operator, as well as the functional derivatives, can be expressed analytically.

Let $X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+500}$ be independent and identically distributed copies of a functional predictor X. The estimators are based on the training set $X_1, \ldots, X_n$; the 500 remaining functions $X_{n+1}, \ldots, X_{n+500}$ are used to assess the quality of the estimators. The model takes the form $Y := m(X) + \varepsilon$, where ε is an independent, centred and normally distributed error term with variance σ^2_ε. Let $\phi_1, \ldots, \phi_4$ be the first four elements of the Fourier basis; the random function X is equal to the linear combination $\sum_{j=1}^{4} U_j\,\phi_j$ where U_j are independent and identically distributed uniform random variables on $[-1, 1]$. The regression operator is given by $m(X) := \sum_{j=1}^{4}\exp(-U_j^2)$. From the expression for m, one can derive its functional derivative at x that takes the form $m'_x(t) = -2\sum_{j=1}^{4} U_j\,\exp(-U_j^2)\,\phi_j(t)$. Here, the size of the approximating subspace is $J = 4$. We consider $n = 100, 150, \ldots, 500$. Finally, the noise-to-signal ra-

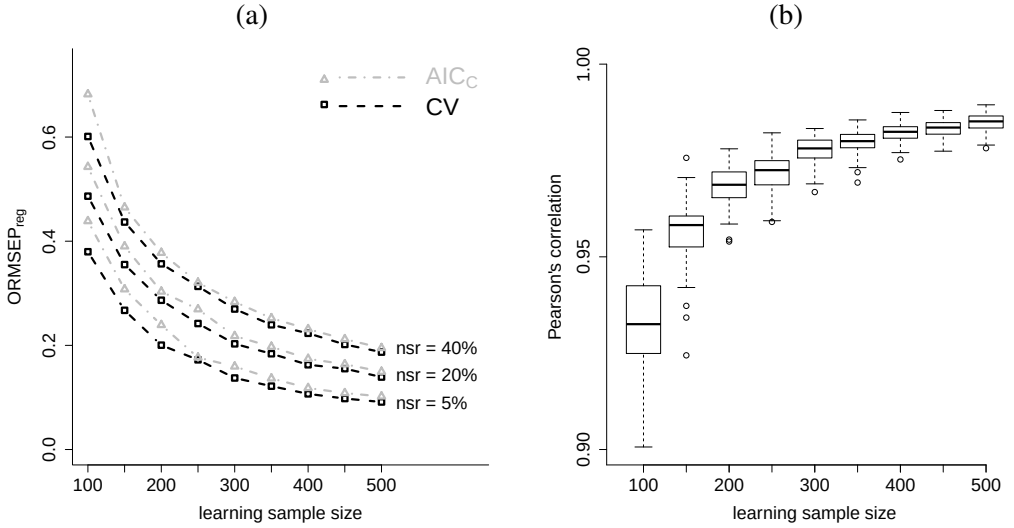

Fig. 1: For different learning sample sizes, (a) in squares and triangles we see the means of $ORMSEP_{reg}$ over 100 runs, depending on noise-to-signal ratios nsr and bandwidth selection methods CV/AIC$_C$; and (b) empirical correlations between true values and their predictions obtained with AIC$_C$ and nsr=0.05.

tio (nsr) is also controlled by setting $\sigma_\varepsilon^2 := nsr \times \mathrm{Var}\{m(X)\}$, with $nsr = 0.05, 0.2, 0.4$. (M1) refers to this simulated model. Because we focus on the bandwidth selection, in the present model in the estimation procedure we deliberately use the true approximating subspace, i.e. $J = 4$ and the first four Fourier basis elements $\phi_1, \ldots, \phi_4$.

The first step is to assess the quality of the bandwidth selection for estimating the regression operator. To this end, the optimal bandwidth h_{reg} is the one minimizing the CV (cross-validation) or AIC$_C$ (corrected Akaike information criterion). Because in (M1) the true regression operator m is known, one can compute the oracle relative mean squared error of prediction

$$ORMSEP_{reg} := \sum_{i=n+1}^{n+500} \{m(X_i) - \hat{m}(X_i)\}^2 \; / \; \sum_{i=n+1}^{n+500} \{m(X_i) - \overline{m}(X)\}^2 \qquad (2)$$

where $\overline{m}(X) := 500^{-1} \sum_{i=n+1}^{n+500} m(X_i)$. Repeating the simulation scheme 100 times in various situations, Figure 1 assesses the quality of the local linear estimator of the regression operator. The consistency of the estimator appears clearly in Figure 1(a) even for a large noise-to-signal ratio. Concerning the bandwidth selection, both methods CV and AIC$_C$ provide similar results, although CV seems to outperform AIC$_C$ slightly. Each boxplot in Figure 1(b) corresponds to the distribution of 100 empirical correlations computed on 100 scatterplots $\{(m(X_i), \hat{m}(X_i)) \colon i = 1, \ldots, n\}$. It reflects high quality of prediction even for small sample size and the consistency.

