Biometrika (2019), **xx**, x, *p*. 1

Printed in Great Britain

Approximating fragmented functional data by segments of Markov chains

BY A. DELAIGLE, P. HALL

School of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010, Australia.

A.Delaigle@ms.unimelb.edu.au

SUMMARY

We consider curve extension and linear prediction for functional data observed only on a part of their domain, in the form of fragments. We suggest an approach based on a combination of Markov chains and nonparametric smoothing techniques, which enables us to extend the observed fragments and construct approximated prediction intervals around them, construct mean and covariance function estimators, and derive a linear predictor. The procedure is illustrated on real and simulated data.

Some key words: Cross-validation; functional principal components; incomplete curve.

1. INTRODUCTION

We consider functional data $X_1, \ldots, X_n$, where each X_i is a random curve defined on an interval $\mathcal{I}_0$. It is common to assume that the curves are completely observed, or that they are observed on a grid of points that are closely spaced within $\mathcal{I}_0$. However, in some problems we can observe each curve X_i on only one or a small number of small subsets of $\mathcal{I}_0$.

One can distinguish principally two types of partially observed functional data. The first case, often referred to as the sparse functional case, corresponds to the situation where each curve is observed at a small number of points that are distributed randomly over $\mathcal{I}_0$. In the second case, which we refer to as the fragmentary functional case, each curve is observed at points that cover a subset of $\mathcal{I}_0$ in such a way that the observations can reasonably be treated as fragments of curves; see the Supplementary Material.

As demonstrated by Delaigle and Hall (2013), the methods appropriate for analysing these two types of incomplete data are different. In the sparse functional case, because the points where each curve is observed are randomly distributed over the whole interval $\mathcal{I}_0$, one can use relatively standard smoothing techniques to estimate the mean and covariance functions. See for example James et al. (2000) and Yao et al. (2005), who also suggest a way of reconstructing the functions on $\mathcal{I}_0$ using a predictor based on the assumption that the data are normally distributed. Those methods are usually inappropriate for the fragmentary case, where each curve is observed on only one or two small subsets of $\mathcal{I}_0$. Delaigle and Hall (2013) suggested a method for reconstructing the functions by adjoining, to each fragment, shifted versions of other observed fragments. However, their method is designed only for situations where we observe only one fragment per curve, and it forces each of the reconstructions to have exactly the shape of an observed fragment. Kraus (2015) considered a related problem, but his setting is closer to that of Yao et al. (2005), since, like there, his fully nonparametric approach requires fragments to cover the whole interval $\mathcal{I}$. There has also been interest in forecasting future observations, but not in our setting of fragmentary data; see for example Liebl (2013) and Goldberg et al. (2014).

In this paper we propose a new approach to reconstructing curves from a sample of fragments, with a particular focus on small fragments. In a first step we discretise the problem in both time and space. For a curve $X(t)$, $t \in \mathcal{I}_0$, we refer to t as time and to $X(t)$ as space. At each discrete point t we model the transition of X from that point to the next by a Markov chain. To estimate the transition probabilities at each discrete t, we borrow information from neighbouring data values using a smoother that takes into account the shapes of the fragments. Most of our attention is dedicated to short and sparse fragments. There, limited by the poor quality of the data, we use low-order Markov chains. In Section 5·2 we extend our approach to higher-order chains that are able to capture more complex structures, but can be employed only when the fragments are longer and less sparse.

Markov modelling enjoys a remarkable diversity of applications, including to climate science and fluid dynamics (Franzke, 2008), to medicine (Strachan et al., 2009), to genomics (Yau et al., 2010), and to many other fields. In these settings, Markov modelling leads to computationally feasible solutions where other approaches often fail. Sometimes this entails compromises. For example, in fluid dynamics, Markov models are attractive even though they require compromises relating to theoretical issues. In particular, when developing continuous-time Markov process models of the atmosphere as a fluid, it can be necessary to consider instantaneous, infinite accelerations at various points. See, for example, Legg and Raupach (1982).

The potential of Markov modelling to supply solutions that might otherwise be out of reach is also useful in applications to functional data. For example, the covariance function, Γ, can be particularly difficult to estimate from fragmented functional data if we use conventional methods, such as that in Yao et al. (2005). Indeed, their curve extension approach relies on the construction of a kernel estimator of Γ, which itself requires observing at least several individual curves at each pair of time points $(s, t) \in \mathcal{I}_0 \times \mathcal{I}_0$. While this is possible in their sparse data setting, it is not possible in the fragmentary context. The Markov modelling approach introduced in this paper enables us to estimate Γ by a relatively simple, explicit estimator. Another attractive property of our approach compared to related methods developed in the sparse setting is that it produces an estimator of the conditional distribution of the missing function values, which makes possible further analysis of conditional characteristics. For example, it enables us to construct approximate prediction intervals around the reconstructed curves, as we discuss in Section 3·2.

Legg and Raupach (1982), among other authors, have pointed to the advantages of using Markov chain rather than Markov process models. A Markov chain is a discrete-time random process, whereas a Markov process, if it does not have jump discontinuities, is a diffusion, and in particular is a continuous process without a first derivative; Ghosh (2011). Therefore the discretisation that we shall introduce in the second paragraph of Section 2 should not be seen as producing grids that will become infinitesimally fine as sample size increases. On the other hand, while it is required in practice, in theory the space discretisation is not really needed, and we discuss this in Section 5·3. Markov models can be used widely in functional data analysis, for example as an aid to building classifiers, but we shall restrict our attention here to their application to covariance estimation, and for solving prediction problems, in the case of fragmented data.

2. MODELS AND DATA

We record independent and identically distributed data (X_i, Y_i), where X_i, an explanatory variable, is a function supported on a compact interval $\mathcal{I}_0 = [a, b]$, and Y_i is a scalar response. However, while the Y_is are all known, we observe X_i only on the interval $\mathcal{I}_i = [A_i, B_i] \subseteq \mathcal{I}_0$, for

$i = 1, \ldots, n$; the case of several disjoint intervals will be discussed in Section 5·1. Our goal is to reconstruct the unobserved parts of the curves X_i, predict the value of Y given that a newly observed X takes the value x, observed on $\mathcal{I} = [A, B] \subseteq \mathcal{I}_0$, and estimate the mean and covariance functions of X. We assume throughout that the fragments of the functions X_i are observed without noise. The procedures we suggest can be adapted to the case where the X_is are observed with additive noise by using deconvolution techniques employed in nonparametric errors-in-variables problems, as in Carroll and Hall (1988) and Delaigle et al. (2008). This problem will be dealt with elsewhere.

We propose to employ Markov chain models for inference based on a discretised version of the process X. Our discretisation is effected in both time and space. We start by constructing a grid of points t_j on the time, or horizontal, axis, $\mathcal{I}_0^{\text{disc}} = \{t_1, \ldots, t_{m_1}\} \subseteq \mathcal{I}_0$, where $a \leq t_1 < \cdots < t_{m_1} \leq b$. The superscript disc denotes discrete. It will simplify notation if the endpoints a and b of $\mathcal{I}_0$ are elements of $\mathcal{I}_0^{\text{disc}}$, and so we make that assumption. Throughout this work, writing $\mathcal{J}$ for a general subinterval of $\mathcal{I}_0$, we let $\mathcal{J}^{\text{disc}} = \mathcal{J} \cap \mathcal{I}_0^{\text{disc}}$.

Next we construct a grid of points $z_1, \ldots, z_{m_2}$ on the state space, or vertical, axis, and reduce the data function $X_i(t)$, $t \in \mathcal{I}_i$, to the set of point pairs $\{t_j, Z_i(t_j)\}$, $t_j \in \mathcal{I}_i^{\text{disc}}$, where $Z_i(t_j)$ takes the value z_k if $(z_{k-1} + z_k)/2 < X(t_j) \leq (z_k + z_{k+1})/2$, and where we define $z_0 = -\infty$ and $z_{m_2+1} = \infty$; see the Supplementary Material. If $Z_i(t_j) = z_k$ then we say that the chain Z_i is in state z_k at time t_j. The t_js and z_ks need not be regularly spaced, but often they will be, and to simplify our discussion we shall assume from now on that they are. We denote the distance between two consecutive t_ks and z_ks by Δt and Δz, respectively. The estimators we shall propose in Section 3·1 involve smoothing, so we can take our discrete grids to be relatively fine. See Section 6·1 for details of practical implementation.

Our notation takes, for simplicity, the time points t_j at which we discretise the function X_i to not depend on i. We anticipate that being the case in practice, not least because we can interpolate among times at which X_i is actually measured. Only minor complications arise if we take the t_js to depend on i, as long as the numbers and spacings do not vary widely.

We first focus on the case where the fragments are sparse and short. There, the data are so poor that progress can only be made if the data vary in a simple way. In this setting, we propose modeling the discretised data using first-order Markov chains. Specifically, we assume that the random process $Z(t)$, for $t \in \mathcal{I}_0^{\text{disc}}$, is a first-order Markov chain, which is true if and only if, whenever $s, t \in \mathcal{I}_0^{\text{disc}}$ with $s < t$, and for all subsets $\mathcal{B}$ of the countable set of states:

$$\text{pr}\{Z(t) \in \mathcal{B} \mid \mathcal{F}_s\} = \text{pr}\{Z(t) \in \mathcal{B} \mid Z(s)\}\,, \qquad (1)$$

where $\mathcal{F}_s$ denotes the sigma-field generated by $Z(u)$ for all $u \in [a, s]^{\text{disc}}$.

The fact that we use this simple model does not imply that we believe that all discretised functional data can be approximated reasonably well by first-order Markov chains. Rather, our solution provides a reasonable approximation in cases where the population consists mostly of curves whose value at a time t is roughly dictated by values in the very recent past. If the observed fragments are longer and less sparse, we can model data with more complex structures by using higher-order Markov chains; see Section 5·2.

Equivalently to (1), the random process $Z(t)$, for $t \in \mathcal{I}_0^{\text{disc}}$, is a first Markov chain if and only if the past and the future are independent, conditional on the present. Since this statement is symmetric in the notions of past and future, then property (1) holding whenever $a \leq s < t \leq b$ is equivalent to the following: Whenever $a \leq t < s \leq b$, where $s, t \in \mathcal{I}_0^{\text{disc}}$,

$$\text{pr}\{Z(t) \in \mathcal{B} \mid \mathcal{G}_s\} = \text{pr}\{Z(t) \in \mathcal{B} \mid Z(s)\}\,, \qquad (2)$$

where $\mathcal{G}_s$ is the sigma-field generated by $Z(u)$ for $u \in [s, b]^{\text{disc}}$.

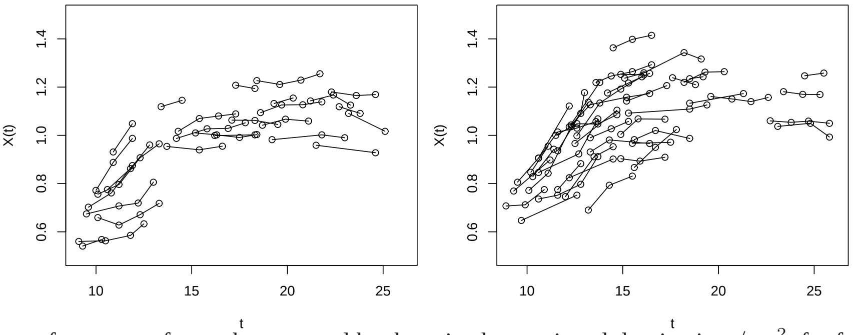

Fig. 1. Curve fragments of growth, measured by the spine bone mineral density, in g/cm^2, for females from the Hispanic (left) and Black ethnic group (right) described in Bachrach et al. (1999).

