Biometrika (2021), **108**, 1, *pp.* 167–181
Printed in Great Britain
doi: 10.1093/biomet/asaa041
Advance Access publication 21 July 2020

Functional regression on the manifold with contamination

BY ZHENHUA LIN

Department of Statistics and Applied Probability, National University of Singapore, 117546 Singapore
stalz@nus.edu.sg

AND FANG YAO

Department of Probability and Statistics, School of Mathematical Sciences, Center for Statistical Science, Peking University, Beijing 100871, China
fyao@math.pku.edu.cn

SUMMARY

We propose a new method for functional nonparametric regression with a predictor that resides on a finite-dimensional manifold, but is observable only in an infinite-dimensional space. Contamination of the predictor due to discrete or noisy measurements is also accounted for. By using functional local linear manifold smoothing, the proposed estimator enjoys a polynomial rate of convergence that adapts to the intrinsic manifold dimension and the contamination level. This is in contrast to the logarithmic convergence rate in the literature of functional nonparametric regression. We also observe a phase transition phenomenon related to the interplay between the manifold dimension and the contamination level. We demonstrate via simulated and real data examples that the proposed method has favourable numerical performance relative to existing commonly used methods.

Some key words: Contaminated functional data; Functional nonparametric regression; Intrinsic dimension; Local linear manifold smoothing; Phase transition.

1. INTRODUCTION

Regression with a functional predictor is of central importance in the field of functional data analysis, and has been advanced by Ramsay & Silverman (1997, 2002) and many other researchers. The early development of functional regression focused on functional linear models (Cardot et al., 1999; Yao et al., 2005b; Yuan & Cai, 2010). Extensions of linear models include generalized linear regression (Cardot & Sarda, 2005; Müller & Stadtmüller, 2005), additive models (Müller & Yao, 2008) and quadratic models (Yao & Müller, 2010), among others. In these works, specific forms of the regression model are prescribed, which are regarded as functional parametric regression models (Ferraty & Vieu, 2006) that entail efficient estimation procedures, and hence are well studied in the literature.

In contrast, functional nonparametric regression, which does not impose structural constraints on the regression function, has received less attention. The first landmark in the development of nonparametric functional data analysis was the monograph of Ferraty & Vieu (2006). Recent advances in this direction include the Nadaraya–Watson estimator (Ferraty et al., 2012) and

the k-nearest-neighbour estimator (Kudraszow & Vieu, 2013). The development of functional nonparametric regression has been hindered by a theoretical barrier, which is formulated in Mas (2012) and linked to the small ball probability problem (Delaigle & Hall, 2010). Essentially, in a rather general setting, the minimax rate of nonparametric regression on a generic functional space is slower than any polynomial of the sample size, which differs markedly from the polynomial minimax rates for many functional parametric regression procedures, see, e.g., Hall & Keilegom (2007), and Yuan & Cai (2010) for functional linear regression. These endeavours in functional nonparametric regression do not exploit the intrinsic structure that is common in practice. For instance, Chen & Müller (2012) suggested that functional data often have a low-dimensional manifold structure which can be utilized for more efficient representation. In this article, we exploit the nonlinear low-dimensional structure for functional nonparametric regression.

Our method, which we call functional regression on the manifold, assumes the model

$$Y = g(X) + \varepsilon, \tag{1}$$

where Y is a scalar response, X is a functional predictor sampled from an unknown manifold $\mathcal{M}$, ε is an error term that is independent of X, and g is some unknown functional to be estimated. In reality, the functional predictor X is rarely fully observed. To accommodate this common scenario, we assume that X is recorded on a grid of points with noise. The model (1) features a manifold structure $\mathcal{M}$ that underlies the functional predictor X and is assumed to be a finite-dimensional, but potentially nonlinear submanifold of the function space $\mathcal{L}^2(D)$, the space of square-integrable functions defined on a compact domain $D \subset \mathbb{R}$. For background on both finite-dimensional and infinite-dimensional manifolds, we refer readers to Lang (1995, 1999).

Data analysis with a manifold structure has been extensively studied in the statistical literature. For example, techniques have been devised to learn an unknown manifold based on a point cloud, such as locally linear embedding (Roweis & Saul, 2000; Wu & Wu, 2018), isomap (Tenenbaum et al., 2000), diffusion maps (Coifman et al., 2005), t-SNE (van der Maaten & Hinton, 2008) and many other methods. Supervised learning on an unknown manifold has also been investigated, such as estimation of functions defined on a manifold (Aswani et al., 2011; Cheng & Wu, 2013; Sober et al., 2020) and estimation of the gradient of such functions (Mukherjee et al., 2010). In addition, data analysis on a known manifold has been studied, such as fundamentals related to the Fréchet mean (Bhattacharya & Patrangenaru, 2003, 2005; Bhattacharya & Lin, 2017), manifold-valued function estimation (Yuan et al., 2012; Lin et al., 2016; Cornea et al., 2017; Lin et al., 2019), manifold-valued principal component analysis (Huckemann et al., 2010; Panaretos et al., 2014), classification on manifolds (Yao & Zhang, 2020) and nonparametric manifold-valued inference (Patrangenaru & Ellingson, 2015).

However, research specifically relating functional data to manifolds is scarce. Zhou & Pan (2014) investigated functional principal component analysis on an irregular domain. Chen & Müller (2012) and Lila & Aston (2016) considered the representation and principal component analysis of functional data sampled from a manifold. Manifold-valued random functions were studied by Su et al. (2014), Dai & Müller (2018) and Lin & Yao (2019). To the best of our knowledge, the present paper is the first to consider a manifold structure in functional regression where a global representation of the low-dimensional functional predictor X can be inefficient. For illustration, the Supplementary Material gives an example of a random process taking values in a one-dimensional submanifold of $\mathcal{L}^2([0, 1])$ while having an infinite number of components in its Karhunen–Loève expansion.

When estimating the regression functional g in (1), we explicitly account for the hidden manifold structure by estimating the tangent spaces of the manifold. Specifically, we first recover

the observed functional predictors from their discrete or noisy measurements, and then adopt the local linear manifold smoothing technique of Cheng & Wu (2013). While our approach and that of Cheng & Wu (2013) share the same intrinsic manifold set-up, their method differs fundamentally in the ambient aspect, which raises challenging issues unique to functional data. First, functional data naturally live in an infinite-dimensional ambient space, while the Euclidean data considered by Cheng & Wu (2013) have a finite ambient dimension. Second, the effects of noise and sampling in the observed functional data need to be dealt with explicitly, since functional data are discretely and noisily recorded in practice, which leads to contamination of the functional predictor. This contamination issue is not encountered in the situation studied by Cheng & Wu (2013) and has been considered only for linear regression of multivariate data by Aswani et al. (2011) and Loh & Wainwright (2012). Moreover, the contamination has an intrinsic dimension that grows with the sample size and thus is coupled with the ambient infinite dimensionality.

The main contributions of this article are as follows. First, by exploiting structural information of the predictor, our approach produces an effective estimation procedure that adapts to the unknown manifold structure and the contamination level, while maintaining the flexibility of functional nonparametric regression. Second, by careful theoretical analysis, we confirm that the regression functional g can be estimated at a polynomial convergence rate with respect to the sample size, especially when only the contaminated functional predictors are available. This provides a new angle on functional nonparametric regression that is subject to a logarithmic rate (Mas, 2012). Third, the contamination of predictors is explicitly accounted for and is shown to be an integral part of the convergence rate, which has not been well studied even in classical functional linear regression (Hall & Keilegom, 2007). Finally, we discover that the polynomial convergence rate exhibits a phase transition phenomenon, depending on the interplay between the manifold dimension and the contamination level. This type of phase transition had not previously been observed in functional regression, and is of at least the same importance as those concerning the estimation of mean or covariance functions (e.g., Cai & Yuan, 2011; Zhang & Wang, 2016). Moreover, in the course of our theoretical development, we obtain some useful results that are of independent interest, such as consistency of the estimated intrinsic dimension and tangent spaces of the manifold in the presence of contamination.