The selection of the bandwidth for estimating the functional derivative is much more challenging because there is no standard criterion to minimize. An original bandwidth selection based on the wild bootstrap procedure is proposed. It aims to build a pilot estimator of the functional derivative:

(i) use h_{reg} for estimating the model error $\hat{\varepsilon}_i := Y_i - \hat{m}(X_i)$ for $i = 1, \ldots, n$,

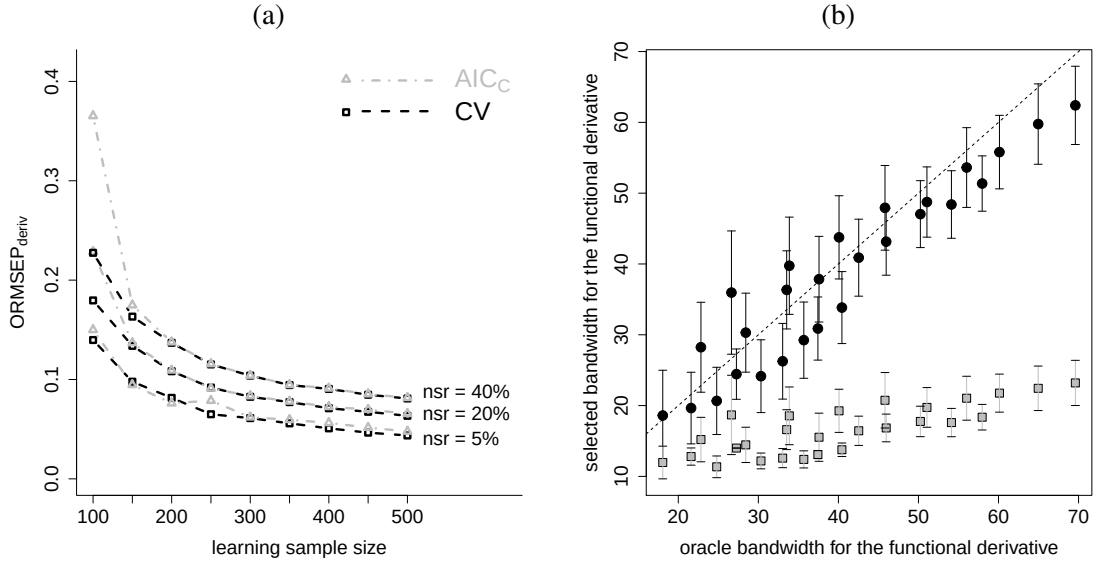

Fig. 2: (a) In squares and triangles we see the mean of $ORMSEP_{deriv}$ over 100 runs, depending on different noise-to-signal ratios nsr, learning sample sizes, and bandwidth selection methods CV/AIC$_C$; (b) h_{deriv}^{oracle} versus h_{deriv}, and h_{deriv}^{oracle} versus h_{reg}. Solid circles and squares represent respectively the averages of h_{deriv} and h_{reg} over 100 runs for each of the 9×3 pairs $(n,\, nsr)$; whiskers stand for standard deviation. On the horizontal axis, averages of the oracle bandwidths h_{deriv}^{oracle} are displayed.

(ii) given independent and identically distributed centred random variables $V_1, \ldots, V_n$ independent of $\hat{\varepsilon}_i$ such that their first moments equal 1, compute the bootstrapped errors $\varepsilon_i^{(b)} := \hat{\varepsilon}_i \times V_i$,

(iii) derive a bootstrapped sample $\mathcal{S}^{(b)} := \left\{X_i,\ Y_i^{(b)} := \hat{m}(X_i) + \varepsilon_i^{(b)}\right\}_{i=1,\ldots,n}$ and compute the bootstrapped estimator $\hat{m}'^{(b)}_{X_i}$ from $\mathcal{S}^{(b)}$.

Repeat steps (ii) and (iii) independently B times and denote $\hat{m}'^{boot}_{X_i} := B^{-1}\sum_{b=1}^{B} \hat{m}'^{(b)}_{X_i}$ that we name the pilot estimator of the functional derivative at X_i. The optimal bandwidth h_{deriv} is defined as the one minimizing (1). In order to assess the relevance of this bandwidth choice, a simulation study with model (M1) is conducted with $B = 100$. Similarly as in the study of the estimator of the regression operator, Figure 2(a) displays the means of the oracle relative mean squared error

$$ORMSEP_{deriv} := \sum_{i=n+1}^{n+500} \left\|m'_{X_i} - \hat{m}'_{X_i}\right\|^2 \Big/ \sum_{i=n+1}^{n+500} \left\|m'_{X_i} - \overline{m}'_X\right\|^2 \tag{3}$$

where $\overline{m}'_X := 500^{-1}\sum_{i=n+1}^{n+500} m'_{X_i}$. Firstly, selecting h_{deriv} with CV or AIC$_C$ has no significant impact on the prediction quality of the estimator of functional derivatives. Secondly, this plot demonstrates the consistency of the local linear estimator of the functional derivative. Its rate of convergence seems to be slightly slower than the one observed for the local linear estimator of the regression operator, as supported by the asymptotic results. Another way to assess the bandwidth choice for estimating the functional derivative is to compare the selected bandwidth h_{deriv} itself with the oracle one h_{deriv}^{oracle} minimizing the oracle relative mean squared error $ORMSE_{deriv}$