Similarly, if $\mathcal{B}_1$ and $\mathcal{B}_2$ are subsets of the set of states, if $\mathcal{H}_{s_1 s_2}$ denotes the sigma-field generated by $Z(u)$ for $u \in [s_1, s_2]^{\text{disc}}$, and if $s_1, s_2, u_1, u_2 \in \mathcal{I}_0^{\text{disc}}$, then

$$\text{pr}\big\{Z(u_1) \in \mathcal{B}_1,\ Z(u_2) \in \mathcal{B}_2 \,\big|\, \mathcal{H}_{s_1 s_2}\big\} = \prod_{j=1}^{2} \text{pr}\big\{Z(u_j) \in \mathcal{B}_j \,\big|\, Z(s_j)\big\} \tag{3}$$

when $a < u_1 < s_1 \leq s_2 < u_2 < b$. Moreover, the left-hand side of (3) equals $\text{pr}\{Z(u_1) \in \mathcal{B}_1,\ Z(u_2) \in \mathcal{B}_2 \,|\, Z(s_1)\}$ when $a \leq u_1, u_2 < s_1 \leq s_2 \leq b$, and equals $\text{pr}\{Z(u_1) \in \mathcal{B}_1,\ Z(u_2) \in \mathcal{B}_2 \,|\, Z(s_2)\}$ when $a \leq s_1 \leq s_2 < u_1, u_2 \leq b$.

3. ESTIMATORS AND LINEAR PREDICTORS

3·1. *Transition probabilities and their estimators*

In order to estimate the quantities that interest us, we need to start by estimating the transition probabilities that govern the discrete-time process $Z(t)$, for $t \in \mathcal{I}_0^{\text{disc}}$. These probabilities are defined by $p(t_j, z_{k_1}, z_{k_2}) = \text{pr}\{Z(t_{j+1}) = z_{k_2} \,|\, Z(t_j) = z_{k_1}\}$ and $q(t_{j+1}, z_{k_1}, z_{k_2}) = \text{pr}\{Z(t_j) = z_{k_2} \,|\, Z(t_{j+1}) = z_{k_1}\}$ for each j, k_1 and k_2. In particular, the $p(t_j, z_{k_1}, z_{k_2})$s represent transitions forward in time, whereas the quantities $q(t_{j+1}, z_{k_1}, z_{k_2})$ represent backward transitions.

A naive approach to estimating these probabilities would be to use maximum likelihood estimators. Let $N(t_j, z, z') = \sum_{i=1}^{n} I\{Z_i(t_j) = z, Z_i(t_{j+1}) = z', A_i \leq t_j < B_i\}$ and $N(t_j, z) = \sum_{z'} N(t_j, z, z')$. We show in the Supplementary Material that the maximum likelihood estimators of the $p(t_j, z, z')$s are given by

$$\hat{p}_{\text{ML}}(t_j, z, z') = N(t_j, z, z')/N(t_j, z)\,, \tag{4}$$

where the right-hand side of (4) is interpreted as zero when $N(t_j, z) = 0$.

While this approach is simple, our data are sparse, and in practice the quantities $\hat{p}_{\text{ML}}(t_j, z, z')$ are each generally computed from too few observations. To illustrate this, we consider partially observed growth curves of 153 females from four ethnic groups referred to as Asians, Blacks, Hispanics and Caucasians, from a dataset described in Bachrach et al. (1999), where growth was measured by the spine bone mineral density. There, each woman was typically followed for only a few years, whence the fragmentary nature of the data, depicted in Fig. 1 for the Hispanic and Black ethnic groups. In each group, the fragments are so sparse that for most values of (t_j, z, z'), $\hat{p}_{\text{ML}}(t_j, z, z')$ is equal to zero, whereas the remaining few equal one.

Fortunately, in a functional data context it is reasonable to assume that if two points (t_j, z, z') and (t_k, u, u') are close, then $|p(t_j, z, z') - p(t_k, u, u')|$ is small. This motivates us to introduce smoothing in time and space and estimate $p(t_j, z, z')$ by

$$\hat{p}(t_j, z, z') = \hat{A}(t_j, z, z') \Big/ \sum_{z'} \hat{A}(t_j, z, z') \,, \tag{5}$$

where $\hat{A}(t_j, z, z')$ denotes a smoothed version of the $N(t_k, z, z')$s and where the right-hand side of (5) is interpreted as zero when $\sum_{z'} \hat{A}(t_j, z, z') = 0$.

To define $\hat{A}(t_j, z, z')$, we note that the numbers $N(t_k, z, z')$ will tend to be unduly large, or small, depending on whether the interval $\mathcal{I}_i$ often covers the pair (t_k, t_{k+1}), or covers it infrequently, respectively. To counter the effect of this source of bias, we replace $N(t_k, z, z')$ by

$$\tilde{N}(t_k, z, z') = \frac{\sum_{i=1}^n I\big\{Z_i(t_k) = z, Z_i(t_{k+1}) = z', A_i \le t_k < B_i\big\}}{\sum_{i=1}^n I\big(A_i \le t_k < B_i\big)} \,, \tag{6}$$

and take

$$\hat{A}(t_j, z, z') = \sum_{u,u'} \sum_{k=1}^{m_1-1} \tilde{N}(t_k, u, u')\, w(t_j, t_k)\, \omega\big\{(z, z'), (u, u')\big\} \,, \tag{7}$$

where the weights w and ω are defined below. We take the contribution to the right-hand side of (7), for index k, to equal zero if $\sum_{i=1}^n I\big(A_i \le t_k < B_i\big) = 0$.

Let K be a symmetric, positive and continuous kernel function, and let $h \ge 0$ and $g \ge 0$ be two bandwidths. For $h > 0$, put $K_h(u) = h^{-1}K(u/h)$, and for $h = 0$, let $K_0(u) = 1\{u = 0\}$, and define K_g analogously. We define our weights w and ω by $w(t_j, t_k) = K_h\big(|t_j - t_k|\big)$ and

$$\omega\big\{(z, z'), (u, u')\big\} = K_g\Big\{\alpha\big(|z - u| + |z' - u'|\big) + (1 - \alpha)|z - z' - (u - u')|\Big\} \,,$$

where $\alpha \in [0, 1]$ is a parameter used to control the spatial direction in which smoothing is applied. In particular, the smaller α, the more smoothing will focus on data fragments for which the difference quotient $(u' - u)/\Delta t$ is similar to $(z' - z)/\Delta t$.

Probabilities $q(t_j, z, z')$ of backwards transitions are estimated similarly, producing estimators $\hat{q}(t_j, z, z')$. Our estimators can be easily adapted to satisfy some constraints. For example, to reconstruct functions that are monotone increasing, we can set $\hat{p}(t_j, z, z')$ to zero for all $z' < z$, and rescale the probabilities so that their sum over z' equals one, with similar changes for $\hat{q}$.

3·2. *Imputing the missing parts of curves*

Let X denote a curve observed on $\mathcal{I} = [A, B] \subset \mathcal{I}_0 = [a, b]$, where it takes the value x. That is, for $s \in \mathcal{I} \subset I_0$, $X(s) = x(s)$. We propose estimating the unobserved parts $X(t)$ for $t \in \mathcal{I}_0 \setminus \mathcal{I}$ by an estimator of the predictor $E\{X(t) \mid X(s),\ s \in \mathcal{I}\}$ constructed from the discretised process $Z(t)$, $t \in \mathcal{I}_0^{\text{disc}}$. As before, assume for simplicity that $a, b, A, B \in \mathcal{I}_0^{\text{disc}}$, and define, for $t \in \mathcal{I}_0^{\text{disc}}$,

$$\nu_1(t \mid Z, \mathcal{I}) = E\big\{Z(t) \,\big|\, Z(s)\,,\ s \in \mathcal{I}^{\text{disc}}\big\} = \begin{cases} Z(t), & t \in \mathcal{I}^{\text{disc}}\,, \\ E\{Z(t) \mid Z(A)\}\,, & a \le t < A\,, \\ E\{Z(t) \mid Z(B)\}\,, & B < t \le b\,, \end{cases} \tag{8}$$

where we have used (1) and (2) to obtain the last equality. For $t \in \mathcal{I}_0^{\text{disc}} \setminus \mathcal{I}^{\text{disc}}$ we predict the unobserved $X(t)$ by an estimator $\hat{\nu}_1(t \mid Z, \mathcal{I})$ of $\nu_1(t \mid Z, \mathcal{I})$. For $t \in \mathcal{I}_0 \setminus \mathcal{I}_0^{\text{disc}}$, we can use, for example, linear interpolation.

To construct $\hat{\nu}_1$, we derive explicit formulae for $E\{Z(t)|Z(A)\}$ and $E\{Z(t)|Z(B)\}$. Assuming that $A = t_j$ and $t = t_{j-r+1}$ (or $B = t_j$ and $t = t_{j+r-1}$) for some j and r , we have

$$E\{Z(t) \mid Z(A)\} = \sum_{\ell=1}^{m_2} \left\{ \sum_{\text{paths to } z_\ell} \prod_{k=1}^{r-1} q(t_{j-k+1}, z_{j_k}, z_{j_{k+1}}) \right\} z_\ell \,, \quad a \le t < A \,, \tag{9}$$

$$E\{Z(t) \mid Z(B)\} = \sum_{\ell=1}^{m_2} \left\{ \sum_{\text{paths to } z_\ell} \prod_{k=1}^{r-1} p(t_{j+k-1}, z_{j_k}, z_{j_{k+1}}) \right\} z_\ell \,, \quad B < t \le b \,, \tag{10}$$

where the summation $\sum_{\text{paths to } z_\ell}$ is over all paths $z^0 = z_{j_1} \mapsto z_{j_2} \mapsto \cdots \mapsto z_{j_r} = z_\ell$ that lead from state z^0 to z_ℓ in just $r-1$ steps, with z^0 denoting $Z(A)$ or $Z(B)$ in the cases of (9) and (10), respectively. Our estimator $\hat{\nu}_1(t \mid Z, \mathcal{I})$ of $\nu_1(t \mid Z, \mathcal{I})$ is obtained by replacing each p and q in (9) and (10) by the estimators $\hat{p}$ and $\hat{q}$ derived in Section 3·1. While it may seem complex to find all the paths that lead to z_ℓ, and thus to compute (9) and (10), we show in the Supplementary Material that these can be computed very simply via sums and products of matrices.