2. Estimation of functional regression on the manifold

2.1. *Step I: recovery of functional data*

We assume that each predictor X_i is observed at m_i design points $T_{i1}, \ldots, T_{im_i} \in D$. We denote the observed value at T_{ij} by $X_{ij}^* = X_i(T_{ij}) + \zeta_{ij}$, where ζ_{ij} is random noise with mean zero and is independent of all X_i and T_{ij}. The collection $\mathbb{X}_i = \{(T_{i1}, X_{i1}^*), \ldots, (T_{im_i}, X_{im_i}^*)\}$ represents all measurements for the realization X_i, and $\{\mathbb{X}_1, \ldots, \mathbb{X}_n\}$ constitutes the observed data for the predictor. We clarify that although each trajectory X_i as a whole function resides on the manifold $\mathcal{M}$, the m_i-dimensional vector $\mathbb{V}_i = \{X_i(T_{i1}), \ldots, X_i(T_{im_i})\}$ does not. Consequently, the manifold assumption in Cheng & Wu (2013) is violated for $\mathbb{V}_i$.

When $\inf_i m_i$ is sufficiently large or grows with the sample size, a scenario commonly referred to as the dense design, we may recover each function X_i based on the observed data $\mathbb{X}_i$ by individual smoothing estimation. Popular smoothing techniques include the local linear smoother (Fan, 1993) and spline smoothing (Ramsay & Silverman, 2005). By applying one of these methods, one obtains an estimate $\hat{X}_i$ of X_i, referred to as the contaminated version of X_i, which is used in subsequent steps to estimate g. To be specific, we consider the local linear estimate of $X_i(t)$ given by $\hat{b}_1$ such that

$$(\hat{b}_1, \hat{b}_2) = \underset{(b_1,b_2)\in\mathbb{R}^2}{\arg\min} \frac{1}{m_i} \sum_{j=1}^{m_i} \{X_{ij}^* - b_1 - b_2(T_{ij} - t)\}^2 K\left(\frac{T_{ij} - t}{h_i}\right),$$

where K is a compactly supported symmetric density function and h_i is the bandwidth. It can be shown that $\hat{b}_1 = (R_0S_2 - R_1S_1)/(S_0S_2 - S_1^2)$, where

$$S_r(t) = \frac{1}{m_i h_i} \sum_{j=1}^{m_i} K\left(\frac{T_{ij} - t}{h_i}\right) \left(\frac{T_{ij} - t}{h_i}\right)^r,$$

$$R_r(t) = \frac{1}{m_i h_i} \sum_{j=1}^{m_i} K\left(\frac{T_{ij} - t}{h_i}\right) \left(\frac{T_{ij} - t}{h_i}\right)^r X_{ij}^*$$

for $r = 0, 1$ and 2.

The estimate $\hat{b}_1$ does not have a finite mean squared error, as its denominator is zero with positive probability for a finite sample. To overcome this issue, we adopt the technique of ridging (Fan, 1993; Seifert & Gasser, 1996; Hall & Marron, 1997) to estimate $X_i(t)$ by the following ridged local linear estimate:

$$\hat{X}_i(t) = \frac{R_0S_2 - R_1S_1}{S_0S_2 - S_1^2 + \delta \mathbb{1}_{\{|S_0S_2 - S_1^2| < \delta\}}}, \tag{2}$$

where $\delta > 0$ is a sufficiently small constant that depends on m_i, such as $\delta = m_i^{-2}$.

When $\sup_i m_i$ is relatively small or bounded by a constant, a scenario commonly referred to as the sparse design, the procedure proposed by Yao et al. (2005a) can be used to recover individual X_i. We refer readers to the Supplementary Material for details of such a procedure.

2.2. *Step II: estimation of the manifold dimension and tangent space*

To characterize the manifold structure, we first estimate the intrinsic dimension d of the manifold $\mathcal{M}$. We adopt the maximum likelihood estimator proposed by Levina & Bickel (2004), replacing the unobservable X_i with the contaminated version $\hat{X}_i$. For a given $x \in \mathcal{M}$, define $\hat{G}_i(x) = \|x - \hat{X}_i\|_{\mathcal{L}^2}$ and let $\hat{G}_{(k)}(x)$ be the kth order statistic of $\hat{G}_1(x), \ldots, \hat{G}_n(x)$. Then the intrinsic dimension d is estimated by

$$\hat{d} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{d}_k \tag{3}$$

with

$$\hat{d}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{d}_k(\hat{X}_i), \quad \hat{d}_k(x) = \left\{ \frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{\hat{G}_{(k)}(x) + \Delta}{\hat{G}_{(j)}(x) + \Delta} \right\}^{-1}, \tag{4}$$

where Δ is a positive constant depending on n, and k_1 and k_2 are tuning parameters. The constant Δ regularizes $\hat{d}_k(x)$ to overcome the additional variability introduced by the contamination of the predictor. We conveniently set $\Delta = 1/\log \bar{m}$ with $\bar{m} = n^{-1} \sum_{i=1}^{n} m_i$ and refer readers

to Levina & Bickel (2004) for the choices of k_1 and k_2. When the observed data are sparsely sampled, the distance $\hat{G}_i(x)$ can be better estimated by the procedure of Peng & Müller (2008).

Now we proceed to estimate the tangent space at the given point x as follows.

Step 1. Determine a neighbourhood of x, $\hat{\mathcal{N}}_{\mathcal{L}^2}(h_{\text{pca}}, x) = \{\hat{X}_i : \|x - \hat{X}_i\|_{\mathcal{L}^2} < h_{\text{pca}}, i = 1, \ldots, n\}$, where $h_{\text{pca}} > 0$ is a tuning parameter.

Step 2. Compute the local empirical covariance function

$$\hat{\mathcal{C}}_x(s,t) = \frac{1}{|\hat{\mathcal{N}}_{\mathcal{L}^2}(h_{\text{pca}}, x)|} \sum_{\hat{X} \in \hat{\mathcal{N}}_{\mathcal{L}^2}(h_{\text{pca}}, x)} \{\hat{X}(s) - \hat{\mu}_x(s)\}\{\hat{X}(t) - \hat{\mu}_x(t)\} \tag{5}$$

and obtain the eigenfunctions $\hat{\varphi}_1, \hat{\varphi}_2, \ldots, \hat{\varphi}_{\hat{d}}$ corresponding to the $\hat{d}$ leading eigenvalues, where $\hat{\mu}_x = |\hat{\mathcal{N}}_{\mathcal{L}^2}(h_{\text{pca}}, x)|^{-1} \sum_{\hat{X} \in \hat{\mathcal{N}}_{\mathcal{L}^2}(h_{\text{pca}}, x)} \hat{X}$ is the local mean function and $|\hat{\mathcal{N}}_{\mathcal{L}^2}(h_{\text{pca}}, x)|$ denotes the number of observations in $\hat{\mathcal{N}}_{\mathcal{L}^2}(h_{\text{pca}}, x)$.

Step 3. Estimate the tangent space at x by $\hat{T}_x\mathcal{M} = \text{span}\{\hat{\varphi}_1, \hat{\varphi}_2, \ldots, \hat{\varphi}_{\hat{d}}\}$, the linear space spanned by the first $\hat{d}$ estimated eigenfunctions.