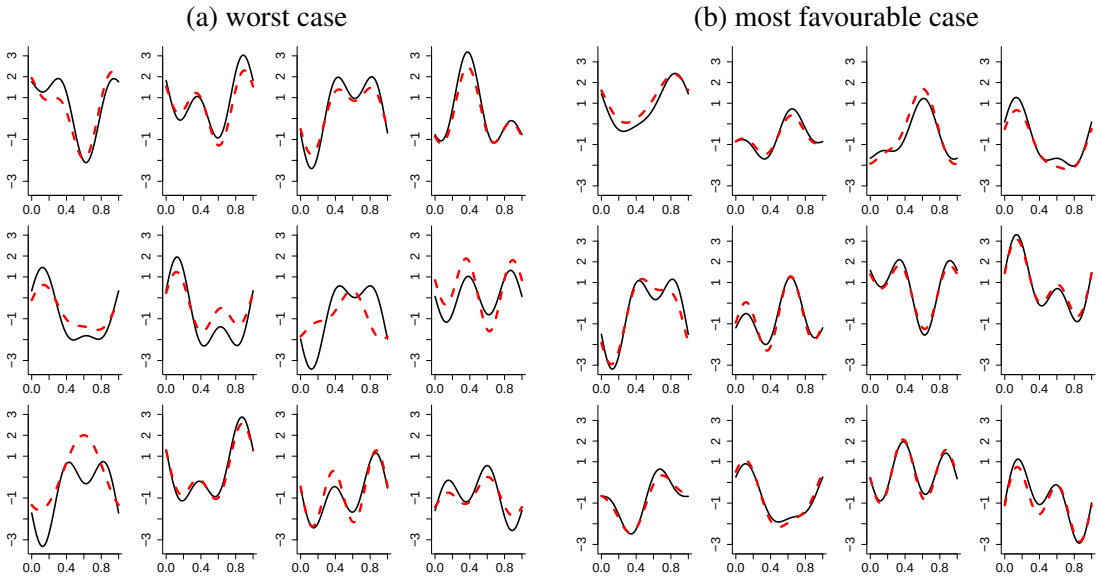

Fig. 3: Functional derivatives m'_{X_i} in solid lines and their predictions $\hat{m}'_{X_i}$ with h_{deriv} in dashed lines.

from (3) computed from the random sample functions $X_1, \ldots, X_n$. Given a learning sample size n and a noise-to-signal ratio nsr, 100 simulated datasets are drawn from (M1). Our local linear estimating procedure provides 100 triples of bandwidths h_{reg}, h_{deriv} and h_{deriv}^{oracle}. In Figure 2(b), h_{deriv} works quite well even if it underestimates slightly the oracle bandwidth h_{deriv}^{oracle}. On the other hand, h_{reg} fails drastically, especially for larger sample sizes n. That illustrates the importance of selecting a specific bandwidth for the estimation of functional derivatives. Figure 3 compares a selection of true functional derivatives $m'_{X_1}, \ldots, m'_{X_n}$ with the corresponding predictions $\hat{m}'_{X_1}, \ldots, \hat{m}'_{X_n}$ (a) in the worst case of $n = 100$ and $nsr = 0.4$, and (b) in the most favourable situation of $n = 500$ and $nsr = 0.05$. Even in the worst situation, the predictions remain adequate. It is worth noting that the bootstrap bandwidth selection introduces additional randomness into the local linear estimation of the functional derivatives. However, Table S.1 given in the supplement indicates clearly that our procedure provides very stable results.

To summarize this section devoted to bandwidth selection, we may conclude that: (i) cross–validation is a useful method for determining h_{reg}, the bandwidth for estimating the local linear regression operator, and (ii) the bootstrap procedure which provides the bandwidth for estimating the functional derivative works well, even for small learning samples size and high noise-to-signal ratios.

What about the automatic choice of the approximation subspace? So far we focused only on the bandwidth selection. To make our method fully automatic, one has to determine also the approximating subspace spanned by $\phi_1, \ldots, \phi_J$ and its dimension J. As explained in the last part of Section 3, functional principal component analysis is a very useful tool for expanding a random function onto the eigenfunctions of the covariance operator. Let $\phi_1, \ldots, \phi_J$ be the eigenfunctions of the covariance operator of the functional predictor X associated to the J largest eigenvalues, and let $\hat{\phi}_1, \ldots, \hat{\phi}_J$ be their estimates from the empirical covariance operator. To make the estimating procedure fully automatic, we may proceed in three steps: (i) compute the first J eigenfunctions $\hat{\phi}_1, \ldots, \hat{\phi}_J$ of the empirical covariance operator with J large enough,

(ii) carry out the local linear estimation of the regression operator with h_{reg} and J_{opt} obtained by minimizing the CV criterion with respect to $J_{opt} \in \{0, 1, \ldots, J\}$ and h_{reg}, (iii) determine h_{deriv} using the bootstrap procedure by minimizing (1) and compute the corresponding local linear estimator of the functional derivative. The reader will find in Section S.1·3 of the supplementary document a comprehensive study assessing the robustness with respect to the estimation of J_{opt}.