An attractive aspect of our approach is that it also provides an estimator of the conditional distribution of $Z(t)|Z(A)$, for $t < A$ and of $Z(t)|Z(B)$, for $t > B$. In particular, using the same notation as above, we have

$$\hat{\mathrm{pr}}\{Z(t) = z_\ell \mid Z(A)\} = \sum_{\text{paths to } z_\ell} \prod_{k=1}^{r-1} \hat{q}(t_{j-k+1}, z_{j_k}, z_{j_{k+1}}) \,, \quad a \le t < A \,,$$

$$\hat{\mathrm{pr}}\{Z(t) = z_\ell \mid Z(B)\} = \sum_{\text{paths to } z_\ell} \prod_{k=1}^{r-1} \hat{p}(t_{j+k-1}, z_{j_k}, z_{j_{k+1}}) \,, \quad B < t \le b \,.$$

We deduce estimators of the conditional cumulative distribution function by taking, for each $z \in \mathbb{R}$ $\hat{\mathrm{pr}}\{Z(t) \le z \mid Z(A)\} = \sum_{\ell=1}^{m_1} I(z_\ell \le z)\, \hat{\mathrm{pr}}\{Z(t) = z_\ell \mid Z(A)\}$ and $\hat{\mathrm{pr}}\{Z(t) \le z \mid Z(B)\} = \sum_{\ell=1}^{m_1} I(z_\ell \le z)\, \hat{\mathrm{pr}}\{Z(t) = z_\ell \mid Z(B)\}$. For any $\beta \in (0,1)$, let $\hat{q}_\beta\{Z(t) \mid Z(A)\}$ and $\hat{q}_\beta\{Z(t) \mid Z(B)\}$ denote the estimators of the conditional quantiles of level β found by inverting these estimators of the conditional cumulative distribution functions. We can construct approximate pointwise prediction intervals of level β for $Z(t)$ by taking $[\hat{q}_{\beta/2}\{Z(t) \mid Z(A)\}, \hat{q}_{1-\beta/2}\{Z(t) \mid Z(A)\}]$ if $t < A$ and $[\hat{q}_{\beta/2}\{Z(t) \mid Z(B)\}, \hat{q}_{1-\beta/2}\{Z(t) \mid Z(B)\}]$ if $t > B$.

To illustrate how effective our method can be, we applied it to the growth data introduced in Section 3·1. For each ethnic group, using only the data from that group, we imputed the missing parts of the curves by $\hat{\nu}_1(t \mid Z, \mathcal{I})$, where, to compute $\hat{p}$ and $\hat{q}$, we chose the parameters as described in Section 6·1. Since, despite small fluctuations, bone mineral density does not usually start decreasing until after 25 years of age, we imposed the monotonicity constraint on $\hat{p}$ and $\hat{q}$ discussed at the end of Section 3·1, but did not alter the few observed non-monotone fragments. To highlight the attractiveness of our method compared to that of Delaigle and Hall (2013), in Fig. 2 we show, for both methods, the reconstructed curves for the Hispanic and the Black ethnic groups. As illustrated in the graphs, a drawback of Delaigle and Hall's (2013) approach is that their reconstructed curves are made of observed fragments, shifted up and down. As a result, if atypical fragments are observed in sparse regions, many of their reconstructed curves inherit this atypicality, whereas our approach is more flexible.

As discussed above, another advantage of our approach is that it enables to construct approximate prediction intervals. To illustrate this, in Fig. 2, we depict the reconstructed curves and the approximate 95% pointwise prediction intervals for three individuals from the Hispanic and the

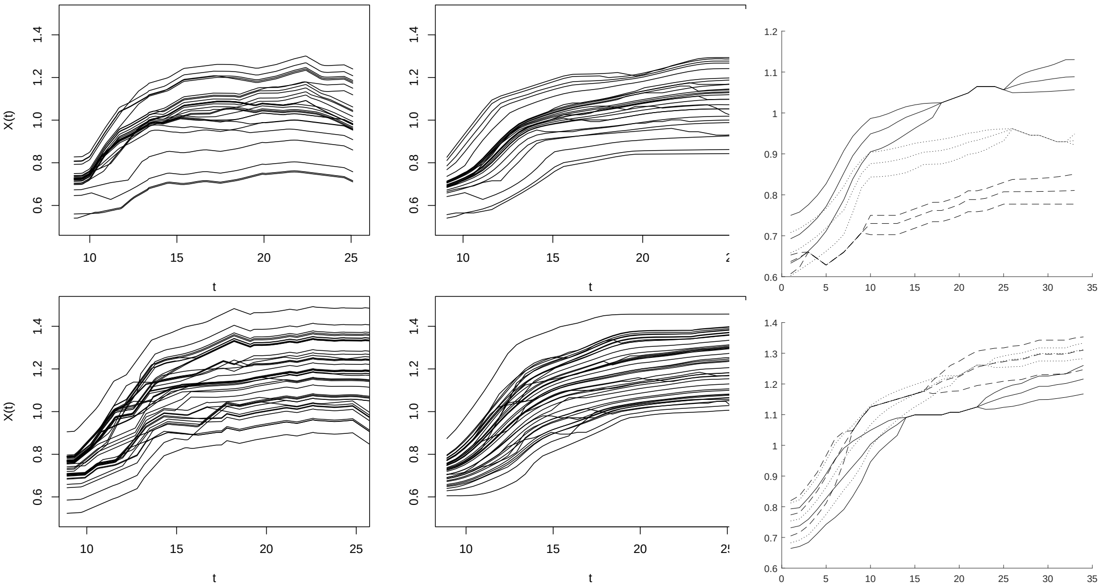

Fig. 2. Reconstruction of growth curves, measured by spine bone mineral density, in g/cm^2, from the incomplete data depicted in Fig. 1, for females from the Hispanic (top) and Black (bottom) ethnic groups described in Bachrach et al. (1999), using Delaigle and Hall's (2013) method (left) or our new approach (middle). Approximate 95% prediction intervals and reconstructed curves using our new approach, for three individuals from the Hispanic group (top) and the Black group (bottom); each individual is shown in a different line type.

Black groups. Each individual is represented with a different line type, with no ambiguity between the reconstructed curves and the upper and lower limits of their prediction intervals as the former always lies between the latter two. Of course, at points t where the curve was observed, there is no prediction interval to construct and we show only the curve itself. We can see that, quite naturally, the length of each prediction interval at a time t increases as t gets further and further away from the interval where the fragment was observed.

As already highlighted by Delaigle and Hall (2013), the methods dedicated to the sparse setting rather than the fragmentary setting, such as those of James et al. (2000) and Yao et al. (2005), do not manage to reconstruct the curves in an attractive way and thus are not discussed here.

3·3. *Estimating the mean and covariance functions*

Our approach can also be used to estimate the mean function $\alpha_1(t) = E\{X(t)\}$ and the covariance function

$$\Gamma(t,u) = E\big[\{X(t) - \alpha_1(t)\}\,\{X(u) - \alpha_1(u)\}\big] = \alpha_2(t,u) - \alpha_1(t)\,\alpha_1(u)\,, \tag{11}$$

where $\alpha_2(t,u) = E\{X(t)\,X(u)\}$. We can estimate readily the function α_1 from the data X_i using the empirical mean $\hat\alpha_1(t) = n^{-1}(t)\,\sum_{i=1}^n X_i(t)\,I(t \in \mathcal{I}_i)$, where $n(t) = \sum_i I(t \in \mathcal{I}_i)$ and it is understood that we regularise the estimator $\hat\alpha_1(t)$ for values of t for which $n(t)$ is too small. However, in the most interesting case where the data fragments are relatively sparse, at each t, $\hat\alpha_1(t)$ defined as above is typically computed from just a few fragments, and is not a very good estimator of α_1. We propose the following Markov-based estimator of α_1, for $t \in \mathcal{I}_0^{\rm disc}$:

$$\hat\alpha_1(t) = \frac{1}{n}\sum_{i=1}^n \hat\nu_1(t \,|\, Z_i, \mathcal{I}_i)\,, \tag{12}$$

where $\hat\nu_1$ is the estimator of ν_1 at (8) derived in Section 3·2.

Estimating α_2 is more complex. Let $\mathcal{I} = [A, B]$ and Z denote generic values of $\mathcal{I}_i$ and Z_i, and assume for simplicity that $a, b, A, B \in \mathcal{I}_0^{\text{disc}}$. For $t, u \in \mathcal{I}_0^{\text{disc}}$, define

$$\nu_2(t, u \mid Z, \mathcal{I}) = \begin{cases} Z(t)\, Z(u)\,, & t, u \in \mathcal{I}^{\text{disc}}, \\ Z(t)\, \nu_1(u \mid Z, \mathcal{I}^{\text{disc}})\,, & t \in \mathcal{I}^{\text{disc}},\, u \notin \mathcal{I}^{\text{disc}}, \\ Z(u)\, \nu_1(t \mid Z, \mathcal{I}^{\text{disc}})\,, & u \in \mathcal{I}^{\text{disc}},\, t \notin \mathcal{I}^{\text{disc}}, \\ E\{Z(t)\, Z(u) \mid Z(A)\}\,, & a \le t, u < A, \\ E\{Z(t)\, Z(u) \mid Z(B)\}\,, & B < t, u \le b, \\ E\{Z(t) \mid Z(A)\}\, E\{Z(u) \mid Z(B)\}\,, & a \le t < A \le B < u \le b\,. \end{cases} \tag{13}$$

In the first and last three cases in (13), we permit t and u to be identical. Much in the same way as we derived formulae (9) and (10), if $a \le t \le u < A$, and assuming that $A = t_j$, $u = t_{j-r_1}$ and $t = t_{j-r_1-r_2}$ for some j, r_1 and r_2, we have

$$E\big\{Z(t)\, Z(u) \,\big|\, Z(A)\big\} = \sum_{\ell_1=1}^{m_2} \sum_{\ell_2=1}^{m_2} \left\{ \sum_{\text{paths to } z_{\ell_1}, z_{\ell_2}} \prod_{k=1}^{r} q(t_{j-k+1}, z_{i_k}, z_{i_{k+1}}) \right\} z_{\ell_1}\, z_{\ell_2}\,, \tag{14}$$

where $r = r_1 + r_2$ and the summation $\sum_{\text{paths to } z_{\ell_1}, z_{\ell_2}}$ is over all paths $z^0 = z_{i_1} \mapsto z_{i_2} \mapsto \cdots \mapsto z_{i_{r+1}} = z_{\ell_2}$ that lead from state z^0, representing $z(A)$, to state z_{ℓ_1} after r_1 steps, and thence to z_{ℓ_2} after another r_2 steps. The case $B < t \le u \le b$ is similar, and the other components of the formula at (13) are trivial. As for equations (9) and (10), (14) can be computed very simply via sums and products of matrices; see the Supplementary Material.

Replacing each p and q, in (14) and its analogues, by the estimators $\hat{p}$ and $\hat{q}$ derived in Section 3·1, we obtain an estimator $\hat{\nu}_2$ of ν_2 and we can compute $\hat{\alpha}_2(t, u) = n^{-1} \sum_{i=1}^{n} \hat{\nu}_2(t, u \mid Z_i, \mathcal{I}_i)$. For $t, u \in \mathcal{I}_0^{\text{disc}}$, we estimate the covariance $\Gamma(t, u)$ at (11) by

$$\hat{\Gamma}(t, u) = \hat{\alpha}_2(t, u) - \hat{\alpha}_1(t)\, \hat{\alpha}_1(u)\,. \tag{15}$$

Remark 1. Estimating the mean is much easier than estimating the covariance, and unlike the latter, the former does not necessarily require methods designed specifically for the fragmentary setting. In particular, even though it was designed for the sparse setting, the smooth mean estimator of Yao et al. (2005) can also work in the fragmentary setting. However, as pointed out by Delaigle and Hall (2013), even when Yao et al.'s (2005) mean estimator works well, their covariance estimator, also designed for the sparse setting, cannot give reasonable results in our fragmentary context. Since our covariance estimator exploits our Markov assumption, to avoid the awkwardness of estimating the mean and covariance functions with incompatible approaches we suggest estimating the mean by the estimator at (12). Delaigle and Hall (2013) also proposed estimators of the mean and covariance functions designed specifically for the fragmentary setting. However their approach works only in the case of a single fragment per curve and it suffers from problems similar to those of their curve reconstruction algorithm.

