Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **00**, 1–17
https://doi.org/10.1093/jrsssb/qkae116

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series B Statistical Methodology

Original Article

Dynamic modelling of sparse longitudinal data and functional snippets with stochastic differential equations

Yidong Zhou and Hans-Georg Müller

Department of Statistics, University of California, Davis, CA 95616, USA

Address for correspondence: Hans-Georg Müller, Department of Statistics, University of California, Davis, CA 95616, USA. Email: hgmueller@ucdavis.edu

Abstract

Sparse functional/longitudinal data have attracted widespread interest due to the prevalence of such data in social and life sciences. A prominent scenario where such data are routinely encountered are accelerated longitudinal studies, where subjects are enrolled in the study at a random time and are only tracked for a short amount of time relative to the domain of interest. The statistical analysis of such functional snippets is challenging since information for far-off-diagonal regions of the covariance structure is missing. Our main methodological contribution is to address this challenge by bypassing covariance estimation and instead modelling the underlying process as the solution of a data-adaptive stochastic differential equation. Taking advantage of the interface between Gaussian functional data and stochastic differential equations makes it possible to efficiently reconstruct the target process by estimating its dynamic distribution. The proposed approach allows one to consistently recover forward sample paths from functional snippets at the subject level. We establish the existence and uniqueness of the solution to the proposed data-driven stochastic differential equation and derive rates of convergence for the corresponding estimators. The finite sample performance is demonstrated with simulation studies and functional snippets arising from a growth study and spinal bone mineral density data.

Keywords: accelerated longitudinal study, dynamic distribution, empirical dynamic, growth monitoring, sparse functional data

1 Introduction

Functional data are commonly viewed as i.i.d. samples of realizations of an underlying smooth stochastic process, which is typically observed at a discrete grid of time points. Such data are common and routinely arise in longitudinal studies. Functional data analysis has received much attention over recent decades; functional principal component analysis (Castro et al., 1986; K. Chen & Lei, 2015; Hall & Hosseini-Nasab, 2006; Kleffe, 1973) and functional regression (Hall & Horowitz, 2007; Ramsay & Silverman, 2005) have emerged as key tools. Detailed reviews can be found in Ramsay and Silverman (2005), Hsing and Eubank (2015), and Wang et al. (2016). One area where there are still important open questions concerns the impact of the study design on the analysis. We develop here a novel type of analysis for functional snippets, which correspond to very sparse sampling designs that arise often in accelerated longitudinal studies, by establishing a connection to stochastic differential equations (SDE).

From a general perspective, functional data are collected through various study designs, where one can differentiate between fully observed, densely and sparsely sampled functional data (Zhang & Wang, 2016). Fully observed functional data occur in continuous sensor signal recordings and dense designs when measurements at a large number of well-spaced time points are available. Sparse designs are characterized by the availability of only a small number of measurements. A common sparse design occurs when sparse time points are distributed over the entire domain for

Received: September 30, 2023. **Revised:** November 17, 2024. **Accepted:** November 23, 2024

each subject, with a smooth density that is strictly positive over the time domain where data are collected (Hall et al., 2006; Li & Hsing, 2010; Yao et al., 2005).

In this article, the focus is on a second type of sparse design that occurs in accelerated longitudinal studies (Galbraith et al., 2017), where subjects are enrolled in the study at a random time within the time domain and are only tracked for a limited amount of time relative to the domain of interest. Such accelerated longitudinal designs are appealing to practitioners in social and life sciences as they minimize the time and resources required to collect data for each subject, especially when data gathering is costly, intrusive or difficult. Formally, denoting the domain of interest by $\mathcal{T} = [a, b]$, the ith subject is only observed on a sub-interval $[A_i, B_i] \subset \mathcal{T}$, where $B_i - A_i \leq \eta(b - a)$ for all i and $\eta \in (0, 1)$ is a constant. When the constant η is much smaller than 1, these are functional snippets (Lin et al., 2021).

Partially observed functional data also arise in the form of functional fragments (Kneip & Liebl, 2020; Kraus, 2015; Liebl & Rameseder, 2019), where the constant η may approach 1. The presence of large fragments makes such functional fragments easier to handle since the design plot (Yao et al., 2005) is typically fully or nearly fully covered by the design points, thereby enabling the estimation of the covariance surface directly from the data. In contrast, all of the design points for functional snippets fall within a narrow band around the diagonal area, while the domain of interest is much larger than this band. It is therefore not possible to infer the covariance surface of the functional data with the usual non-parametric approach and this impedes the implementation of functional principal component analysis and all related methods. The only known solution is to impose additional and typically very strong and often unverifiable assumptions about the nature of the covariance. Such assumptions have been made to justify various forms of covariance completion that have included parametric, semiparametric and other approaches (see, e.g. Delaigle & Hall, 2016; Delaigle et al., 2021; Descary & Panaretos, 2019; Lin & Wang, 2022; Lin et al., 2021; Rice & Silverman, 1991).