2.3. *Step III: local linear regression on the tangent space*

Finally, we utilize the local manifold structure by projecting all the $\hat{X}_i$ onto the estimated tangent space $\hat{T}_x\mathcal{M}$, obtaining the local coordinate $\hat{\xi}_i = (\langle \hat{X}_i, \hat{\varphi}_1 \rangle, \ldots, \langle \hat{X}_i, \hat{\varphi}_{\hat{d}} \rangle)^{\mathrm{T}}$ for $\hat{X}_i$. Then, the estimate of $g(x)$ is given by

$$\hat{g}(x) = e_1^{\mathrm{T}} (\hat{Q}^{\mathrm{T}} \hat{W} \hat{Q})^{-1} \hat{Q}^{\mathrm{T}} \hat{W} \mathcal{Y}, \quad \hat{Q} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \hat{\xi}_1 & \hat{\xi}_2 & \cdots & \hat{\xi}_n \end{pmatrix}^{\mathrm{T}}, \tag{6}$$

where $\hat{W} = \text{diag}\{K_{h_{\text{reg}}}(\|x - \hat{X}_1\|_{\mathcal{L}^2}), K_{h_{\text{reg}}}(\|x - \hat{X}_2\|_{\mathcal{L}^2}), \ldots, K_{h_{\text{reg}}}(\|x - \hat{X}_n\|_{\mathcal{L}^2})\}$ with $K_h(t) = K(t/h)/h^{\hat{d}}$ and bandwidth h_{reg}, $\mathcal{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, and $e_1^{\mathrm{T}} = (1, 0, \ldots, 0)$ is an $n \times 1$ vector. Here, the matrix $\hat{Q}$ incorporates the estimated geometric structure that is encoded by the local eigenbasis $\hat{\varphi}_1, \ldots, \hat{\varphi}_{\hat{d}}$. We emphasize that in the above estimation procedure, which is illustrated in Fig. 1(a), all the steps are based on the contaminated sample $\{\hat{X}_1, \ldots, \hat{X}_n\}$ rather than the unavailable functions $\{X_1, \ldots, X_n\}$. When the predictor x is also measured only at m_x discrete points $t_1, \ldots, t_{m_x}$, we impute it by the procedures in § 2.1 and replace x in (4)–(6) with the imputed curve $\tilde{x}$ to obtain an estimate of $g(\tilde{x})$.

2.4. *Tuning parameter selection*

There are several tuning parameters to be determined in our estimation procedure. For the parameters k_1 and k_2 in (3) for estimating the intrinsic dimension, $k_1 = 10$ and $k_2 = 20$ are suggested by Levina & Bickel (2004). However, we have found that $k_1 = 20$ and $k_2 = 30$ generally work better in our setting, perhaps partially because of the contamination, which requires a relatively large local neighbourhood to offset it.

For the individual smoothing presented in § 2.1, we employ the following leave-one-out cross-validation to select the bandwidth h_i (Fan & Gijbels, 1996; Lee& Solo, 1999). Let $\hat{X}_{i,h,-j}(x)$ be the leave-one-out estimate of $X_i(t)$, i.e., the estimate computed according to (2) using all

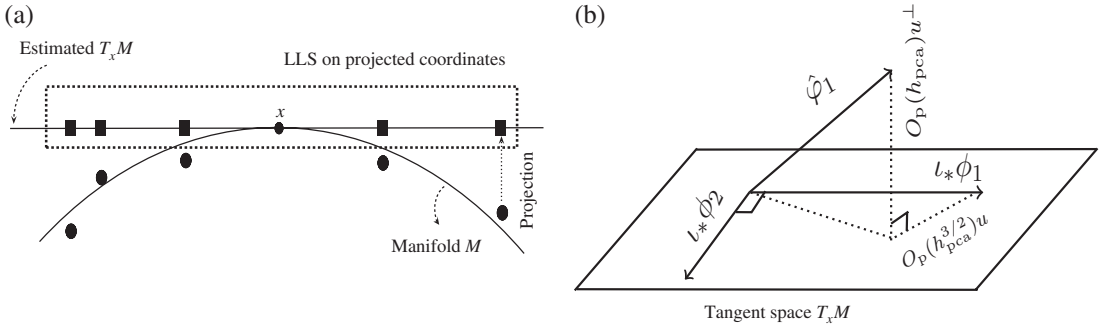

Fig. 1. (a) An illustration of functional regression on a manifold, where dots represent the observed $\hat{X}_i$ and squares the projected $\hat{x}_i$. (b) An illustration of the asymptotic result of (8) for $d = 2$.

of $(T_{i1}, X^*_{i1}), \ldots, (T_{im_i}, X^*_{im_i})$ except (T_{ij}, X^*_{ij}). We then select h_i from a pool of candidates to minimize the cross-validation error $\mathrm{CV}(h) = \sum_{j=1}^{m_i} \{X^*_{ij} - \hat{X}_{i,h,-j}(T_{ij})\}^2$.

For the bandwidths h_{pca} in (5) and h_{reg} in (6), we choose the pair $(h_{\mathrm{pca}}, h_{\mathrm{reg}})$ from a pool $\mathbb{H}$ of candidate pairs to minimize the leave-one-out cross-validation error $\mathrm{CV}(h_{\mathrm{pca}}, h_{\mathrm{reg}}) = \sum_{i=1}^{n} \{Y_i - \hat{g}_{h_{\mathrm{pca}}, h_{\mathrm{reg}}, -i}(\hat{X}_i)\}^2$, where $\hat{g}_{h_{\mathrm{pca}}, h_{\mathrm{reg}}, -i}$ denotes the leave-one-out estimate of g with parameters $(h_{\mathrm{pca}}, h_{\mathrm{reg}})$ without using the pair $(\hat{X}_i, Y_i)$. The pool $\mathbb{H}$ will be constructed in such a way that every $\hat{\mathcal{N}}_{\mathcal{L}^2}(h_{\mathrm{pca}}, \hat{X}_i)$ contains at least $\hat{d} + 1$ samples for every pair $(h_{\mathrm{pca}}, h_{\mathrm{reg}})$ in $\mathbb{H}$, in order to ensure sufficient data for local estimation.

3. THEORETICAL PROPERTIES

We focus on the scenario where $\inf_i m_i$ increases with the sample size n, and defer the case of $\sup_i m_i \leqslant m_0 < \infty$ to future research because of the additional technical challenges associated with it. Without loss of generality, assume $m_i \asymp m$, where $a_n \asymp b_n$ means $0 < \liminf a_n/b_n < \limsup a_n/b_n < \infty$. We further assume that the ζ_{ij}, and similarly T_{ij} and X_i, are independent and identically distributed. We emphasize that the theoretical development below can be modified to accommodate fixed designs, weak dependence and heterogeneous distributions. However, because achieving such generality would involve considerably more technicalities without adding further insight, it is not pursued here.

The discrepancy between $\hat{X}_i$ and X_i, quantified by $\|\hat{X}_i - X_i\|_{\mathcal{L}^2}$, is called the contamination of X_i. The decay of this contamination is intimately linked to the consistency of our estimates of the intrinsic dimension, the tangent space, and eventually the regression functional $g(x)$. Moreover, the convergence rate of $\hat{g}(x)$ is found to exhibit a phase transition phenomenon, depending on the interplay between the intrinsic dimension and the decay of the contamination. To set the stage, we start with a property of contamination in recovery of functional data by the individual smoothing approach in § 2.1. Specifically, we study the pth moment of contamination when $\hat{X}_i$ is the ridged local linear estimate in (2). Our result below for an arbitrary pth moment has not appeared before in the literature; see Fan (1993) for the $p = 2$ case only.