3·4. *Linear prediction of Y given a new fragment*

Our next goal is to construct a predictor of Y for a new fragment $\{X(t),\ t \in \mathcal{I} \subset \mathcal{I}_0\}$ whose Y is unknown, using a sample of fragmentary data $\big(\{X_i(t), t \in \mathcal{I}_i \subset \mathcal{I}_0\}, Y_i\big)$, for $i = 1, \ldots, n$. The pair (X, Y) has the distribution of the pairs (X_i, Y_i), where X and X_i refer to the unobserved complete curves. As usual, the form of our predictor depends on the way in which we model the relationship between Y and X. A variety of models, parametric or nonparametric, could be considered, but given the sparse nature of our data, we cannot reasonably contemplate using sophisticated models. If we were able to observe the curves $X_1, \ldots, X_n$ on the whole interval $\mathcal{I}_0$, one of the simplest models we could consider would be the functional linear model,

$$Y = \theta_0 + \int_{\mathcal{I}_0} \theta\, X + \epsilon\,, \quad E\big\{\epsilon \,\big|\, X(t)\,,\ t \in \mathcal{I}_0\big\} = 0\,, \tag{16}$$

where θ_0 is a scalar and θ is a function; see for example Cai and Hall (2006), Apanasovich and Goldstein (2008), Crambes et al. (2009), Delaigle et al. (2009) and Lee and Park (2012).

Since we observe only fragments $X_i(t)$, $t \in \mathcal{I}_i \subset \mathcal{I}_0$, we suggest using an approach similar to those employed in the previous sections. Specifically we use the model

$$Y = \theta_0 + \sum_{j=1}^{m_1} \theta_j\, Z(t_j) + \epsilon\,, \quad E\big\{\epsilon \,\big|\, Z(t)\,,\ t \in \mathcal{I}_0^{\rm disc}\big\} = 0\,, \tag{17}$$

where Z is the discrete process introduced in Section 2 and $\theta_0, \ldots, \theta_{m_1}$ are unknown scalars. Then, we predict Y by $\hat{Y} = \hat{\mu}_{\mathcal{I}}(Z)$, where $\hat{\mu}_{\mathcal{I}}(Z)$ is an estimator of

$$\mu_{\mathcal{I}}(Z) = E\Big\{Y \,\Big|\, Z(t)\,,\ t \in \mathcal{I}^{\rm disc}\Big\}\,. \tag{18}$$

Note the distinction between $\mathcal{I}^{\rm disc}$ in (18) and $\mathcal{I}_0{}^{\rm disc}$ in (17).

Using (17), and recalling the definition of ν_1 at (8), we can write $\mu_{\mathcal{I}}(Z) = E\big\{Y \,\big|\, Z(t)\,,\ t \in \mathcal{I}^{\rm disc}\big\} = \theta_0 + \sum_{j=1}^{m_1} \theta_j\, \nu_1(t_j \,|\, Z, \mathcal{I})$. An estimator $\hat{\nu}_1$ of ν_1 was derived in Section 3·2 and to estimate $\hat{\mu}_{\mathcal{I}}$, it suffices to construct estimators of $\theta_0, \ldots, \theta_{m_1}$. While we could estimate these parameters by minimising the sum of squares

$$\sum_{i=1}^{n} \left\{Y_i - \theta_0 - \sum_{j=1}^{m_1} \theta_j\, \hat{\nu}_1(t_j \,|\, Z_i, \mathcal{I}_i)\right\}^2, \tag{19}$$

with $\hat{\nu}_1$ as in Section 3·2, often m_1 is large, and better results can be obtained if we reduce dimension. This can be done in a standard way, as follows. Let $\theta = \big(\theta_1, \ldots, \theta_{m_1}\big)^{\rm T}$, $\nu_{1i} = \big(\hat{\nu}_1(t_1 \,|\, Z_i, \mathcal{I}_i), \ldots, \hat{\nu}_1(t_{m_1} \,|\, Z_i, \mathcal{I}_i)\big)^{\rm T}$, let $\psi_1, \psi_2, \ldots, \psi_{m_1}$ be an orthonormal basis of $\mathbb{R}^{m_1}$, and write $\theta = \sum_{k=1}^{m_1} \beta_k\, \psi_k$, $\nu_{1i} = \sum_{k=1}^{m_1} \gamma_{ik}\, \psi_k$, where $\beta_k = \theta^{\rm T}\, \psi_k$ and $\gamma_{ik} = \nu_{1i}^{\rm T}\, \psi_k$. Keeping only the first m terms of these two series, where the choice of m will be discussed in Section 6·1, we replace the objective function at (19) by

$$S_m(\theta_0; \beta_1, \ldots, \beta_m) = \sum_{i=1}^{n} \left\{Y_i - \theta_0 - \sum_{k=1}^{m} \beta_k \gamma_{ik}\right\}^2. \tag{20}$$

For a given m we compute estimators $\hat{\theta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ of $\theta_0, \beta_1, \ldots, \beta_m$ by minimising (20). Finally we estimate $\mu_{\mathcal{I}}(Z)$ by $\hat{\mu}_{\mathcal{I}}(Z) = \hat{\theta}_0 + \sum_{k=1}^{m} \hat{\beta}_k \gamma_k$, where $\gamma_k = \big(\hat{\nu}_1(t_1 \,|\, Z, \mathcal{I}), \ldots, \hat{\nu}_1(t_{m_1} \,|\, Z, \mathcal{I})\big)\, \psi_k$.

3·5. *Principal component basis*

One possible choice for the orthonormal basis $\psi_1, \ldots, \psi_{m_1}$ employed in Section 3·4 is the sequence $\hat{\phi}_1, \ldots, \hat{\phi}_{m_1}$ obtained through empirical spectral decomposition of the covariance, as follows. For estimation of this basis in the fully observed functional context, see for example Hall and Hosseini-Nasab (2006). For $t_k, t_\ell \in \mathcal{I}_0^{\rm disc}$, the kernels Γ and $\hat{\Gamma}$, introduced in Section 3·3, can be expressed via singular value decompositions as

$$\Gamma(t_k, t_\ell) = \sum_{j=1}^{\infty} \lambda_j \, \phi_j(t_k) \, \phi_j(t_\ell) \,, \quad \hat{\Gamma}(t_k, t_\ell) = \sum_{j=1}^{m_1} \hat{\lambda}_j \, \hat{\phi}_j(t_k) \, \hat{\phi}_j(t_\ell) \,, \tag{21}$$

where (λ_j, ϕ_j) are (eigenvalue, eigenfunction) pairs, $(\hat{\lambda}_j, \hat{\phi}_j)$ are (eigenvalue, eigenvector) pairs, and the eigenvalues are ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots$. In the case of Γ the eigenvalues are all nonnegative, although this is not necessarily true for$\hat{\Gamma}$, and the second expansion at (21) should be interpreted as including terms that correspond to negative eigenvalues. Comparing the two expansions in (21), we view $\hat{\theta}_j$ and $\hat{\phi}_j$ as approximations to θ_j and ϕ_j, respectively. Of course, $\hat{\phi}_j$, a vector, is interpreted as a discrete approximation to the function, ϕ_j.

4. THEORETICAL PROPERTIES

4·1. *Setting and conditions*

Recall that we have discretised time as $\mathcal{I}_0^{\text{disc}} = \{t_1, \ldots, t_{m_1}\} \subseteq \mathcal{I}_0$, where $a \leq t_1 < \cdots < t_{m_1} \leq b$, and discretised space as $z_0 < \cdots < z_{m_2+1}$, where $z_0 = -\infty$ and $z_{m_2+1} = \infty$ but z_1 and z_{m_2} are finite. Recall too that $Z(t_j) = z_k$ if $(z_{k-1} + z_k)/2 < X(t_j) \leq (z_k + z_{k+1})/2$, and that we consider the functions X_i to have been recorded originally on $\mathcal{I}_0$, but that we are able to observe X_i only on $\mathcal{I}_i = [A_i, B_i] \subseteq \mathcal{I}_0$. We assume that:

(a) The X_is are identically distributed, (b) the pairs (A_i, B_i) are identically distributed, and (c) for each pair of consecutive points (t_j, t_{j+1}) in our time grid, $\text{pr}\big([t_j, t_{j+1} \in [A, B]\big) > 0$, where (A, B) has the common distribution of the pairs (A_i, B_i). (22)

Parts (a) and (b) of (22) are conventional, and it is straightforward to appreciate that if (c) fails to hold then it is not possible to estimate the probabilities of transiting from any given time point t_j to an adjacent one, in either direction. We assume that

The random functions X_i, and the random intervals $\mathcal{I}_i = [A_i, B_i]$, are completely independent. (23)

The implications of the assumed independence in (23) are subtle. If (A, B) denotes a generic pair (A_i, B_i), and is not independent of X, then estimators of transition probabilities can be seriously biased, as the following example shows. Assume that (22)(c) holds, and that the following property obtains for particular values of j and k: $[t_j, t_{j+1}] \subseteq [A, B]$ only when the event

$$\mathcal{E} = \Big\{X(t_j) \in \Big(\frac{z_{k-1} + z_k}{2}, \frac{z_k + z_{k+1}}{2}\Big]\Big\} \cap \Big\{X(t_{j+1}) \in \Big(\frac{z_k + z_{k+1}}{2}, \frac{z_{k+1} + z_{k+2}}{2}\Big]\Big\} \,,$$

or equivalently $\mathcal{E} = \{Z(t_j) = z_k\} \cap \{Z(t_{j+1}) = z_{k+1}\}$, holds. We shall call this property (P). Except in the pathological case where $\text{pr}(\mathcal{E}) = 1$, (P) requires that the quantities (A, B) and X not be independent. If (P) holds, and if the process is in fact a first-order Markov chain, then our estimator of $p(t_j, z_k, z_{k+1})$ typically will converge to too large a value. A consequence of this example, and others like it, is that if (23), or a similar condition, does not hold, then no approach to estimating the transition probabilities can be guaranteed to produce consistent estimators, even if the stochastic process Z is a true Markov chain.

4·2. *Theoretical results*

Our methodology uses statistical smoothing to extract as much information as possible from the data. A conventional theoretical account of the performance of that technique would involve so-called infill asymptotics, where the number of data points per unit interval of time was permitted to diverge, retaining the Markov property at each step and also in the asymptotic limit. However, the latter argument is not attractive here, since the continuous process to which a Markov chain converges is a diffusion process, and has only half a derivative. Consequently, theoretical large-sample convergence rates are extremely slow, suggesting a level of performance that is much lower than we find in practice. Therefore, we keep the grid sizes m_1 and m_2 fixed, which implies that our process Z remains discrete in the limit as $n \to \infty$.