4·2. *A comparative study*

We now conduct a simulation study, in which the finite sample performance of the local linear estimator is compared to its competitors from the literature. In the simulated datasets, we extend model (M1) to consider a whole spectrum of scenarios, from a linear to a nonlinear additive one.

To make the estimating procedure more challenging, we add structural perturbation to the functional predictors X. Let $\phi_1, \ldots, \phi_{2D}$ be the first $2D$ elements of the Fourier basis; for an integer $D \geq 1$ given, a noisy functional predictor $X \coloneqq \sum_{j=1}^{D} U_j \, \phi_j + \eta$ is built with $\eta \coloneqq \sum_{j=D+1}^{2D} V_j \, \phi_j$ where U_j and V_j are independent and identically distributed uniform random variables defined on $[-1, 1]$ and $[-b, b]$, respectively. The second part η provides a structural noise that is controlled by the ratio $\rho \coloneqq \mathrm{E}\left(\|\eta\|^2\right) / \mathrm{E}\left(\|X\|^2\right) = b^2/(1+b^2)$. Given any $\rho \in (0, 1)$, one can always find a corresponding bound b for simulating the functional predictors. We now consider $Y \coloneqq m_a(X) + \varepsilon$ where $m_a(X) \coloneqq (1-a)\langle \beta, X \rangle + a \sum_{j=1}^{D} \exp(-U_j^2)$ and $\beta \coloneqq \sum_{j=1}^{D} \phi_j$. The choice $a = 0$ corresponds to a standard functional linear model, whereas $a = 1$ represents a nonlinear regression model. For all choices of a, this simulated model termed (M2) is additive (Müller & Yao, 2008, 2010), meaning that the regression operator can be expressed as a sum of components where each component is a function that depends only on a single principal score of the regressor X. Additivity allows direct computation of the functional derivative of m_a; it takes the form $m'_{a,x}(t) = (1-a)\beta(t) - 2a \sum_{j=1}^{D} U_j \exp(-U_j^2) \, \phi_j(t)$. In this section we use $D = 4$; in the online supplement also results for the choice $D = 15$ can be found. For any X_j in the testing sample, the predictive performance of our local linear estimators $m(X_j)$ and m'_{X_j} is compared with:

(L) *Functional linear regression estimator*: the standard linear regression model applied to the projections of all the involved centred functional data into the first J basis functions (Reiss & Ogden, 2007). The estimator of $m(x)$ is the intercept estimated by this model. A sensible estimator of the Riesz representation of the functional derivative can be obtained as $\Phi(t)^{\mathsf{T}} \hat{\mathrm{b}}_L$, where $\hat{\mathrm{b}}_L$ is the estimate of the non-intercept terms in the linear model with J regressors and the intercept. Expansion into the eigenbasis estimated from the random sample functions is considered.

(LC) *Functional local constant Nadaraya-Watson kernel estimator*:

$$\hat{m}(X_j) \coloneqq \sum_{i=1}^{n} Y_i \, K\left(\|X_j - X_i\| / h\right) / \sum_{i=1}^{n} K\left(\|X_j - X_i\| / h\right)$$

for K a kernel function, and h a bandwidth (Ferraty & Vieu, 2006). The local constant estimator does not allow direct estimation of the functional derivative m'_{X_j}.

(LL) *Functional local linear regression estimator*: the estimators of $m(X_j)$ and m'_{X_j} proposed in this paper with expansions into the eigenbasis given by the empirical covariance operator of the random sample curves. All parameters are automatically selected.

(FAM) *Functional additive model estimator*: the centred functional data are first projected into the univariate spaces given by their first J estimated eigenfunctions $\hat{\phi}_1, \ldots, \hat{\phi}_J$ to obtain their principal component scores. For all $j = 1, \ldots, J$, local polynomial estimates $\hat{f}_j$ and $\hat{f}_j{}'$ of the regression function and its derivative, respectively, in the model of centred responses against the univariate scores are obtained. The additive regression operator m is estimated by the sum of the functional values $\hat{f}_j$ evaluated at the principal scores of x, plus the average of the responses Y_i. The final estimator of the functional derivative m_x is the sum of functions $\hat{\phi}_j$ weighted by the corresponding estimated derivatives $\hat{f}_j{}'$ evaluated at the principal scores of x. According to Müller & Yao (2008, 2010), J is chosen so that the first J estimated eigenfunctions explain $90\ \%$ of the variability in the data.

(GPR) *Gaussian process regression:* A Bayesian method of nonparametric smoothing based on the assumption that the unknown regression functional m_a comes as a realisation of a Gaussian process with a given covariance structure that depends on a small number of hyper-parameters θ. These parameters are estimated using the maximum likelihood, and the new response at a given regressor X is then predicted as the mean of the conditional distribution of the response given X and the estimate of θ. We use the implementation described in Wang & Xu (2019) with the covariance functional $k(X_i, X_j) = v \exp(-w \|X_i - X_j\|)$, with hyper-parameters $v > 0$, $w > 0$, and $\sigma_\varepsilon^2 > 0$, the last one being the conditional variance of the error term. This method does not allow direct estimation of functional derivatives.