Let $\Sigma(\nu, L)$ denote the Hölder class with exponent ν and Hölder constant L, which represents the set of $\lfloor \nu \rfloor$-times differentiable functions F whose derivative $F^{(\ell)}$ for $\ell = \lfloor \nu \rfloor$ satisfies $|F^{(\ell)}(t) - F^{(\ell)}(s)| \leqslant L|t-s|^{\nu-\ell}$ for $s, t \in D$, where $\lfloor \nu \rfloor$ denotes the largest integer strictly smaller than ν. We require the following mild assumptions, and we assume $h_i \asymp h_0$ without loss of generality.

Assumption 1. The kernel K is differentiable with a bounded derivative, and is such that $\int_{-1}^{1} K(u)\,\mathrm{d}u = 1$, $\int_{-1}^{1} uK(u)\,\mathrm{d}u = 0$ and $\int_{-1}^{1} |u|^p K(u)\,\mathrm{d}u < \infty$ for all $p > 0$.

Assumption 2. The sampling density f_T is bounded away from zero and infinity, i.e., for some constants $C_{T,1}, C_{T,2} \in (0, \infty)$, $C_{T,1} = \inf_{t\in D} f_T(t) \leqslant \sup_{t\in D} f_T(t) = C_{T,2}$.

Assumption 3. We have that $X \in \Sigma(\nu, L_X)$, where $L_X > 0$ is a random quantity and the constant $\nu \in (0, 2]$ quantifies the smoothness of the process.

Assumption 4. For all $r \geqslant 1$, $E(\sup_t |X(t)|^r) < \infty$, $E(L_X)^r < \infty$ and $E(|\zeta|^r) < \infty$.

The condition $E(\sup_t |X(t)|^r) < \infty$ holds rather generally (Li & Hsing, 2010; Zhang & Wang, 2016), compared with a stronger assumption on X given in Hall et al. (2006, (A.1)). The following proposition is an immediate consequence of Lemma S1 in the Supplementary Material, so its proof is omitted.

PROPOSITION 1. *For any $p \geqslant 1$, assume $E(|\zeta|^p) < \infty$. Under Assumptions 1–3, for the estimate $\hat{X}$ in (2) with $h_0 \asymp m^{-1/(2\nu+1)}$ and $\delta = m^{-2}$, we have*

$$\{E(\|\hat{X} - X\|_{\mathcal{L}^2}^p \mid X)\}^{1/p} = O\{m^{-\nu/(2\nu+1)}\}\Big\{\sup_t |X(t)| + L_X\Big\}. \tag{7}$$

Furthermore, if Assumption 4 also holds, then $\{E(\|\hat{X} - X\|_{\mathcal{L}^2}^p)\}^{1/p} = O\{m^{-\nu/(2\nu+1)}\}$.

When X is deterministic as in nonparametric regression, the rate in (7) for $p = 2$ coincides with that in Tsybakov (2008). In addition, the pth order of the contamination $\|\hat{X}_i - X_i\|_{\mathcal{L}^2}$ decays at a polynomial rate that depends on ν, but not on the order p.

To analyse the asymptotic properties of $\hat{g}(x)$, we make the following assumptions.

Assumption 5. The probability density f of X on $\mathcal{M}$ satisfies $C_{f,1} = \inf_{x\in\mathcal{M}} f(x) \leqslant \sup_{x\in\mathcal{M}} f(x) = C_{f,2}$ for some constants $0 < C_{f,1} \leqslant C_{f,2} < \infty$.

Assumption 6. The regression functional g has a bounded second derivative.

For Assumption 5, since the functional predictor resides on a low-dimensional manifold, the existence of a density can be safely assumed. We also make the following assumption on the imputed trajectories in § 2.1.

Assumption 7. The $\hat{X}_1, \ldots, \hat{X}_n$ are independent and identically distributed. For some $\beta \in (0, \infty)$ and all $p \geqslant 1$, $\{E(\|\hat{X} - X\|_{\mathcal{L}^2}^p \mid X)\}^{1/p} \leqslant C_p m^{-\beta} \eta(X)$ for some constant C_p depending only on p and some nonnegative function $\eta(X)$ depending only on X such that $E[\{\eta(X)\}^p] < \infty$.

Under Assumptions 1–4, by Proposition 1, the imputed functions $\hat{X}_1, \ldots, \hat{X}_n$ obtained by individual smoothing via the local linear estimation (2) satisfy Assumption 7 with $\beta = \nu/(2\nu + 1)$. Therefore, Assumption 7 could be replaced with the more concrete Assumptions 1–4. The conditions can be relaxed to accommodate heterogeneous data distributions and weakly dependent functional data by modifying the proofs. Also, it is possible to accommodate imputed functions that are obtained by borrowing information across individuals (e.g., Yao et al., 2005a); this is beyond the scope of the present paper, but is an interesting topic for future research.

The contamination of the predictor X renders the true neighbourhood $\mathcal{N}_{\mathcal{L}^2}(h_{\mathrm{pca}}, x) = \{X_i : \|X_i - x\|_{\mathcal{L}^2} < h_{\mathrm{pca}}\}$ inaccessible. However, it can be shown that the contaminated neighbourhood

$\hat{\mathcal{N}}_{\mathcal{L}^2}(h_{\mathrm{pca}}, x)$ is a good estimate; see the Supplementary Material for details. Consequently, the local manifold structure can be consistently estimated in the sense of the following theorem.

THEOREM 1. *Suppose that Assumptions 5 and 7 hold.*

(i) *Then* $\hat{d}$ *is a consistent estimator of* d *when* $\min\{k_1, k_2\} \to \infty$ *and* $\max\{k_1, k_2\}/m \to 0$.
(ii) *If* $h_{\mathrm{pca}} \to 0$ *and* $h_{\mathrm{pca}} \gtrsim \max\{m^{-\beta+\epsilon}, n^{-1/(d+2)}\}$ *for an arbitrarily small but fixed constant* $\epsilon > 0$, *then the eigenbasis* $\{\hat{\varphi}_k\}_{k=1}^d$ *derived from* $\hat{\mathcal{C}}_x$ *in* (5) *is close to an orthonormal basis* $\{\phi_k\}_{k=1}^d$ *of* $T_x\mathcal{M}$ *in the sense that for each* $x \in \mathcal{M}$,

$$\hat{\varphi}_k = \phi_k + O_{\mathrm{p}}(h_{\mathrm{pca}}^{3/2})u_k + O_{\mathrm{p}}(h_{\mathrm{pca}})u_k^{\perp} \quad (k = 1, \ldots, d), \tag{8}$$

where $u_k \in T_x\mathcal{M}$, $u_k^{\perp} \perp T_x\mathcal{M}$ *and* $\|u_k\|_{\mathcal{L}^2} = \|u_k^{\perp}\|_{\mathcal{L}^2} = 1$.

In light of Theorem 1(i), we shall present subsequent results by conditioning on the event $\hat{d} = d$. For part (ii), which is illustrated in Fig. 1(b), the condition $h_{\mathrm{pca}} \gtrsim m^{-\beta+\epsilon}$ suggests that h_{pca} will be larger than the contamination by an arbitrarily small polynomial order of m. This is required to ensure that the discrepancy between the estimated local neighbourhood $\hat{N}_{\mathcal{L}^2}(h_{\mathrm{pca}}, x)$ and the uncontaminated neighbourhood $\mathcal{N}_{\mathcal{L}^2}(h_{\mathrm{pca}}, x) = \{X_i : \|x - X_i\|_{\mathcal{L}^2} < h_{\mathrm{pca}},\ i = 1, \ldots, n\}$ is asymptotically negligible, suggested by a lemma in the Supplementary Material. The curvature at x is a constant that is absorbed into the O_{p} terms and so does not influence the asymptotic rate. However, in practice it is often more difficult to estimate the tangent structure at a point with larger curvature.