Our Markov assumption implies that the discretised process $Z(t)$, for $t \in \mathcal{I}_0^{\rm disc}$, is modeled parametrically; with this model, only a finite number of transition probabilities needs to be estimated. In particular, although smoothing is needed for small samples due to the sparsity of the data, asymptotically our estimator with $h = g = 0$ and $\alpha = 1$ enjoys the optimal $n^{-1/2}$ convergence rate. This result, which will be established in Theorem 1, contrasts with standard smoothing results for continuous random variables, where a bandwidth is not usually allowed to be equal to zero. Thus, unlike more conventional problems, the most interesting and challenging aspects of the fragmentary data, which require us to use smoothing techniques, arise only in small samples. As sample size increases, these difficulties disappear and the asymptotic theory becomes that of a standard parametric problem. In Section 5·3, we shall discuss an alternative approach based on a continuous space model, which has theoretical properties similar to those of standard kernel estimators.

For $j = 1, \ldots, m_1 - 1$ and $k, \ell = 1, \ldots, m_2$, define $p_0(t_j, z_k, z_\ell) = \mathrm{pr}\{Z(t_{j+1}) = z_\ell \mid Z(t_j) = z_k\}$, where this probability is computed using the true probability measure for the random process Z, i.e. without assuming a Markov model. Our theorem addresses the performance of the estimators $\hat{p}$ and $\hat{q}$ of transition probabilities p and q, suggested in Section 3·1. Its proof can be found in the Supplementary Material.

THEOREM 1. *Assume that* (22) *and* (23) *hold. Then for* $j = 1, \ldots, m_1 - 1$ *and* $k, \ell = 1, \ldots, m_2$*, where* m_1 *and* m_2 *are fixed, if we take* $h = g = 0$ *and* $\alpha = 1$*, our estimator at* (5) *satisfies* $\hat{p}(t_j, z_k, z_\ell) = p_0(t_j, z_k, z_\ell) + O_p(n^{-1/2})$*. Moreover, the analogues for transition probabilities* $q(t_j, z_k, z_\ell)$ *also hold.*

Recall the definition of the mean estimator $\hat{\alpha}_1(t)$ at (12), and assume that the conditions of Theorem 1 and the first-order Markov assumption hold. Then it follows from the theorem, and from the weak law of large numbers that, for $t \in \mathcal{I}_0^{\rm disc}$, $\hat{\alpha}_1(t) = \alpha_1(t) + O_P(n^{-1/2})$. Likewise, for $\hat{K}(s,t)$ defined at (15) and $s, t \in \mathcal{I}_0^{\rm disc}$, we have $\hat{K}(s,t) = K(s,t) + O_P(n^{-1/2})$. Similarly, for all $1 \le j, k \le m_1$ such that $j \neq k$, and for $\ell = 1, \ldots, m_2$, the estimator of $E\{Z(t_k) \mid Z(t_j) = z_\ell\}$ derived in Section 3·2 is root-n consistent. In conventional linear prediction for functional linear regression, the linear model at (16), and its counterparts below that formula, are infinite dimensional quantities, a property that they inherit from the infinite dimensional nature of the singular value decomposition of a covariance function. However, in the discrete context of a Markov chain model, discussed in Section 3·4, the covariance is a finite covariance matrix, the model is finite dimensional, and, as noted above, the covariance can be estimated root-n consistently. It comes as no surprise that, for the discrete Markov chain model discussed in Section 3·4, the predictive model is also discrete, and in particular the function θ in (16) is replaced in (17) by a vector the length of which does not diverge with sample size. In particular, $(\theta_1, \ldots, \theta_{m_1})$ in (17) can also be estimated root-n consistently.

5. EXTENSIONS

5·1. *Multiple disjoint fragments*

Our method can be extended to cases where several disjoint fragments of each curve X_i are observed. Recall that we work with the discretised process $Z(t)$, $t \in \{t_1, \ldots, t_{m_1}\}$. A fragment is a sequence of observations $Z(t_j), Z(t_{j+1}), \ldots, Z(t_{j+r})$ for some j and r. Two disjoint fragments correspond to the case where the sequence of observations is of the type $Z(t_j), Z(t_{j+1}), \ldots, Z(t_{j+r_1})$, $Z(t_{j+r_2}), Z(t_{j+r_2+1}), \ldots, Z(t_{j+r_2+r_3})$, for some numbers j, r_1, r_2 and r_3 such that $r_2 > r_1 + 1$. While we could construct smoothed versions of the maximum likelihood estimators of the forward and backward transition probabilities, this time taking into account all the possible paths that connect the various observed fragments of a given curve, this approach is rather complex. Instead we propose estimating the transition probabilities using the smoothing approach introduced in Section 3·1, treating all observed fragments as independent fragments.

On the other hand, when imputing the missing values of a curve X, in order to be able join the fragments in a reasonable manner we need to revise our estimator of ν_1. Likewise, in order to estimate the covariance functions K we need to modify our estimator of ν_2, taking into account the fact that several fragments of the same curve have been observed. We describe in detail a modification for the case where two fragments of a given curve have been observed. This can be generalised to the case of more than two fragments per curve.

Suppose we observe a curve X on two disjoint intervals $\mathcal{I} = [A, B] \subset \mathcal{I}_0 = [a, b]$ and $\mathcal{J} = [C, D] \subset \mathcal{I}_0$. As before, let Z denote the discretised version of X, defined on $\mathcal{I}_0^{\text{disc}}$; and assume that $a \le A < B < C < D \le b$ lie in $\mathcal{I}_0^{\text{disc}}$. To impute the missing values of $X(t)$, for $t \in \mathcal{I}_0^{\text{disc}} \setminus (\mathcal{I}^{\text{disc}} \cup \mathcal{J}^{\text{disc}})$, we use the predictor

$$\begin{aligned} \nu_1(t \mid Z, \mathcal{I}) =& E\big\{Z(t) \,\big|\, Z(s)\,,\ s \in \mathcal{I}^{\text{disc}} \cup \mathcal{J}^{\text{disc}}\big\} \\ =& \begin{cases} Z(t)\,, & t \in \mathcal{I}^{\text{disc}} \cup \mathcal{J}^{\text{disc}}, \\ E\{Z(t) \mid Z(A)\}\,, & a \le t < A, \\ E\{Z(t) \mid Z(D)\}\,, & D < t \le b, \\ E\{Z(t) \mid Z(B), Z(C)\}\,, & B < t < C\,. \end{cases} \end{aligned} \tag{24}$$

The first three terms can be handled in the same way as in the case of one fragment, so here we focus on the last term on the right-hand side of (24).

Write $B = t_j$, $t = t_{j+r_1}$ and $C = t_{j+r}$ for some j, $r_1 \ge 1$ and $r > r_1$. We have $E\{Z(t_{j+r_1}) | Z(t_j) = z_{j_0}, Z(t_{j+r}) = z_{j_r}\} = \big\{ \sum_{j_1,\ldots,j_{r-1}=1}^{m_2} z_{j_{r_1}} \prod_{k=1}^{r} p(t_{j+k-1}, z_{j_{k-1}}, z_{j_k})\big\} / \big\{ \sum_{j_1,\ldots,j_{r-1}=1}^{m_2} \prod_{k=1}^{r} p(t_{j+k-1}, z_{j_{k-1}}, z_{j_k})\big\}$. Similarly, for the covariance function the only difference compared to the case of one fragment per individual is the covariance between two points that are between the two fragments of the same individual. Specifically, for $1 \le r_1 \le r_2 < r_3$ we have $E\{Z(t_{j+r_1})Z(t_{j+r_2}) | Z(t_j) = z_{j_0}, Z(t_{j+r_3}) = z_{j_{r_3}}\} = \big\{ \sum_{j_1,\ldots,j_{r_3-1}=1}^{m_2} z_{j_{r_1}} z_{j_{r_2}} \prod_{k=1}^{r_3} p(t_{j+k-1}, z_{j_{k-1}}, z_{j_k})\big\} / \big\{ \sum_{j_1,\ldots,j_{r_3-1}=1}^{m_2} \prod_{k=1}^{r_3} p(t_{j+k-1}, z_{j_{k-1}}, z_{j_k})\big\}$. In practice, we compute all these equations in a matrix form; see the Supplementary Material.

5·2. *Higher-order Markov chains*

In the previous sections, we approximated the discretised process by first-order Markov chains, which can be fitted even if the data consist of a small number of very short fragments, as in the growth data example. On the other hand, this simple model is a reasonable approximation to the truth only if, at each time t_{k+1}, the main shape of the curves depends mostly on their shape at the previous time t_k. If not, for example if the population consists of overlapping increasing and decreasing curves, so that given that $Z(t_k) = z_j$, with high probability, $Z(t_{k+1})$ can either increase or decrease, then the first-order Markov approximation will not work well in practice. In such instances, if we observe only few very short fragments, sophisticated models cannot be fitted well from the data.

If the fragments are less sparse, more complex structures can be captured through higher-order Markov chains. Under an mth-order Markov assumption, using the notation $p(t_k, z_{j_0}, \ldots, z_{j_m}) = \mathrm{pr}\{Z(t_{k+m}) = z_{j_m} \mid Z(t_{k+m-1}) = z_{j_{m-1}}, \ldots, Z(t_k) = z_{j_0}\}$, we have $\mathrm{pr}\{Z(t_{k+m}) = z_{j_m} \mid Z(t_{k+m-1}) = z_{j_{m-1}}, \ldots, Z(t_1) = z_{j_{1-k}}\} = p(t_k, z_{j_0}, \ldots, z_{j_m})$. Let $\tilde{N}(t_\ell, u_0, \ldots, u_m)$ be equal to

$$\frac{\sum_{i=1}^n I\big\{Z_i(t_\ell) = u_0, \ldots, Z_i(t_{\ell+m}) = u_m, A_i \le t_\ell < t_{\ell+m} \le B_i\big\}}{\sum_{i=1}^n I\big\{A_i \le t_\ell < t_{\ell+m} \le B_i\big\}}$$

if $\sum_{i=1}^n I\big\{A_i \le t_\ell < t_{\ell+m} \le B_i\big\} \neq 0$, and to zero otherwise. Generalising our estimator at (5) to mth-order Markov chains, we take

$$\widehat{p}(t_k, z_{j_0}, \ldots, z_{j_m}) = \hat{A}(t_k, z_{j_0}, \ldots, z_{j_m}) / \sum_{z_{j_m}} \hat{A}(t_k, z_{j_0}, \ldots, z_{j_m})\,,$$

where

$$\hat{A}(t_k, z_{j_0}, \ldots, z_{j_m}) = \sum_{\ell=1}^{m_1 - m} \sum_{u_0, \ldots, u_m} \tilde{N}(t_\ell, u_0, \ldots, u_m) \\ \times w(t_k, t_\ell)\, \omega\big\{(z_{j_0}, \ldots, z_{j_m}), (u_0, \ldots, u_m)\big\}\,,$$

with w as in Section 3·1 and ω a generalisation of the weight from Section 3·1; for example $\omega\{(z_{j_0}, \ldots, z_{j_m}), (u_0, \ldots, u_m)\} = K_g\big[\alpha\{|z_{j_0} - u_0| + \ldots + |z_{j_m} - u_m|\} + (1 - \alpha)\{|u_m - u_{m-1} - (z_{j_m} - z_{j_{m-1}})| + \ldots + |u_1 - u_0 - (z_{j_1} - z_{j_0})|\}/m\big]$.