For the modelling of time-dynamic systems, empirical dynamics for functional data (Müller & Yao, 2010) is an approach to recover the underlying dynamics from repeated observations of the trajectories that are generated by the dynamics, including a non-linear version (Verzelen et al., 2012). These approaches do not cover functional snippets. To the best of our knowledge, Dawson and Müller (2018) is the only existing dynamic approach aimed at the analysis of functional snippets, where the underlying dynamics are investigated through an autonomous differential equation for longitudinal quantile trajectories, requiring the underlying process to be monotonic (Abramson & Müller, 1994; Vittinghoff et al., 1994). This approach aims at estimating the conditional quantile trajectories given an initial condition, rather than the whole dynamic distribution which is our goal here.

Specifically, we aim to reconstruct the latent stochastic process that generates the observed functional snippets by recovering its time-evolving distributions, which we refer to as dynamic distribution. To overcome the challenge posed by snippets, we model the underlying process as the solution of a data-adaptive SDE. The dynamic distribution of the target process, containing all information about the underlying dynamics, is then estimated by stepwise forward integration. There is previous research (Comte & Genon-Catalot, 2020; Denis et al., 2021; Mohammadi et al., 2023) where functional data analysis has been utilized to infer SDEs, primarily focusing on the estimation of drift and diffusion coefficients with parametric components. Our approach does not follow these approaches and is entirely different, as our emphasis is the modelling of functional snippets and the recovery of the underlying stochastic process from such highly incomplete data. To accomplish this, we employ SDEs in a novel and fully non-parametric way.

The proposed SDE approach is not only new but in contrast to various covariance completion approaches is non-parametric and does not involve functional principal component analysis. The latter requires to recover the complete covariance surface, which is straightforward for dense designs (He et al., 2000), but for functional snippets in principle is impossible, unless one is prepared to impose strong assumptions on the global structure of the covariance surface that in general cannot be verified. The utility of the proposed approach for statistical practice is illustrated for growth and bone mineral density data in Section 6, where it is shown to aid in growth monitoring and more generally distinguishing individuals with abnormal development patterns. The proposed SDE approach also enables predictions for individuals with only one observation, where the individual-specific dynamic distribution far into the future can be predicted. The rate of convergence for the corresponding conditional distributions is derived in terms of the Wasserstein metric. The main assumption of the proposed approach is that the underlying process is Gaussian, which is a common assumption in functional data analysis.

The specific contributions of this article are, first, to provide an alternative perspective to characterize functional snippets using SDEs; second, to recover future distributions of individual subjects under minimal assumptions; third, to provide an approach that works for minimal snippets, where only two adjacent measurements may be available for each subject; fourth, to demonstrate existence and uniqueness of the solution of the data-adaptive SDE, along with the rate of convergence for the corresponding estimate; fifth, to illustrate the wide applicability of the proposed dynamic modelling approach with growth snippets from Nepalese children and for bone mineral density data.

The rest of this article is organized as follows. In Section 2, we introduce the proposed dynamic model, while Section 3 covers estimation procedures. Theoretical results are established in Section 4. Simulations and applications for a Nepal growth study data and spinal bone mineral density data are discussed in Sections 5 and 6, respectively. Finally, we conclude with a brief discussion in Section 7.

2 Learning dynamic distribution via stochastic differential equations

2.1 Stochastic differential equations and diffusion processes

A typical (Itô) SDE takes the form

$$\begin{cases} dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t, & t \in \mathcal{T}, \\ X_0 = x_0, \end{cases} \tag{1}$$

where $X_t = X(t)$ is a stochastic process on $(\Omega, \mathcal{F}, P)$, b and σ are the drift and diffusion coefficients, respectively, and B_t is a Brownian motion (also known as Wiener process). The initial value x_0 can be either deterministic or random, independent of the Brownian motion B_t. It is known that a unique solution of (1) exists if the Lipschitz condition

$$|b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leq C|x - y| \quad \text{for all } x, y \in \mathbb{R}, t \in \mathcal{T}. \tag{2}$$

and the linear growth condition

$$|b(t, x)| + |\sigma(t, x)| \leq C(1 + |x|) \quad \text{for all } x \in \mathbb{R}, t \in \mathcal{T}, \tag{3}$$

hold for some constant $C > 0$ (Øksendal, 2003, chapter 5.2). In fact, if coefficients b and σ satisfy the Lipschitz and linear growth conditions, then any solution X_t is a diffusion process on $\mathcal{T}$ with drift coefficient b and diffusion coefficient σ (Panik, 2017, p. 154). A diffusion process is a continuous-time Markov process that has continuous sample paths, which can be defined by specifying its first two moments together with the requirement that there are no instantaneous jumps over time. We can write the formulae for the drift and diffusion coefficients of a diffusion process in the following form:

$$b(t, x) = \lim_{s \to t^+} \frac{1}{s - t} E(X_s - X_t \mid X_t = x) \tag{4}$$

and

$$\sigma^2(t, x) = \lim_{s \to t^+} \frac{1}{s - t} E\{(X_s - X_t)^2 \mid X_t = x\}.$$