We are now ready to state the results on the estimated regression functional. Recall that $\hat{g}(x)$ in (6) is obtained by applying the local linear smoother to the coordinates of contaminated predictors within the estimated tangent space at x. It is well known that the local linear estimator does not suffer from boundary effects, i.e., the first-order behaviour of the estimator on the boundary is the same as in the interior (Fan, 1992). However, the contamination of the predictor has a different impact, and we shall address the interior and boundary cases separately. Let $\mathscr{X} = \{(X_1, \hat{X}_1), \ldots, (X_n, \hat{X}_n)\}$ and $\mathcal{M}_h = \{x \in \mathcal{M} : \inf_{y \in \partial\mathcal{M}} \mathfrak{d}(x, y) \leqslant h\}$, where $\partial\mathcal{M}$ denotes the boundary of $\mathcal{M}$ and $\mathfrak{d}(\cdot, \cdot)$ is the distance function on $\mathcal{M}$. For points sufficiently far away from the boundary of $\mathcal{M}$, we have the following result about the convergence rate of the estimator $\hat{g}(x)$.

THEOREM 2. *Suppose that Assumptions 1 and 5–7 hold. Let* $x \in \mathcal{M} \backslash \mathcal{M}_{h_{\mathrm{reg}}}$ *and* h_{pca} *satisfy the conditions of Theorem 1*(ii). *For an arbitrarily small but fixed constant* $\epsilon > 0$, *suppose that* $h_{\mathrm{reg}} \to 0$, $h_{\mathrm{reg}} > h_{\mathrm{pca}}$ *and* $\min\{nh_{\mathrm{reg}}, m^{\beta} h_{\mathrm{reg}}^{5/3+\epsilon}\} \to \infty$. *Then*

$$E\big[\{\hat{g}(x) - g(x)\}^2 \mid \mathscr{X}\big] = O_{\mathrm{p}}\left(h^4 + \frac{1}{m^{2\beta} h_{\mathrm{reg}}^{2+2\epsilon}} + \frac{1}{nh^d}\right). \tag{9}$$

In addition, if $h_{\mathrm{pca}} \asymp \max\{m^{-\beta}, n^{-1/(d+2)}\}$, *and if* $h_{\mathrm{reg}} \asymp n^{-1/(d+4)}$ *when* $m \gtrsim n^{(3+\epsilon)/\{\beta(d+4)\}}$ *and* $h_{\mathrm{reg}} \asymp m^{-\beta/(3+\epsilon)}$ *otherwise, then*

$$E\big[\{\hat{g}(x) - g(x)\}^2 \mid \mathscr{X}\big] = O_{\mathrm{p}}\left\{n^{-4/(d+4)} + m^{-4\beta/(3+\epsilon)}\right\}. \tag{10}$$

We highlight the following observations from this theorem. First, according to our analysis in the Supplementary Material, the first two terms on the right-hand side of (9) correspond to

the bias while the last term stems from the variability of the estimator. This suggests that, under the conditions of the theorem, the contamination has an impact on the asymptotic bias, but not on the variance. Second, the convergence rate of $\hat{g}(x)$ is a polynomial of the sample size n and the sampling rate m. This is in contrast with traditional functional nonparametric regression methods which do not exploit the intrinsic structure and thus cannot attain a polynomial rate of convergence.

Third, the rate in (10) consists of two terms, one related to the intrinsic dimension d and the sample size n, and the other related to m and β which together characterize the contamination of the predictor. As $\epsilon > 0$ is arbitrary, the transition between these two terms occurs at the rate $\mathfrak{m}_0 \asymp n^{3/\{\beta(d+4)\}}$. When the sampling rate falls below $\mathfrak{m}_0$, the contamination term dominates the convergence rate in (10); otherwise, the intrinsic dimension and sample size determine the rate. This phase transition, though similar in spirit to that of Cai & Yuan (2011) and Zhang & Wang (2016), has a different interpretation, as follows. When the contamination level is low, the manifold structure can be estimated reliably and utilized for regression. In contrast, when the contamination is at a high level, for example when m or β is small, the manifold structure is buried by noise and cannot be well exploited. Finally, we observe that the phase transition threshold $\mathfrak{m}_0$ increases with the intrinsic dimension d that indicates the complexity of the manifold. This interesting finding suggests that, although a complex manifold makes the estimation more challenging, for example leading to a slower rate, such a manifold is more resistant to contamination.

In our set-up, the actual observed predictor is $\mathbb{X}_i = \{(T_{i1}, X^*_{i1}), \ldots, (T_{im}, X^*_{im_i})\}$, an m_i-dimensional random vector. Moreover, the distribution of this random vector is fully supported on $\mathbb{R}^{m_i}$ owing to the presence of the noise ζ_{ij}, and hence the support of the distribution of the recovered trajectory $\hat{X}_i$ is also m_i-dimensional. Smoothness of the functional data could help to tighten the distribution of $\hat{X}_i$, but will not reduce its dimension. As m_i goes to infinity, the curse of dimensionality could become a serious concern. In this sense, the polynomial rate and phase transition phenomenon in Theorem 2 are remarkable: when $\inf_i m_i$ exceeds a certain threshold, by exploiting the low-dimensional manifold structure, the growing dimension of the contamination can be defeated with the aid of smoothness.

The following theorem characterizes the behaviour of $\hat{g}$ on the boundary of $\mathcal{M}$.

THEOREM 3. *Suppose that Assumptions 1 and 5–7 hold. Let $x \in \mathcal{M}_{h_{\mathrm{reg}}}$ and h_{pca} satisfy the conditions of Theorem 1*(ii)*. For an arbitrarily small but fixed constant $\epsilon > 0$, suppose that $h_{\mathrm{reg}} \to 0$, $h_{\mathrm{reg}} > h_{\mathrm{pca}}$ and $\min\{nh_{\mathrm{reg}}, m^{\beta} h_{\mathrm{reg}}^{3+\epsilon}\} \to \infty$. Then*

$$E\big[\{\hat{g}(x) - g(x)\}^2 \mid \mathcal{X}\big] = O_{\mathrm{p}}\left(h^4 + \frac{1}{m^{2\beta} h_{\mathrm{reg}}^{4+2\epsilon}} + \frac{1}{nh^d}\right).$$

In addition, if $h_{\mathrm{pca}} \asymp \max\{m^{-\beta}, n^{-1/(d+2)}\}$, and if $h_{\mathrm{reg}} \asymp n^{-1/(d+4)}$ when $m \gtrsim n^{(4+\epsilon)/\{\beta(d+4)\}}$ and $h_{\mathrm{reg}} \asymp m^{-\beta/(4+\epsilon)}$ otherwise, then

$$E\big[\{\hat{g}(x) - g(x)\}^2 \mid \mathcal{X}\big] = O_{\mathrm{p}}\left\{n^{-4/(d+4)} + m^{-4\beta/(4+\epsilon)}\right\}. \tag{11}$$

Comparing the above theorem with Theorem 2, we see that the effect of the intrinsic dimension on convergence is the same regardless of where $\hat{g}$ is evaluated on the manifold. However, the effect of contamination behaves differently, due to the fact that the second-order behaviour of the local linear estimator depends on the location and needs to be considered when there is contamination of X. Moreover, we see that the phase transition occurs at $\mathfrak{m}_1 = n^{4/\{\beta(d+4)\}} \gg \mathfrak{m}_0$, and when the

contamination dominates, the convergence is slightly slower for boundary points than for interior points. This is the price we pay for the boundary effect when the predictor is contaminated, which is in contrast with the classical result on the local linear estimator (Fan, 1993).

4. SIMULATION STUDY

To demonstrate the performance of our method, we conduct simulation studies for three different manifolds, namely, the three-dimensional rotation group SO(3), the Klein bottle and the mixture of two Gaussian densities.