Note that both the kernel estimator (LC) and the linear regression estimator (L) are special cases of the local linear estimator (LL) — for $J = 0$ we recover the kernel estimator, and for a kernel $K(t)$ continuous at $t = 0$ from the right, the local linear smoother approaches the standard functional linear regression estimator as the bandwidth h tends to infinity.

For all competitors, the Epanechnikov kernel $K(t) = 0.75(1 - t^2)$ for $t \in [-1, 1]$ is used. The bandwidths, as well as the dimension J in the functional (local) linear regression, are chosen by a leave-one-out cross-validation procedure.

The learning and testing sample sizes are set to 500. Two perturbations are considered: the noise-to-signal ratio nsr of the regression model and the structural perturbation ρ acting on the regressors. Parameters $(nsr,\ \rho)$ are set to $(0.05,\ 0.05)$ and $(0.4,\ 0.4)$, respectively, corresponding to a low/high perturbation level and $a = 0,\ 0.25,\ 0.5, 0.75,\ 1$ successively. Results for other combinations of perturbation levels are provided in the online supplement. 100 runs are performed in each case. To assess the prediction quality, we use the $ORMSEP$ criteria from (2) and (3), except for $a = 0$ and the derivative, where the denominator of $ORMSEP_{deriv}$ is null since $m'_{0,x} \equiv \beta$ for any x. In the latter case, we report only the numerator of $ORMSEP_{deriv}$ from (3). Mean and standard deviation (in brackets) can be found in Tables 1 and 2, each table corresponding to a particular noise level.

From the results of the simulation study we conclude the following: (i) the functional linear estimator (L) is a good method for the estimation of the regression functional when the model is linear, or close to linear. In the situation when the model is strongly nonlinear, the estimator fails as expected. Even for models close to linear, (LL) however estimates the functional derivatives more accurately. (ii) The local linear estimator of the regression operator convincingly outperforms the local constant estimator in all considered scenarios. (iii) Method (GPR) appears to be highly competitive, especially with nonlinear models. It, however, does not allow us to estimate the functional derivatives. (iv) In most scenarios (FAM) performs worse than the local linear estimator, for both regression operator and functional derivatives. This corroborates the good finite

Table 1: Model (M2) with $nsr = 0.05$ and $\rho = 0.05$.

		$a = 0$	$a = 0.25$	$a = 0.5$	$a = 0.75$	$a = 1$
Reg.	L	0.001 (0.000)	0.014 (0.001)	0.108 (0.008)	0.531 (0.031)	1.010 (0.012)
	LC	0.048 (0.008)	0.048 (0.006)	0.068 (0.008)	0.155 (0.019)	0.265 (0.030)
	LL	0.001 (0.001)	0.007 (0.002)	0.022 (0.003)	0.059 (0.007)	0.098 (0.011)
	FAM	0.036 (0.028)	0.040 (0.025)	0.071 (0.026)	0.178 (0.035)	0.290 (0.050)
	GPR	0.008 (0.001)	0.008 (0.001)	0.011 (0.002)	0.017 (0.003)	0.024 (0.004)
Deriv.	L	**1.850 (0.028)**	1.075 (0.132)	1.012 (0.015)	1.010 (0.017)	1.004 (0.008)
	LL	**0.116 (0.118)**	0.280 (0.133)	0.140 (0.113)	0.059 (0.041)	0.063 (0.064)
	FAM	**0.477 (0.202)**	5.876 (10.083)	1.118 (2.892)	0.348 (0.275)	0.259 (0.110)

Table 2: Model (M2) with $nsr = 0.4$ and $\rho = 0.4$.

		$a = 0$	$a = 0.25$	$a = 0.5$	$a = 0.75$	$a = 1$
Reg.	L	0.007 (0.003)	0.021 (0.004)	0.119 (0.011)	0.540 (0.038)	1.009 (0.014)
	LC	0.157 (0.022)	0.162 (0.022)	0.216 (0.026)	0.410 (0.039)	0.618 (0.045)
	LL	0.017 (0.007)	0.028 (0.009)	0.089 (0.015)	0.254 (0.032)	0.386 (0.033)
	FAM	0.069 (0.044)	0.073 (0.037)	0.122 (0.040)	0.243 (0.048)	0.378 (0.053)
	GPR	0.030 (0.006)	0.034 (0.008)	0.049 (0.008)	0.092 (0.012)	0.127 (0.015)
Deriv.	L	1.850 (0.023)	1.203 (0.093)	1.029 (0.013)	1.011 (0.005)	1.003 (0.003)
	LL	0.846 (0.160)	1.058 (0.306)	0.626 (0.170)	0.474 (0.138)	0.459 (0.085)
	FAM	0.826 (0.983)	12.316 (20.955)	1.771 (1.806)	0.574 (0.643)	0.448 (0.530)

sample properties of the local linear estimator observed before, as all models considered in (M2) in the simulation study satisfy the additivity condition, under which (FAM) was designed. Note that for (LL) the additivity of the regression operator is not required. (v) Another practical issue regarding the behaviour of the estimators is their numerical stability. In the complete results of this simulation study, given in the supplementary material, we observe that (FAM) tends to be numerically unstable, especially for small learning sample sizes. The instabilities occur mostly when the principal scores of a predictor lie outside the range of the scores of the data, in which case either the functional values f_j, or the derivatives f_j' have to be extrapolated. Remarkably, (LL) does not appear to suffer from such drawbacks.