As the order m increases, the Markov model becomes more sophisticated and can capture more complex structures, but more data are required to fit it, and the observed fragments need to be longer. However, second-order Markov chains can already be quite useful, since they enable us to distinguish curves with an upward trend from those with a downward trend. The techniques derived in Sections 3·2 to 3·4 can be extended to higher-order Markov chains. Here we show how to extend to second-order chains the imputing algorithm from Section 3·2; similar ideas can be applied for higher-order chains and for the methods developed in Sections 3·3 and 3·4. Under the second-order Markov assumption, letting $\mathcal{I}^{\mathrm{disc}} = [A, B] = [t_j, t_k]$ with $k > j$, (8) becomes

$$E\big\{Z(t) \,\big|\, Z(s)\,,\ s \in \mathcal{I}^{\mathrm{disc}}\big\} = \begin{cases} Z(t)\,, & t \in \mathcal{I}^{\mathrm{disc}} \\ E\{Z(t) \mid Z(t_j), Z(t_{j+1})\}\,, & a \le t < t_j \\ E\{Z(t) \mid Z(t_{k-1}), Z(t_k)\}\,, & t_k < t \le b\,, \end{cases}$$

where quantities such as the one at the second and third lines of the last equation can estimated using approaches similar to the one in Section 3·2. For example, in the case of the third line, letting $t = t_{k+r}$ with $r \ge 1$, we have $E\{Z(t) \mid Z(t_{k-1}) = z_{\ell_{-1}}, Z(t_k) = z_{\ell_0}\} = \sum_{\ell_1, \ldots, \ell_r = 1}^{m_2} z_{\ell_r} \prod_{i=1}^r \mathrm{pr}\{Z(t_{k+i}) = z_{\ell_i} \mid Z(t_{k+i-1}) = z_{\ell_{i-1}}, Z(t_{k+i-2}) = z_{\ell_{i-2}}\}$, which can be estimated by replacing the probabilities by the estimators derived above.

5·3. *Continuous space model*

In the last paragraph of the introduction, we have discussed the advantages of using a discrete time Markov chain instead of a continuous time Markov process. On the other hand, in theory it is not necessary to discretise space, and in this section we discuss an alternative approach that uses a continuous space model. As we shall see, it leads to estimators that are very similar to those derived using our discrete model.

In the continuous space model, instead of assuming that the discretised process $Z(t)$ is a first-order Markov chain, we make the first-order Markov assumption on the process $X(t)$, where $t \in \{t_1, \ldots, t_{m_1}\}$. That is, we assume that, for each positive integer j ad r such that $j + r \leq m_1$, the conditional density of $X(t_{j+r})$ given $X(t_1), \ldots, X(t_j)$ satisfies $f_{X(t_{j+r})|X(t_j),\ldots,X(t_1)} = f_{X(t_{j+r})|X(t_j)}$. As in Section 3·2, let X denote a curve observed on $\mathcal{I} = [A, B] \subset \mathcal{I}_0 = [a, b]$, let $\mathcal{I}_0^{\text{disc}}$ and $\mathcal{I}^{\text{disc}}$ be as defined in that section, and assume for simplicity that $a, b, A, B \in \mathcal{I}_0^{\text{disc}}$.

Under the above continuous version of the Markov assumption, to estimate the unobserved parts $X(t)$ for $t \in \mathcal{I}_0^{\text{disc}} \setminus \mathcal{I}^{\text{disc}}$, we can directly construct an estimator of the predictor

$$E\big\{X(t) \,\big|\, X(s)\,,\ s \in \mathcal{I}^{\text{disc}}\big\} = \begin{cases} X(t) & \text{if } t \in \mathcal{I}^{\text{disc}} \\ E\{X(t) \,|\, X(A)\}\,, & a \leq t < A \\ E\{X(t) \,|\, X(B)\}\,, & B < t \leq b\,. \end{cases} \tag{25}$$

Similarly to our derivations in Section 3·2, assuming that $B = t_j$ and $t = t_{j+r-1}$ for some j and r, we have $E\{X(t) \,|\, X(B) = x_j\} = \int \ldots \int x_{j+r-1} \prod_{k=1}^{r-1} \big[f_{X(t_{j+k})|X(t_{j+k-1})}(x_{j+k}|x_{j+k-1}) \cdot 1\{f_{X(t_{j+k-1})}(x_{j+k-1}) > 0\}\big] \prod_{k=1}^{r-1} dx_{j+k}$, and a similar expression holds for $E\{X(t) \mid X(A)\}$.

In practice these integrals need to be approximated by sums involving the conditional densities, and then the conditional densities must be estimated from the data. No matter which approach we use, the error of the resulting predictor has two parts: a systematic error Err_{int} caused by the numerical integration, which does not tend to zero as n increases because of the limitations imposed by the computational feasibility; and the error caused by the estimation of the conditional densities, which tends to zero as n increases. We illustrate this below through the simple midpoint integral approximation, which is easy to explain and which can be directly connected to our discretised approach from Section 3·2. More sophisticated numerical integration rules can be applied, but the main ideas remain the same.

Reflecting most practical situations, assume that the $X(t_k)$'s are bounded. Specifically, assume that for each k, $z_1 - \Delta z/2 \leq X(t_k) \leq z_{m_2} + \Delta z/2$, where the z_j's and Δz are as in Section 2. Letting $z_{\ell_j} = x_j$, the midpoint integration rule leads to

$$E\{X(t) \,|\, X(B) = x_j\} = (\Delta z)^{r-1} \sum_{\ell_{j+1}=1}^{m_2} \ldots \sum_{\ell_{j+r-1}=1}^{m_2} z_{\ell_{j+r-1}} \prod_{k=1}^{r-1} \big[f_{X(t_{j+k})|X(t_{j+k-1})}(z_{\ell_{j+k}}|z_{\ell_{j+k-1}}) \cdot 1\{f_{X(t_{j+k-1})}(z_{\ell_{j+k-1}}) > 0\}\big] + \text{Err}_{\text{int}}\,, \tag{26}$$

where the integration error Err_{int} is bounded away from zero – while it decreases with Δz at a rate depending on the number of bounded derivatives of the conditional densities, Δz is always bounded away from zero because, for the sums to be computable, m_2 must be finite.

In practice, the conditional densities are unknown and need to be estimated from the data. Assuming that the densities $f_{X(t_{j+k}),X(t_{j+k-1})}$ and $f_{X(t_{j+k-1})}$ are all continuous at the boundary of their support, we can use relatively standard kernel density estimators, where, to overcome the sparsity of the observations at any given time point, we smooth also in time, as we did in the previous sections for the discrete space model. Namely, for each k we can take

$$\hat{f}_{X(t_{k+1})|X(t_k)}(z_{\ell_{k+1}}|z_{\ell_k}) \\ = \frac{\sum_{\ell=1}^{m_1-1}\sum_{i=1}^{n} w(t_k,t_\ell)\,K_g\big\{z_{\ell_k}-X_i(t_\ell)\big\}K_g(z_{\ell_{k+1}}-X_i(t_{\ell+1})\big\}I\{A_i\leq t_\ell<B_i\}}{\sum_{\ell=1}^{m_1-1}\sum_{i=1}^{n} w(t_k,t_\ell)I\{A_i\leq t_\ell<B_i\}K_g\big\{z_{\ell_k}-X_i(t_\ell)\big\}}\,,$$

with w (which depends on a bandwidth h), K and g as in the previous sections. Plugging these conditional density estimators into (26), we get our estimator $\hat{E}\{X(t)\,|\,X(B)\}$ of $E\{X(t)\,|\,X(B)\}$. Comparing with the estimator from Section 3·2, we can see that the main difference is that the conditional transition probabilities are replaced by the conditional densities.

Using standard arguments as in De Gooijer and Zerom (2003), it can be proved that if (22) and (23) hold, the conditional and marginal densities are bounded and twice continuously differentiable with all partial derivatives bounded, K is such that $\int K(x)\,dx=1$, $\int xK(x)\,dx=0$, $0<\int x^2K(x)\,dx<\infty$ and $\int K^2(x)\,dx<\infty$, $h=0$ and $g\asymp n^{-1/6}$, then we have

$$\hat{E}\{X(t)\,|\,X(B)=x\}=E\{X(t)\,|\,X(B)=x\}+\mathrm{Err}_{\mathrm{int}}+O_P(n^{-1/3})\,.$$

Recalling that $\mathrm{Err}_{\mathrm{int}}$ is bounded away from zero even as $n\to\infty$, we deduce that, asymptotically, the prediction error is dominated by the numerical integration error $\mathrm{Err}_{\mathrm{int}}$.

6. Numerical results

6·1. *Details of implementation*

Our procedure depends on several parameters that need to be selected in practice: the size of two grids $t_1<\cdots<t_{m_1}$ and $z_1<\cdots<z_{m_2}$, two bandwidths h and g, and a parameter α. While it would be possible to select all of them in a data-driven way by cross-validation, the smoothing nature of our approach implies that, as long as m_1 and m_2 are relatively large, increasing them further typically has the same effect as reducing h and g. Therefore, we fix m_1 and m_2 at relatively large values, see below, and choose only h, g and α by cross-validation.

We apply the same amount of smoothing in standardised time and space. Given the definition of our weights w and ω in Section 3·1, we implement this by taking $h=g\,\hat{\sigma}_t/(2\hat{\sigma}_Z)$, where $\hat{\sigma}_t$ and $\hat{\sigma}_Z$ are the empirical standard deviations of, respectively, the time points corresponding to all data fragments, and the values $Z_i(t_j)$ corresponding to all data fragments. We choose g and α by cross-validation. Specifically, if our goal is to reconstruct the missing parts of the curves as in Section 3·2, see for example Fig. 2, we choose g and α by minimising $\sum_{i=1}^{n}\sum_{t\in\{A_i,B_i\}}\big|X_i(t)-\hat{\nu}_1^{(-i)}(t\,|\,Z_i,\mathcal{I}_i\setminus\{t\})\big|$, where $\hat{\nu}_1^{(-i)}$ denotes the estimator $\hat{\nu}_1$ computed without using X_i. If our goal is linear prediction, as in Section 3·4, we choose g, α, and m in (20) by minimising $\sum_\ell\{Y_\ell-\hat{\mu}_{\ell,\mathcal{I}_\ell}(Z_\ell)\}^2$, where, for $1\leq\ell\leq n$, $\hat{\mu}_{\mathcal{I}_\ell}^{(-\ell)}=\hat{\theta}_0^{(-\ell)}+\sum_{k=1}^{m}\hat{\beta}_k^{(-\ell)}\gamma_k$, with $\hat{\theta}_0^{(-\ell)}$ and the $\hat{\beta}_k^{(-\ell)}$s obtained by minimising the version of (20) with $\sum_{i=1}^{n}$ replaced by $\sum_{i\neq\ell}$.