For the SO(3) manifold we set $X_i(t) = \sum_{k=1}^{9} z_{ik} b_k(t)$, where $b_{2\ell-1}(t) = \cos\{(2\ell - 1)\pi t/10\}/5^{1/2}$ and $b_{2\ell}(t) = \sin\{(2\ell - 1)\pi t/10\}/5^{1/2}$. To generate the random variables z_{ik}, for a vector $r = (r_1, r_2, r_3)$ and a variable $\theta \in \mathbb{R}$ we define

$$R(r,\theta) = (1 - \cos\theta) rr^{\mathrm{T}} + \begin{pmatrix} \cos\theta & -r_3 \sin\theta & r_2 \sin\theta \\ r_3 \sin\theta & \cos\theta & -r_1 \sin\theta \\ -r_2 \sin\theta & r_1 \sin\theta & \cos\theta \end{pmatrix}.$$

Writing $e_2 = (0,1,0)^{\mathrm{T}}$ and $e_3 = (0,0,1)^{\mathrm{T}}$, we set $(z_{i1}, \ldots, z_{i9})^{\mathrm{T}} = \mathrm{vec}(Z_i)$ with Euler angle parameterization $Z_i = R(e_3, u_i)R(e_2, v_i)R(e_3, w_i)$, where the (u_i, v_i) are uniformly sampled from the two-dimensional sphere $S^2 = [0, 2\pi) \times [0, \pi]$ and the w_i are uniformly sampled from the unit circle $S^1 = [0, 2\pi)$.

For the Klein bottle we set $X_i(t) = \sum_{k=1}^{4} z_{ik} b_k(t)$ with $b_k(t)$ as in the SO(3) setting. We set $z_{i1} = (2\cos v_i + 1)\cos u_i$, $z_{i2} = (2\cos v_i + 1)\sin u_i$, $z_{i3} = 2\sin v_i \cos(u_i/2)$ and $z_{i4} = 2\sin v_i \sin(u_i/2)$, where u_i and v_i are independently sampled from the uniform distribution on $(0, 2\pi)$. Here $(u, v) \mapsto (z_1, z_2, z_3, z_4)$ is a parameterization of the Klein bottle with intrinsic dimension $d = 2$.

For the Gaussian mixture we set $X_i(t) = \exp\{-(t - u_i)^2/2\}/(2\pi)^{1/2} + \exp\{-(t - v_i)^2/2\}/(2\pi)^{1/2}$ with $(v_1, v_2)^{\mathrm{T}}$ uniformly sampled from a circle with diameter 0.5, similar to the form used in Chen & Müller (2012).

The functional predictor X_i is observed at m_i points $T_{i1}, \ldots, T_{im_i}$ in the interval $[0,1]$ with heteroscedastic measurement errors $\zeta_{ij} \sim N(0, \sigma_{ij}^2)$, where σ_{ij} is determined by the signal-to-noise ratio $\mathrm{SNR}_X = \mathrm{var}\{X(T_{ij}) \mid T_{ij}\}/\sigma_{ij}^2 = 4$. The response is generated by $Y_i = 4\sin(4Z_i)\cos(Z_i^2) + 2\Gamma(1 + Z_i/2) + \varepsilon_i$ with $Z_i = \int_0^1 X_i^2(t) t\,\mathrm{d}t$ and $\Gamma(\alpha) = \int_0^\infty s^{\alpha-1}\exp(-s)\,\mathrm{d}s$. The noise ε_i added to the response Y is a centred Gaussian variable with variance σ_ε^2 that is determined by the signal-to-noise ratio $\mathrm{SNR}_Y = \mathrm{var}(Y)/\sigma_\varepsilon^2 = 2$. To see the effect of the manifold structure on regression, we normalize the functional predictor in all settings to the unit scale, i.e., we multiply X by the constant $c = 1/\{E(\|X\|^2)\}^{1/2}$ so that the result satisfies $E(\|X\|^2) = 1$. Such a scaling does not change the geometric structure of the manifolds except for their size. We find empirically that to account for at least 95% of the variance of the data, more than 10 principal components are needed in all settings, i.e., the dimensions of the contaminated data are considerably larger than their intrinsic dimensions.

For evaluation, we generate independent test data of size 5000 and compute the root mean squared error using the test data. In the test data, each predictor is also discretely measured and contaminated by noise in the same way as in the training sample. We compare our method with nonparametric estimators based on functional Nadaraya–Watson smoothing, functional conditional expectation, the functional mode, the functional conditional median, and a multi-method that averages estimates from the methods of functional conditional expectation, the functional

Table 1. *Results of simulation studies for densely observed data: reported are the Monte Carlo averages of root mean squared errors based on* 100 *independent simulation replicates, with the corresponding standard errors in parentheses*

	SO(3) manifold			Klein bottle			Gaussian mixture		
	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$
FLR	22.1 (0.34)	21.8 (0.23)	21.6 (0.20)	61.3 (0.62)	61.2 (0.39)	6.09 (0.35)	29.6 (1.43)	29.0 (1.26)	28.8 (0.99)
FNW	16.2 (0.58)	15.7 (0.43)	15.5 (0.32)	31.8 (4.05)	29.1 (1.79)	28.3 (0.68)	18.7 (1.46)	17.5 (0.83)	17.0 (0.65)
FCE	15.3 (0.66)	14.1 (0.52)	13.2 (0.30)	29.7 (1.46)	27.1 (1.04)	26.1 (0.81)	21.1 (1.32)	20.4 (0.93)	19.8 (0.64)
FMO	25.4 (1.16)	23.0 (0.94)	22.0 (0.85)	46.2 (3.07)	41.2 (2.20)	38.3 (1.87)	35.9 (2.80)	33.6 (2.05)	32.2 (1.61)
FCM	20.2 (0.60)	18.6 (0.52)	17.2 (0.35)	39.1 (2.67)	33.9 (1.61)	30.9 (1.02)	27.3 (1.71)	25.1 (1.05)	23.2 (0.83)
MUL	18.2 (0.59)	16.6 (0.48)	15.4 (0.31)	34.0 (2.13)	30.0 (1.24)	27.7 (0.92)	24.6 (1.49)	23.1 (1.07)	21.8 (0.81)
FREM	10.1 (0.72)	8.16 (0.56)	6.38 (0.25)	16.5 (1.39)	12.3 (1.11)	9.51 (0.74)	10.5 (1.32)	8.08 (0.86)	6.12 (0.75)

FLR, functional linear regression; FNW, functional Nadaraya–Watson smoothing; FCE, functional conditional expectation; FMO, functional mode, FCM, functional conditional median; MUL, multi-method; FREM, the proposed functional regression on the manifold.

mode and the functional conditional median (Ferraty & Vieu, 2006). Functional linear regression is also included to illustrate the impact of a nonlinear relationship. The tuning parameters in these methods, such as the number of principal components for functional linear regression and the bandwidth for the nonparametric methods, are selected by 10-fold cross-validation.