One could reckon that the introduction of the bootstrap procedure cumulates with the selection of two bandwidths and one dimension, and requires a computation that is quite intensive. Nevertheless, the running time of our `R` procedure is surprisingly short — at most about 5 seconds (according to the previously described simulation scheme with a processor Intel Core i7 2.7 GHz with 16 GB RAM) are necessary to carry out the estimation/prediction for both the regression operator and the functional derivative, including FPCA for the basis expansion $\hat{\phi}_1, \ldots, \hat{\phi}_{J_{opt}}$ with automatic computation of J_{opt}, and automatic bandwidths (h_{reg} and h_{deriv}) selection. To complete our comparative study, Table S.29 in the supplementary document indicates that the running times of our local linear estimations are competitive with respect to alternative nonparametric methods.

4·3. *Benchmark growth data analysis*

The Berkeley growth data trace back to the pioneering work of Tuddenham & Snyder (1954) and were reconsidered in Gasser et al. (1984); Kneip & Gasser (1992); Ramsay & Li (1998) and

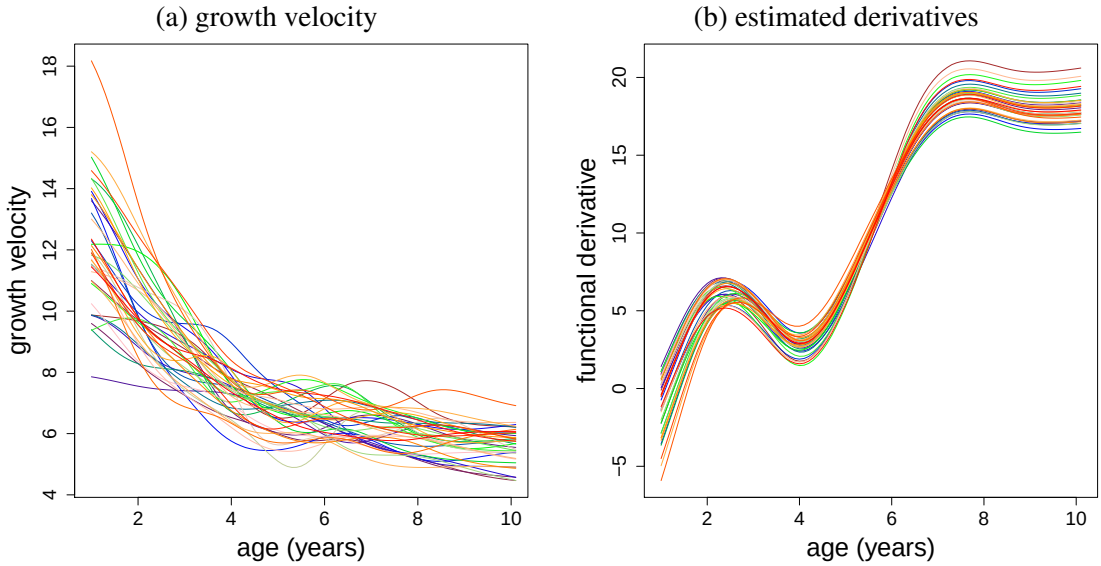

Fig. 4: Berkeley growth dataset: (a) growth velocity profiles; (b) functional derivatives estimated using the local linear approach.

Gervini & Gasser (2005). In Hall et al. (2009) an interesting analysis of this dataset involving estimated functional derivatives was performed. To better understand the growth mechanism, the relationship between the growth velocity profile up to 10 years of age (the functional predictor X) and the adult height (observed at 18 years, scalar response Y) of the boys (39 individuals) was investigated. Here, we consider the same problem using the local linear methodology. Our approach allows us to estimate the functional derivatives corresponding to the individual regressors X directly, which greatly facilitates the interpretation of the results. In Figure 4(a) we see the growth velocity profiles obtained via standard univariate local linear regression and in Figure 4(b) the estimated functional derivatives $\hat{m}'_{X_1}, \ldots, \hat{m}'_{X_{39}}$ are displayed. Focusing on the estimated functional derivatives, a sharp increase at around 5 years of age is observed for all boys. A possible interpretation is that the growth velocity profile prior to the age of 5 has little impact on the adult height of an individual. Compared to the previous analyses of the growth dataset, this finding appears to be original.