Next we discuss the choice of the discrete grids. As argued above, since we choose h, g and α in a data-driven way, the grids can be chosen quite flexibly. We suggest using the following default grids, which, in our experience, often contain enough points. Users who do not feel comfortable about this could choose the size of the grids by cross-validation but we argue that this is unnecessarily complicated. Since the states usually differ more among individuals than do the times, by default we take more states than times. For the states, we employ $m_2=100$ equally spaced points, and to cover a broad range of values we take 90% of those points in $[L,U]$, where $L=\min_{i,t\in\mathcal{I}_i}X_i(t)$ and $U=\max_{i,t\in\mathcal{I}_i}X_i(t)$. That is, we take $z_1=L-(5\,\Delta_X\,/90)$ and

$z_{m_2} = U + (5\,\Delta_X/90)$, where $\Delta_X = U - L$. We take the time points t_j to be equally spaced on the interval $\mathcal{I}_0 = [a, b]$, and satisfying two criteria that prevent us from using too small a grid. First, we use at least $m_1 = 35$ time points. This value is chosen in much the same spirit as Ruppert, Wand and Carroll's (2003) recommendation that 35 knots are often enough for constructing splines in practice. Second, we want the $\mathcal{I}_i$s to contain overall at least two or three consecutive points, although enforcing this constraint strictly could result in m_1 being too large. With this in mind, we take the distance Δt between two consecutive points to be no more than one third of the median fragment length $\text{med}(B_i - A_i)$. Combining these two criteria together, we put $\Delta t = \min\big\{(b-a)/34, \text{med}(B_i - A_i)/3\big\}$.

Of course, for all $a \le t_k < b$ our estimated transition probabilities should be such that if $\max_y p(t_k, y, z_j) > 0$, then $\max_z p(t_{k+1}, z_j, z) > 0$, which is not always satisfied for all values of the smoothing parameters; in such cases some of the probabilities need to be modified. We show how to do this in the Supplementary Material.

6·2. *Simulation study: curve reconstruction*

We applied our curve reconstruction algorithm from Section 3·2 to simulated examples, and compared it with Delaigle and Hall's (2013) technique. To our knowledge, this is the only competing approach designed for short fragments. As indicated earlier, methods for sparse functional data, such as those of James et al. (2000) and Yao et al. (2005), are not appropriate for fragments and, in general, do not work well when applied to fragments. A referee asked us to prove this via simulations, which we did; see the Supplementary Material.

We generated data from several settings, with either one fragment per curve, observed on an interval $\mathcal{I}_i = [A_i, B_i] \in \mathcal{I}_0 = [1, 100]$, or two fragments observed on $\mathcal{I}_i = [A_i, B_i] \cap [C_i, D_i]$. Since Delaigle and Hall's (2013) method can handle only one fragment per curve, we compare our approach with theirs only in the single fragment cases considered below. In each setting we generated $M = 100$ samples of data $(X_i, \mathcal{I}_i)$, for $i = 1, \ldots, n$, where $n = 30$ or $n = 50$, from three models used by Delaigle and Hall (2013) and adapted to the single population setting.

To illustrate our approach with a single fragment per curve, we considered three cases. For each we took $\mathcal{I}_i = [A_i, B_i]$ where $A_i = [U_i]$, with $[\cdot]$ denoting the nearest integer to $\cdot$, and $B_i = \min(A_i + [V_i], 100)$, with $U_i \sim U[1, 95]$, $V_i \sim U[7, 15]$ in case (i) and $V_i \sim U[5, 10]$ in cases (ii) and (iii). In cases (i) and (ii), $X_i(t) = m(t - W_{i1}) + \big\{W_{i2} + W_{i3}\sin(t/W_{i4})\big\}\{W_{i5} + \sin(t\pi/1000)\}$, where, in case (i), $m(t) = \exp\{0.1(t-40)\}/[1 + \exp\{0.1(t-40)\}]$, $W_{i1} \sim U[-5, 10]$, $W_{i2} \sim U[-0.75, 0.75]$, $W_{i3} \sim U[0.02, 0.05]$, $W_{i4} \sim U[2, 3]$, $W_{i5} \sim U[0.1, 0.5]$. In case (ii), $m(t) = \sin(t/15)/\{0.1(t-5)^2 + 1\}$, W_{ik} for $k \ne 2$ as in case (i), and $W_{i2} \sim U[-1, 1]$. In case (iii), $X_i(t) = m(t) + W_{i1} + W_{i2}\sin(t/W_{i3}) + W_{i4}$, where $m(t) = \sin(t/15)/\{0.1(t-5)^2 + 1\}$, $W_{i1} \sim U[-1, 1]$, $W_{i2} \sim U[0.02, 0.05]$, $W_{i3} \sim U[2, 3]$ and $W_{i4} \sim \text{N}(0, 0.04)$. To illustrate our approach with two fragments per curve, we modified cases (i) and (ii) by replacing $\mathcal{I}_i$ there by $\mathcal{I}_i = [A_i, B_i] \cup [C_i, D_i]$, where $A_i = [U_i]$ with $U_i \sim U[1, 95]$, $B_i = \min([A_i + V_i], 100)$ with $V_i \sim U[5, 10]$, and $[C_i, D_i] \sim [A_i, B_i]$.

For $i = 1, \ldots, n$, let $\hat{X}_i$ denote the curves reconstructed either by our method or by that of Delaigle and Hall (2013). To measure the performance of the reconstruction algorithm, we computed for both methods the average absolute error $\text{AAE} = n^{-1}\sum_{i=1}^{n}\int_{\mathcal{I}_0}|\hat{X}_i(t) - X_i(t)|\,dt$. In Table 1 we report, for both methods, the median and interquartile range of AAE computed over the M simulated samples. The results show that in all cases, our new method improves, sometimes considerably, the method of Delaigle and Hall (2013). In particular, while their procedure can sometimes work almost as well as the new procedure for simple cases where the curves are monotone, as in model (i), it has more difficulties with curves whose shape is more complex, such as those from models (ii) and (iii), especially when the number of fragments is small.

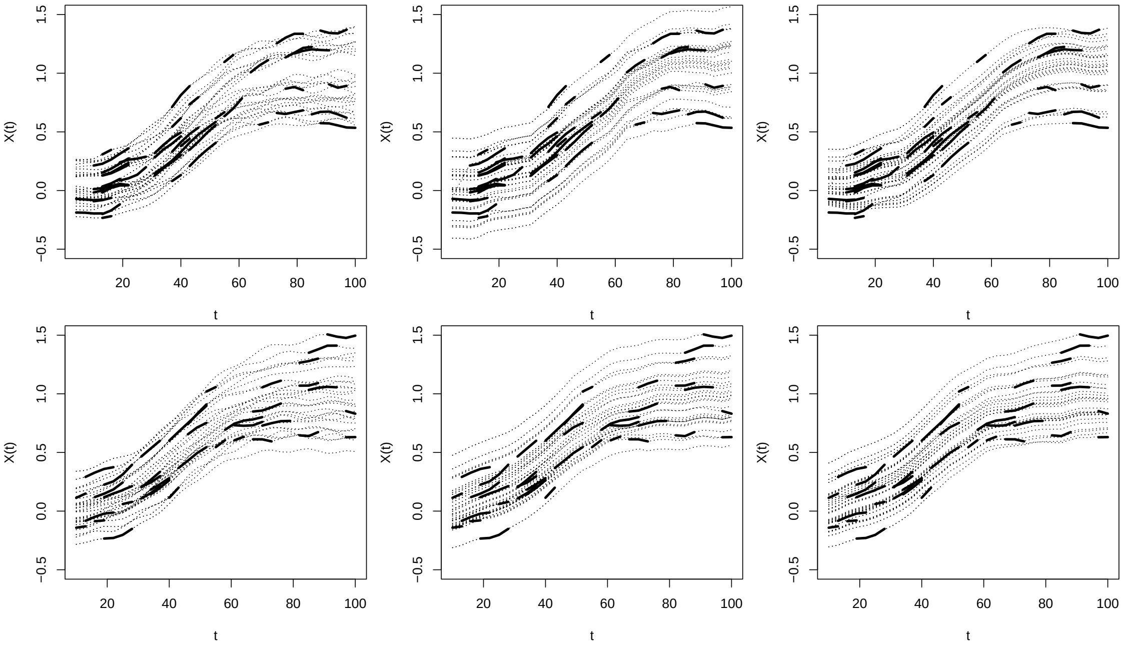

Fig. 3. Reconstruction of $n = 30$ curves for two samples (row 1 and row 2) from model (i) . True curves (left), and reconstructions using Delaigle and Hall's (2013) method (middle) or our new approach (right). The observed fragments are shown in bold.

Table 1. *Median (interquartile range) of M calculated AAEs for the curve reconstruction algorithm in models (i) to (iii), using our suggested approach, denoted here by NEW, in the cases of one and two fragments, denoted here by one frag and two frags, and, for one fragment, the method of Delaigle and Hall (2013) denoted here by DH (2013).*

	Model (i)			Model (ii)			Model (iii)	
	One frag		Two frags	One frag		Two frags	One frag	
n	NEW	DH (2013)	NEW	NEW	DH (2013)	NEW	NEW	DH (2013)
30	5.7 (1.3)	7.3 (4.3)	4.8 (1.0)	12.7 (5.5)	19.4 (12.3)	8.6 (2.4)	9.8 (4.7)	27.1 (25.1)
50	5.3 (0.9)	5.6 (1.3)	4.0 (0.6)	10.9 (2.5)	14.6 (6.6)	7.8 (1.3)	7.2 (2.5)	10.5 (14.9)

To illustrate this graphically, in Figs. 3 and 4 we show samples of $n = 30$ observed fragments from cases (i) to (iii), the corresponding true unobserved curves X_i, and the curves reconstructed by our new procedure and by that of Delaigle and Hall (2013). These samples were chosen to exemplify the properties discussed in the previous paragraph. In Fig. 3, both methods provide good reconstructions of the simple monotone curves corresponding to samples of size $n = 30$ generated by model (i). However, in Fig. 4, we can see that by the very nature of Delaigle and Hall's (2013) algorithm, which is based on copying shifted observed fragments, when the curves are more complex and the sample of fragments is sparse in some regions, the curves can be reconstructed poorly, whereas the smoothing nature of our new approach tends to guard us against such dramatic problems. Both procedures improve as n increases and our approach improves as the number of fragments per data curve increases. In addition to significantly improving Delaigle and Hall's (2013) approach, our method has the advantage that it can be applied to the case where we observe multiple fragments per curve, unlike Delaigle and Hall's (2013) approach.