Here we focus on the scenario of dense functional data, and refer readers to the Supplementary Material for simulation studies with sparsely observed data. Specifically, we set $m_i = m = 100$ and $T_{ij} = t_j$, where $t_1, \ldots, t_m$ are equally spaced over $[0, 1]$. Three sample sizes are considered, $n = 250, 500$ and 1000. We repeat each study 100 times independently, and the results are presented in Table 1. First, we observe that the proposed method shows favourable numerical performance in all simulation settings. Second, as the sample size grows, the reduction in root mean squared error is more prominent for the proposed method than for the other methods. For example, for our method the relative reduction in root mean squared error from $n = 250$ to $n = 500$ is 25.5% and the reduction from $n = 500$ to $n = 1000$ is 22.7%, whereas for the functional Nadaraya–Watson estimator the corresponding reductions are 8.49% and 2.75%. This suggests that the proposed estimator may have a faster convergence rate. Furthermore, it provides evidence for the polynomial rate stated in Theorems 2 and 3. Based on these theorems, the relative reduction is expected to be $1 - (n_1/n_2)^{2/(d+4)}$ as the sample size increases from n_1 to n_2, since the data are sufficiently dense that the convergence rate is dominated by the intrinsic dimension. In the setting of the Klein bottle, it is about 20.6%, and the empirical relative reduction is 22.7% from $n_1 = 500$ to $n_2 = 1000$. Similar observations can be made for the other settings. In contrast, the existing kernel methods manage no better than a logarithmic rate, providing numerical evidence for the theory of Mas (2012). Third, as the intrinsic dimension goes up, the relative reduction in root mean squared error for our estimator decreases, suggesting that the intrinsic dimension plays an important role in determining the convergence rate. Finally, different manifolds result in different constants hidden in the O_p terms in Theorems 2 and 3. For example, according to Table 1, those in the SO(3) setting seem smaller than their counterparts in the Klein bottle setting.

5. Real data examples

We apply our method to the analysis of three real datasets. For the purpose of evaluation, we train our method on 75% of each dataset and reserve the remaining 25% as test data. The root mean squared error is computed on the held-out test data. We repeat this procedure 100 times on random partitions of the datasets; the results are summarized in Table 2.

Table 2. *Results for the real data analysis: reported are the Monte Carlo averages of root mean squared errors based on* 100 *independent simulation replicates, with the corresponding standard errors in parentheses; the values for the diffusion tensor imaging and systolic blood pressure data are scaled by* 0.1 *for visualization*

	FLR	FNW	FCE	FMO	FCM	MUL	FREM
MSP	2.56 (0.43)	2.42 (0.33)	1.97 (0.35)	2.66 (0.46)	2.82 (0.45)	2.31 (0.35)	1.06 (0.34)
DTI	1.14 (0.09)	1.28 (0.12)	1.36 (0.13)	1.78 (0.16)	1.25 (0.14)	1.33 (0.13)	0.96 (0.09)
SBP	1.34 (0.18)	1.57 (0.17)	1.64 (0.16)	2.33 (0.26)	1.68 (0.19)	1.76 (0.17)	1.15 (0.11)

FLR, functional linear regression; FNW, functional Nadaraya–Watson smoothing; FCE, functional conditional expectation; FMO, functional mode, FCM, functional conditional median; MUL, multi-method; FREM, the proposed functional regression on manifold; MSP, meat spectrometric data; DTI, diffusion tensor imaging data; SBP, systolic blood pressure data.

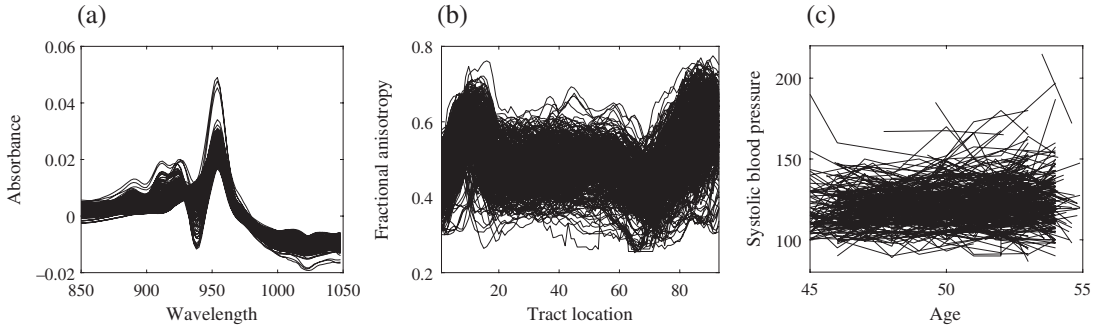

Fig. 2. (a) First derivatives of spectrometric curves of 215 meat samples. (b) Fractional anisotropy profiles of 340 multiple sclerosis patients. (c) Systolic blood pressure measurements of 323 men over time.

The first example aims to predict the fat content of a piece of meat based on a spectrometric curve for the meat from the Tecator dataset consisting of 215 meat samples (Ferraty & Vieu, 2006). For each sample, the spectrometric curve for a piece of finely chopped pure meat was measured at 100 different wavelengths from 850 to 1050 nm. Along with the spectrometric curve, the fat content of each piece of meat was recorded. Compared with the analytic chemistry required to measure the fat content, obtaining a spectrometric curve takes much less time and cost. As in Ferraty & Vieu (2006), we predict the fat content based on the first-derivative curves approximated by the difference quotient between measurements at adjacent wavelengths, shown in Fig. 2(a). Some striking patterns can be seen around the middle wavelengths. The proposed method is able to capture these patterns with a low-dimensional manifold structure. For example, functional linear regression uses on average 15.7 principal components with a standard error of 1.07, while the intrinsic dimension estimated by our method is 5.05 with a standard error of 0.62. Thus, our method predicts the fat content more accurately than the other methods by a significant margin, according to Table 2.

The second example studies the relationship between cognitive function and brain microstructure in the corpus callosum of patients with multiple sclerosis, a demyelinating disease caused by inflammation in the brain. Demyelination refers to damage to the myelin that protects nerve axons and helps nerve signals to travel faster; it occurs in the white matter of the brain and can lead to loss of mobility and cognitive impairment (Jongen et al., 2012). Diffusion tensor imaging, a technique that can produce high-resolution images of white matter by tracing water diffusion within the tissue, is an important method for examining potential myelin damage in the brain.

For example, from such images, some properties of white matter, such as fractional anisotropy of water diffusion, can be derived. It has been shown that fractional anisotropy is related to multiple sclerosis (Ibrahim et al., 2011).

To predict cognitive performance based on fractional anisotropy profiles, we analyse data collected by Johns Hopkins University and the Kennedy-Krieger Institute. The data contain $n = 340$ profiles of multiple sclerosis patients, recorded from a grid of 93 points, and paced auditory serial addition test scores, which quantify cognitive function (Gronwall, 1977). Figure 2(b) shows all the fractional anisotropy profiles, which are considerably more complex than the spectrometric data. The average estimated intrinsic dimension is 5.82 with a standard error of 0.098. By contrast, the average number of principal components for functional linear regression is 11.98 with a standard error of 5.22. According to Table 2, our method provides the most accurate prediction, while the performance of all the other functional nonparametric methods deteriorates substantially.

Our third example concerns prediction of the systolic blood pressure of healthy men and uses anonymous data from the Baltimore Longitudinal Study of Aging. In the study, 1590 healthy male volunteers were scheduled to visit the Gerontology Research Center bi-annually. Their systolic blood pressure and current age were recorded at each visit. The design of the data is sparse and irregular, as many visits were missed by participants or not on the schedule; see Pearson et al. (1997) for more details. We aim to predict the average systolic blood pressure in late middle age, between the ages of 55 and 60, based on the blood pressure trajectory between ages 45 and 55. After excluding subjects who had at most one visit between ages 45 and 55 and no visit between ages 55 and 60, we are left with a subset of the data containing $n = 323$ subjects with on average 4.2 visits per subject, shown in Fig. 2(c). The average of the estimated intrinsic dimensions is 2.4 with a standard error of 0.069, while the average number of principal components for functional linear regression is 4 with a standard error of 2.01. From Table 2 it can be seen that our method outperforms the others significantly.

Acknowledgement

Yao's research was partially supported by the National Natural Science Foundation of China and the Key Laboratory of Mathematical Economics and Quantitative Finance, Peking University, Ministry of Education.

Supplementary material

Supplementary material available at *Biometrika* online includes additional details and simulation studies for sparse functional data, proofs of the main theorems, auxiliary results, and technical lemmas with proofs.