To better understand the shape of the estimated functional derivatives, we propose to model the relationship between the adult height at 18 and the growth velocity up to 10 using a single-functional index model

$$\text{height at } 18 \;=\; g\left(\langle \text{growth velocity up to } 10,\, \beta\rangle\right) \;+\; error,$$

where the link function g and the functional direction β are unknown. The single-functional index model is well suited for the studied problem, as all the estimated functional derivatives share a common shape. Therefore, the average functional derivative is a good representative of the collection of the estimated derivatives. Using the average derivative estimation method described in the introduction of this paper, we can estimate the functional parameter β and the link function g as follows: (1) $\hat{\mathrm{E}}\, m'_X := 39^{-1}\sum_{i=1}^{39} \hat{m}'_{X_i}$ and $\hat{\beta} = \hat{\mathrm{E}}\, m'_X / \|\hat{\mathrm{E}}\, m'_X\|$, (2) based on the sample $(Z_1, Y_1), \ldots, (Z_{39}, Y_{39})$ where $Z_i := \langle X_i, \hat{\beta}\rangle$, one gets an estimator $\hat{g}$ of the link function g by any standard univariate nonparametric regression method.

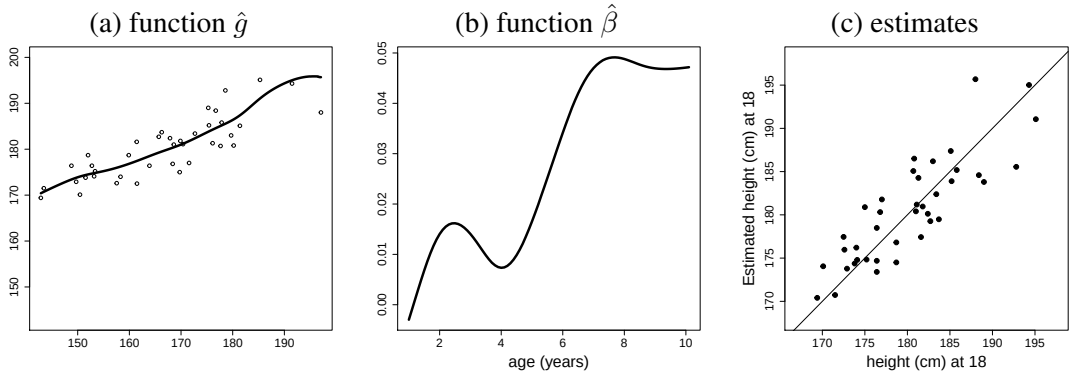

Fig. 5: Berkeley growth dataset: (a) link function $\hat{g}$ estimated by local linear regression; (b) estimated functional index $\hat{\beta}$; (c) observed adult height versus its estimates.

Figures 5(a) and (b) display respectively $\hat{g}$ and $\hat{\beta}$. The shape of the estimated functional index $\hat{\beta}$ reflects the significant jump at around 5 years with a stabilization at 8 years. The results are quite positive; the estimated heights are strongly correlated with the observed adult heights with Pearson's correlation $\simeq 0.85$. Since $\hat{g}$ is a positive and non-decreasing function, the growth velocity after 5 years of age plays a major role in the prediction of the adult height. In order to confirm this interpretation, the single-functional index model was estimated again, but by restricting the growth velocity to the range of 5–8 years of age. As expected, the estimation quality is comparable; similar correlation ($\simeq 0.82$) between the estimated adult heights and the observed ones is obtained. For additional details we refer to Section S.1·7 in the supplementary material.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the helpful suggestions of anonymous reviewers. The work of S. Nagy was supported by the grant 19-16097Y of the Czech Science Foundation, and by the PRIMUS/17/SCI/3 project of Charles University.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online extends the materials from the main paper in two directions. The first part is oriented towards practitioners; it describes additional features of our estimating procedure, the companion R package and its implementation for reproducing analyses. The second part details hypotheses and proofs of theorems with complementary theoretical developments regarding examples of approximating bases and the original condition (H3).

REFERENCES

AIT-SAÏDI, A., FERRATY, F., KASSA, R. & VIEU, P. (2008). Cross-validated estimations in the single-functional index model. *Statistics* **42**, 475–494.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*. Akadémiai Kiadó, Budapest.

ALLEN, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16**, 125–127.