6·3. *Simulation study: prediction*

To illustrate our linear prediction algorithm introduced in Section 3·4, for each of the $M = 100$ samples of data $(X_i, \mathcal{I}_i)$, for $i = 1, \ldots, n$ where $n = 30$, $n = 50$ or $n = 100$, generated

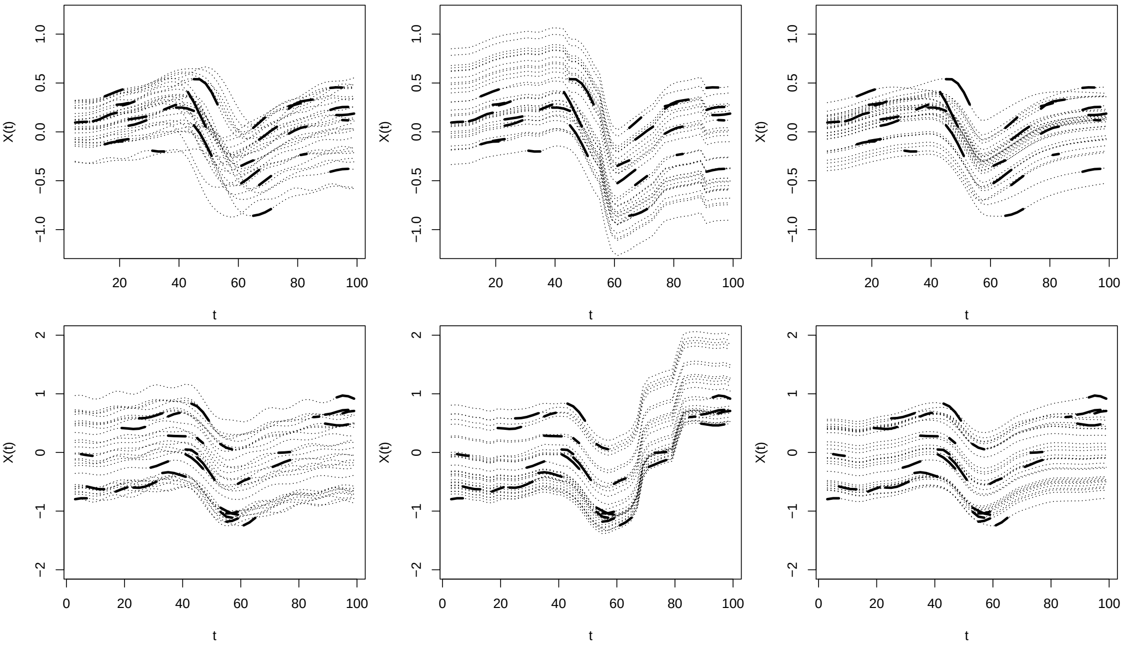

Fig. 4. Reconstruction of $n = 30$ curves for a sample from model (ii) (row 1) or from model (iii) (row 2). True curves (left), and reconstructions using Delaigle and Hall's (2013) method (middle) or our new approach (right). The observed fragments are shown in bold.

Table 2. *Median (interquartile range) of M calculated APEs for predicting values $Y_{\mathrm{NEW},i}$ in models (i) and (ii), using our suggested approach in the cases of one and two fragments, denoted here by one frag and two frags, or using the full non-fragmented data (full).*

		Model (i)			Model (ii)	
n	Full	One frag	Two frags	Full	One frag	Two frags
30	0.09 (0.01)	0.29 (0.09)	0.22 (0.05)	0.09 (0.01)	0.63 (0.14)	0.52 (0.11)
50	0.08 (0.01)	0.26 (0.05)	0.20 (0.03)	0.08 (0.01)	0.58 (0.11)	0.46 (0.06)
100	0.08 (0.01)	0.22 (0.03)	0.17 (0.02)	0.08 (0.01)	0.51 (0.07)	0.41 (0.05)

from models (i) and (ii) in the previous section, we also generated $Y_i = \theta_0 + \int_{\mathcal{I}_0} \theta X_i + \epsilon_i$, $i = 1, \ldots, n$, where $\theta_0 = 0.1$ and $\theta(t) = 5\,\phi\{0.1(t-50)\}$, with ϕ denoting the standard normal density and $\epsilon_i \sim \mathrm{N}(0, 0.1^2)$. Then we applied our predictor from Section 3·4, with the basis chosen as in Section 3·5, to $n_{\mathrm{NEW}} = 100$ new fragmented data $(X_{\mathrm{NEW},i}, \mathcal{I}_{\mathrm{NEW},i})$, for $i = 1, \ldots, n_{\mathrm{NEW}}$. The value of Y, $Y_{\mathrm{NEW},i}$, was not observed.

We compared our predictor with the standard linear functional predictor using the full unfragmented data curves, where the data are projected onto the standard functional principal component basis, with the number of basis functions chosen by cross-validation. Let$\widehat{Y}_{\mathrm{NEW},i}$ denote the predicted $Y_{\mathrm{NEW},i}$ computed using either predictor. To measure performance, we computed, for each predictor, the average prediction error $\mathrm{APE} = n_{\mathrm{NEW}}^{-1} \sum_{i=1}^{n_{\mathrm{NEW}}} |\widehat{Y}_{\mathrm{NEW},i} - Y_{\mathrm{NEW},i}|$. The results are shown in Table 2 where we report, for the full data and for the partial data observed on one or two fragments, the median and interquartile range of APE computed over the M simulated samples. Unsurprisingly, the predictor computed from the full data performs significantly better than that for fragmented data. The performance of our predictor improves as n increases and as the number of fragments per curve increases. Despite the difficulty of the fragmented case, our predictor works reasonably well. See Fig. 5, where we present scatterplots of pairs $(Y_{\mathrm{NEW},i}, \widehat{Y}_{\mathrm{NEW},i})$, for $i = 1, \ldots, 100$, when $\widehat{Y}_{\mathrm{NEW},i}$ is computed using our predictor.

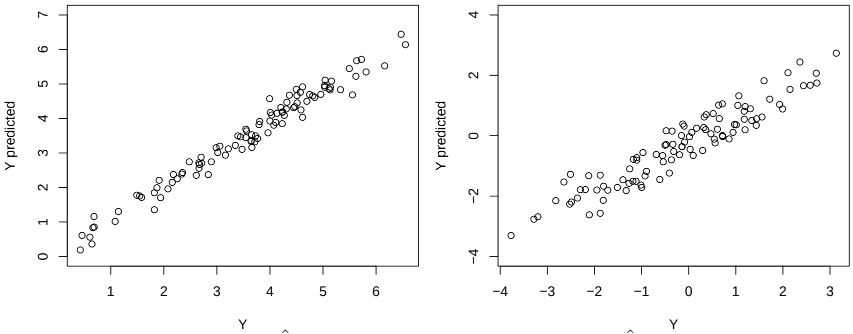

Fig. 5. Scatterplots of pairs $(Y_{\mathrm{NEW},i}, \hat{Y}_{\mathrm{NEW},i})$, for $i = 1, \ldots, 100$, when $\hat{Y}_{\mathrm{NEW},i}$ is our predictor computed from data observed on two fragments generated from model (i) (left) or model (ii) (right), when $n = 50$.

6·4. *Real data illustration*

In addition to the bone mineral density growth dataset used in Section 3·1 to illustrate the imputation of missing parts of curves, we applied our linear prediction algorithm to the Canadian weather stations data from Ramsay and Silverman (2005). This dataset consists of yearly average curves of temperature and precipitations at each of 35 weather stations. As in Ramsay and Silverman's (2005) chapter 15, we let Y_i be the logarithm of total annual precipitation at the ith weather station, and we took X_i to be the yearly average temperature curve at that station. Like them, our predictor of Y is to be based on the linear model at (16), which we approximate through our discrete Markov process as in Section 3·4.

Our goal is to compare the linear predictor based on the full data with our linear predictor applied to fragmented data. To do this, we extracted fragments of curves from the full curves by keeping, for the ith station, the values $X_i(t)$ for $t \in \mathcal{I}_i = [A_i, B_i]$, where $A_i = [U_i]$, with $U_i \sim U[1, 350]$ and $B_i = \min([A_i + V_i], 365)$, with $V_i \sim U[20, 60]$.

We generated $M = 100$ such fragmented samples of size $n = 35$. For each sample generated in this way, and for $i = 1, \ldots, n$, we computed the predictor $\hat{Y}_i$ of Y_i using the data pairs (X_j, Y_j), for $j \neq i$, where each X_j was either the full temperature curve of the jth weather station, or the fragmented version we extracted from it; we used the predictors described in Section 6·3. For each of the M samples, we measured the average prediction error $\mathrm{APE} = n^{-1} \sum_{i=1}^{n} |\hat{Y}_i - Y_i|$. For the predictor based on the full data curves, the APE was equal to 7.81. For the predictor based on fragments, the median and the interquartile range of the M values of the APE were equal to 12 and 3.07, respectively. Given the much reduced data curves we generated by keeping only a fragment of each curve, the degradation in performance of the predictor based on fragments is rather small, suggesting that our predictor performs well.

Acknowledgement

Research supported by the Australian Research Council. Peter passed away on 9 January 2016 in Melbourne, Australia. He was a wonderful man and an extraordinary researcher. He will be sorely missed. We downloaded the bone data at `http://fedc.wiwi.hu-berlin.de/quantnet/index.php?p=show&id=896`

Supplementary Material

Supplementary Material available at *Biometrika* online includes additional illustrations and calculations, an algorithm for practical implementation and the proof of the theorem.

References

Apanasovich, T. & Goldstein, E. (2008). On prediction error in functional linear regression. *Statist. Probab. Lett.*, **78**, 1807–1810.

Bachrach, L. K., Hastie, T. J., Wang, M. C., Narasimhan, B., & Marcus, R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black and Caucasian youth; a longitudinal study. *J. Clinical Endocrinology & Metabolism*, **84**, 4702–4712.

Cai, T. T. & Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.*, **34**, 2159–2179.

Carroll, R. J., & Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, **83**, 1184-–1186.

Crambes, C., Kneip, A. & Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.*, **37**, 35–72.

De Gooijer, J. G. & Zerom, D. (2003). On conditional density estimation. *Stat. Neerl.*, **57**, 159–176.

Delaigle, A. & Hall, P. (2013). Classification using censored functional data. *J. Amer. Statist. Assoc.*, **108**, 1269–1283.

Delaigle, A., Hall, P. & Apanasovich, T. V. (2009). Weighted least squares methods for prediction in the functional data linear model. *Electron. J. Stat.*, **3**, 865–885.

Delaigle, A., Hall, P. & Meister, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.*, **36**, 665–685.

Franzke, C., Crommelin, D., Fischer, A. & Majda, A. (2008). A hidden Markov model perspective on regimes and metastability in atmospheric flows. *J. Climate*, **21**, 1740–1757.

Ghosh, A. P. (2011). *Introduction to Diffusion Processes.* In *Wiley Encyclopedia of Operations Research and Management Science*, Edited by Cochran, J. J., Cox, L. A. Jr., Keskinocak, P., Kharoufeh, J. P. & Smith, J. C.

Goldberg, Y., Ritov, Y. & Mandelbaum, A. (2014) Predicting the continuation of a function with applications to call center data. *J. Statist. Plann. Inf.*, **147**, 53–65.

Hall, P. & Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc.* Ser. B, **68**, 109–126.

James, G., Hastie, T. J. & Sugar, C. (2000). Principal Component Models for Sparse Functional Data. *Biometrika*, **87**, 587–602.

Kraus, D. (2015). Components and completion of partially observed functional data. *J. R. Stat. Soc.* Ser. B, **77**, 777–801.

Lee, E. R. & Park, B. U. (2012). Sparse estimation in functional linear regression. *J. Multivariate Analysis*, **105**, 1–17.

Legg, B. J & Raupach, M. R. (1982). Markov-chain simulation of particle dispersion in inhomogeneous flows: The mean drift velocity induced by a gradient in Eulerian velocity variance. *Boundary-Layer Meteorology*, **24**, 3–13.

Liebl, D. (2013) Modeling and forecasting electricity spot prices: a functional data perspective. *Ann. Appl. Statist.*, **7**, 1562–1592.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, second edn. Springer, New York

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press.

Strachan, I. G., Hughes, N. P., Poonawala, M. H., Mason, J. W. & Tarassenko, L. (2009). Automated QT analysis that learns from cardiologist annotations. *Ann. Noninvasive Electrocardiology*, **14**, Supplement 1, S9–S21.

Yao, F., Müller, H. G. & Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.*, **100**, 577–590.

Yau, C., Papaspillopoulos, O., Roberts, G. O. & Holmes, C. (2010). Bayesian non-parametric hidden Markov models with applications in genomics. *J. Roy. Statist. Soc.* Ser. B, **73**, 37–57.