References

Aswani, A., Bickel, P. & Tomlin., C. (2011). Regression on manifolds: Estimation of the exterior derivative. *Ann. Statist.* **39**, 48–81.

Bhattacharya, R. & Lin, L. (2017). Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. *Proc. Am. Math. Soc.* **145**, 413–28.

Bhattacharya, R. & Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31**, 1–29.

Bhattacharya, R. & Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds. II. *Ann. Statist.* **33**, 1225–59.

Cai, T. & Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Ann. Statist.* **39**, 2330–55.

Cardot, H., Ferraty, F. & Sarda, P. (1999). Functional linear model. *Statist. Prob. Lett.* **45**, 11–22.

CARDOT, H. & SARDA, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *J. Mult. Anal.* **92**, 24–41.
CHEN, D. & MÜLLER, H. (2012). Nonlinear manifold representations for functional data. *Ann. Statist.* **40**, 1–29.
CHENG, M. & WU, H. (2013). Local linear regression on manifolds and its geometric interpretation. *J. Am. Statist. Assoc.* **108**, 1421–34.
COIFMAN, R., LAFON, S., LEE, A. B., MAGGIONI, M., NADLER, B., WARNER, F. & ZUCKER, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Nat. Acad. Sci.* **102**, 7426–31.
CORNEA, E., ZHU, H., KIM, P. & IBRAHIM, J. G. (2017). Regression models on Riemannian symmetric spaces. *J. R. Statist. Soc.* B **79**, 463–82.
DAI, X. & MÜLLER, H.-G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *Ann. Statist.* **46**, 3334–61.
DELAIGLE, A. & HALL, P. (2010). Defining probability density for a distribution of random functions. *Ann. Statist.* **38**, 1171–93.
FAN, J. (1992). Design-adaptive nonparametric regression. *J. Am. 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. (1996). *Local Polynomial Modelling and Its Applications.* London: Chapman and Hall.
FERRATY, F., KEILEGOM, I. V. & VIEU, P. (2012). Regression when both response and predictor are functions. *J. Mult. Anal.* **109**, 10–28.
FERRATY, F. & VIEU, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer.
GRONWALL, D. M. A. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. *Percept. Motor Skills* **44**, 367–73.
HALL, P. & KEILEGOM, I. V. (2007). Two sample tests in functional data analysis starting from discrete data. *Statist. Sinica* **17**, 1511–31.
HALL, P. & MARRON, J. S. (1997). On the shrinkage of local linear curve estimators. *Statist. Comp.* **516**, 11–17.
HALL, P., MÜLLER, H.-G. & WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34**, 1493–517.
HUCKEMANN, S., HOTZ, T. & MUNK, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica* **20**, 1–58.
IBRAHIM, I., TINTERA, J., SKOCH, A., JIRØU, F., HLUSTIK, P., MARTINKOVA, P., ZVARA, K. & RASOVA, K. (2011). Fractional anisotropy and mean diffusivity in the corpus callosum of patients with multiple sclerosis: The effect of physiotherapy. *Neuroradiology* **53**, 917–26.
JONGEN, P., TER HORST, A. & BRANDS, A. (2012). Cognitive impairment in multiple sclerosis. *Minerva Medica* **103**, 73–96.
KUDRASZOW, N. L. & VIEU, P. (2013). Uniform consistency of kNN regressors for functional variables. *Statist. Prob. Lett.* **83**, 1863–70.
LANG, S. (1995). *Differential and Riemannian Manifolds*. New York: Springer.
LANG, S. (1999). *Fundamentals of Differential Geometry*. New York: Springer.
LEE, T. C. & SOLO, V. (1999). Bandwidth selection for local linear regression: A simulation study. *Comp. Statist.* **14**, 515–32
LEVINA, E. & BICKEL, P. (2004). Maximum likelihood estimation of intrinsic dimension. In *Proc. 17th Int. Conf. Neural Information Processing Systems (NIPS'04)*. Cambridge, Massachusetts: MIT Press, pp. 777–84.
LI, Y. & HSING, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38**, 3321–51.
LILA, E. & ASTON, J. A. D. (2016). Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *Ann. Appl. Statist.* **10**, 1854–79.
LIN, L., MU, N., CHEUNG, P. & DUNSON, D. (2019). Extrinsic Gaussian processes for regression and classification on manifolds. *Bayesian Anal.* **14**, 887–906.
LIN, L., ST THOMAS, B., ZHU, H. & DUNSON, D. B. (2016). Extrinsic local regression on manifold-valued data. *J. Am. Statist. Assoc.* **112**, 1261–73.
LIN, Z. & YAO, F. (2019). Intrinsic Riemannian functional data analysis. *Ann. Statist.* **47**, 3533–77.
LOH, P.-L. & WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Ann. Statist.* **40**, 1637–64.
MAS, A. (2012). Lower bound in regression for functional data by representation of small ball probabilities. *Electron. J. Statist.* **6**, 1745–78.
MUKHERJEE, S., WU, Q. & ZHOU, D.-X. (2010). Learning gradients on manifolds. *Bernoulli* **16**, 181–207.
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. Am. Statist. Assoc.* **103**, 1534–44.
PANARETOS, V. M., PHAM, T. & YAO, Z. (2014). Principal flows. *J. Am. Statist. Assoc.* **109**, 424–36.
PATRANGENARU, V. & ELLINGSON, L. (2015). *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. Boca Raton, Florida: CRC Press.

Pearson, J., Morrell, C., Brant, L., Landis, P. & Fleg, J. (1997). Age-associated changes in blood pressure in a longitudinal study of healthy men and women. *J. Gerontol. Med. Sci.* **52**, 177–83.

Peng, J. & Müller, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Statist.* **2**, 1056–77.

Ramsay, J. O. & Silverman, B. W. (1997). *Functional Data Analysis*. New York: Springer.

Ramsay, J. O. & Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer, 2nd ed.

Roweis, S. T. & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–6.

Seifert, B. & Gasser, T. (1996). Finite-sample variance of local polynomials: analysis and solutions. *J. Am. Statist. Assoc.* **91**, 267–75.

Sober, B., Aizenbud, Y. & Levin, D. (2020). Approximation of functions over manifolds: A moving least-squares approach. *arXiv:* 1711.00765v4.

Su, J., Kurtek, S., Klassen, E. & Srivastava, A. (2014). Statistical analysis of trajectories on Riemannian manifolds: Bird migration, hurricane tracking, and video surveillance. *Ann. Appl. Statist.* **8**, 530–52.

Tenenbaum, J. B., de Silva, V. & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–23.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. New York: Springer.

van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–605.

Wu, H.-T. & Wu, N. (2018). Think globally, fit locally under the manifold setup: Asymptotic analysis of locally linear embedding. *Ann. Statist.* **46**, 3805–37.

Yao, F. & Müller, H. G. (2010). Functional quadratic regression. *Biometrika* **97**, 49–64.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.* **100**, 577–90.

Yao, F., Müller, H. G. & Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873–903.

Yao, Z. & Zhang, Z. (2020). Principal boundary on Riemannian manifolds. *J. Am. Statist. Assoc.* **115**, 1435–48.

Yuan, M. & Cai, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38**, 3412–44.

Yuan, Y., Zhu, H., Lin, W. & Marron, J. S. (2012). Local polynomial regression for symmetric positive definite matrices. *J. R. Statist. Soc.* B **74**, 697–719.

Zhang, X. & Wang, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Statist.* **44**, 2281–321.

Zhou, L. & Pan, H. (2014). Principal component analysis of two-dimensional functional data. *J. Comp. Graph. Statist.* **23**, 779–801.

[*Received on* 14 *April* 2019. *Editorial decision on* 10 *January* 2020]