AMATO, U., ANTONIADIS, A. & DE FEIS, I. (2006). Dimension reduction in functional regression with applications. *Comput. Statist. Data Anal.* **50**, 2422–2446.
BAÍLLO, A. & GRANÉ, A. (2009). Local linear regression for functional predictor and scalar response. *J. Multivariate Anal.* **100**, 102–111.
BERLINET, A., ELAMINE, A. & MAS, A. (2011). Local linear regression for functional data. *Ann. Inst. Statist. Math.* **63**, 1047–1075.
BOSQ, D. (2000). *Linear processes in function spaces*, vol. 149 of *Lecture Notes in Statistics*. Springer-Verlag, New York.
BURBA, F., FERRATY, F. & VIEU, P. (2009). k-nearest neighbour method in functional nonparametric regression. *J. Nonparametr. Stat.* **21**, 453–469.
CAI, T. T. & HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2159–2179.
CARDOT, H., FERRATY, F. & SARDA, P. (1999). Functional linear model. *Statist. Probab. Lett.* **45**, 11–22.
CHEN, D., HALL, P. & MÜLLER, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *Ann. Statist.* **39**, 1720–1747.
CHENG, M.-Y., FAN, J. & MARRON, J. S. (1997). On automatic boundary corrections. *Ann. Statist.* **25**, 1691–1708.
CRAMBES, C., KNEIP, A. & SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37**, 35–72.
DAUXOIS, J., POUSSE, A. & ROMAIN, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal.* **12**, 136–154.
DE BOOR, C. (1978). *A practical guide to splines*, vol. 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York-Berlin.
DEVROYE, L. (1978). The uniform convergence of nearest neighbour regression function estimators and their application in optimization. *IEEE Trans. Infom. Theory* **24**, 142–151.
DEVROYE, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.* **9**, 1310–1319.
FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998–1004.
FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.
FAN, J. & GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008–2036.
FAN, J. & GIJBELS, I. (1996). *Local polynomial modelling and its applications*, vol. 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.
FERRATY, F., VAN KEILEGOM, I. & VIEU, P. (2010). On the validity of the bootstrap in non-parametric functional regression. *Scand. J. Stat.* **37**, 286–306.
FERRATY, F. & VIEU, P. (2006). *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York.
GASSER, T., KÖHLER, W., MÜLLER, H.-G., KNEIP, A., LARGO, R., MOLINARI, L. & PRADER, A. (1984). Velocity and acceleration of height growth using kernel estimation. *Ann. Hum. Biol.* **11**, 397–411.
GERVINI, D. & GASSER, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika* **92**, 801–820.
GYÖRFI, L., KOHLER, M., KRZYZAK, A. & WALK, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.
HALL, P. & HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 109–126.
HALL, P. & MARRON, J. S. (1997). On the role of the shrinkage parameter in local linear smoothing. *Probab. Theory Related Fields* **108**, 495–516.
HALL, P., MÜLLER, H.-G. & WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34**, 1493–1517.
HALL, P., MÜLLER, H.-G. & YAO, F. (2009). Estimation of functional derivatives. *Ann. Statist.* **37**, 3307–3329.
HÄRDLE, W. & STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986–995.
HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2001). *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York.
HSING, T. & EUBANK, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.
HURVICH, C. M., SIMONOFF, J. S. & TSAI, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60**, 271–293.
JAMES, G. M. (2002). Generalized linear models with functional predictors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**, 411–432.
JIANG, C.-R. & WANG, J.-L. (2011). Functional single index models for longitudinal data. *Ann. Statist.* **39**, 362–388.
KARHUNEN, K. (1946). Zur Spektraltheorie stochastischer Prozesse. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.* **34**, 1–7.

KNEIP, A. & GASSER, T. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.* **20**, 1266–1305.
KOKOSZKA, P. & REIMHERR, M. (2017). *Introduction to functional data analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL.
LIN, Z., MÜLLER, H.-G. & YAO, F. (2018). Mixture inner product spaces and their application to functional data analysis. *Ann. Statist.* **46**, 370–400.
LOÈVE, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle. *Revue Sci.* **84**, 159–162.
MACK, Y.-P. (1981). Local properties of k-nn regression estimates. *SIAM J. Alg. Discr. Meth.* **2**, 311–323.
MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21**, 255–285.
MÜLLER, H.-G. & STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33**, 774–805.
MÜLLER, H.-G. & YAO, F. (2008). Functional additive models. *J. Amer. Statist. Assoc.* **103**, 1534–1544.
MÜLLER, H.-G. & YAO, F. (2010). Additive modelling of functional gradients. *Biometrika* **97**, 791–805.
R CORE TEAM (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
RAMSAY, J. O. & LI, X. (1998). Curve registration. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60**, 351–363.
RAMSAY, J. O. & SILVERMAN, B. W. (2002). *Applied functional data analysis*. Springer Series in Statistics. Springer-Verlag, New York.
RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional data analysis*. Springer Series in Statistics. Springer, New York, 2nd ed.
RAO, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics* **14**, 1–17.
REDD, A. (2012). A comment on the orthogonalization of B-spline basis functions and their derivatives. *Stat. Comput.* **22**, 251–257.
REISS, P. T. & OGDEN, R. T. (2007). Functional principal component regression and functional partial least squares. *J. Amer. Statist. Assoc.* **102**, 984–996.
RUPPERT, D. & WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.
STONE, C. J. (1977a). Consistent nonparametric regression. *Ann. Statist.* **5**, 595–620.
STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* , 1040–1053.
STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36**, 111–147.
STONE, M. (1977b). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39**, 44–47.
TUDDENHAM, R. D. & SNYDER, M. M. (1954). Physical growth of California boys and girls from birth to eighteen years. *Publications in child development. University of California, Berkeley* **1**, 183–364.
WANG, B. & XU, A. (2019). Gaussian process methods for nonparametric functional regression with mixed predictors. *Comput. Statist. Data Anal.* **131**, 80–90.
WU, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261–1350.
YAO, F., MÜLLER, H.-G. & WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**, 577–590.

[*Received* 10 *October* 2019. *Editorial decision on* 19 *March* 2021]