Note that the diffusion coefficient can be equivalently defined as

$$\sigma^2(t, x) = \lim_{s \to t^+} \frac{1}{s - t} \mathrm{Var}(X_s - X_t \mid X_t = x) \tag{5}$$

since

$$\begin{aligned} \mathrm{Var}(X_s - X_t \mid X_t = x) &= E\{(X_s - X_t)^2 \mid X_t = x\} - \{b(t, x)(s - t) + o(s - t)\}^2 \\ &= E\{(X_s - X_t)^2 \mid X_t = x\} + o(s - t). \end{aligned}$$

Here, $b(t, X_t)$ may be thought of as the instantaneous rate of change in the mean of the process given X_t; and $\sigma^2(t, X_t)$ can be viewed as the instantaneous rate of change of the squared fluctuations of the process given X_t (Kloeden & Platen, 1999, chapter 1.7). For a more detailed treatment, we refer to Section S.1.1 of the online supplementary material.

Diffusion processes originate in physics as mathematical models of the motions of individual molecules undergoing random collisions with other molecules (Pavliotis, 2014). Brownian motion is the simplest and most pervasive diffusion process. Several more complex processes can be constructed from standard Brownian motion, including the Brownian bridge, geometric Brownian motion and the Ornstein–Uhlenbeck process (Uhlenbeck & Ornstein, 1930). When drift and diffusion components of a diffusion process are moderately smooth functions, its transition density satisfies partial differential equations, i.e. the Kolmogorov forward (Fokker–Planck) and the Kolmogorov backward equation.

2.2 Alternative formulation of stochastic differential equations

We assume that the observed snippets are generated by an underlying stochastic process X_t defined on some compact domain $\mathcal{T} \subset \mathbb{R}$ with mean function $\mu(t) = E(X_t)$ and covariance function $\Sigma(s, t) = \mathrm{Cov}(X_s, X_t)$. Without loss of generality, $\mathcal{T}$ is taken to be $[0, 1]$ in the sequel. Suppose $\{X_{t,1}, \ldots, X_{t,n}\}$ is an independent random sample of X_t, where n is the sample size. In practice, each $X_{t,i}$ is only recorded at subject-specific N_i time points $T_{i1}, \ldots, T_{iN_i}$ and the observed data are $Y_{ij} = X_{T_{ij},i}$ for $j = 1, \ldots, N_i$. We assume that $N_i > 1$ for the subjects used to learn the SDE as subjects with only one measurement do not carry information about the local covariance structure. The snippet nature is reflected by the restriction that $|T_{ij} - T_{ik}| \leq \delta$ for all i, j, k, and some constant $\delta \in (0, 1)$. The focus of this article is to infer stochastic dynamics of the underlying stochastic process X_t from data pairs (T_{ij}, Y_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, N_i$. Specifically, we are interested in estimating sample paths of X_t starting from some initial time point given a starting value. The proposed approach borrows information from subjects with at least two measurements in order to recover the subject-specific dynamic distribution far into the future for each participant, even for those with a single measurement; their data do not contribute to the model fitting step. To illustrate the effectiveness of the proposed method for snippets with minimal numbers of observations, we consider the case $N_i = 2$ for simplicity. However, the proposed method is not restricted to this case and utilizes all data when more than two measurements are available for some or all subjects. Additional details are in Section 6.1 and Section S.5.3 of the online supplementary material.

The underlying stochastic process X_t is assumed to follow a general SDE as per (1). In real data applications, the drift and diffusion coefficients in (1) are typically unknown. To recover the underlying dynamics of X_t, instead of attempting to directly estimate the drift and diffusion terms, which is challenging for functional snippet data, we plug in representations (4) and (5) of drift and diffusion coefficients to obtain the following alternative version of the SDE,

$$\begin{cases} dX_t = \dfrac{\partial}{\partial s} E(X_s \mid X_t)\Big|_{s=t} \cdot dt + \left\{ \dfrac{\partial}{\partial s} \mathrm{Var}(X_s \mid X_t)\Big|_{s=t} \right\}^{1/2} \cdot dB_t, \quad t \in \mathcal{T}, \\ X_0 = x_0. \end{cases} \tag{6}$$

Note that s is taken to be strictly greater than t when calculating the partial derivatives of $E(X_s \mid X_t)$ and $\mathrm{Var}(X_s \mid X_t)$ with respect to s, in which case the diffusion coefficient is well-defined and not equal to 0. The SDE (6) is the key tool to obtain sample paths of X_t given an initial condition by means of a recursive procedure, where under Gaussianity at each step the distribution of X_t is constructed using the estimation of conditional means $E(X_s \mid X_t)$ and conditional variances $\mathrm{Var}(X_s \mid X_t)$.

Examples of the SDE (6) include Brownian motion, the Ho–Lee model (Ho & Lee, 1986) and the Ornstein–Uhlenbeck process (Uhlenbeck & Ornstein, 1930), among others. Different models postulate different forms of b and σ. The Brownian motion B_t, with extensive applications in physics and electrical engineering, is a special case with zero drift and unit diffusion. The Ho–Lee model $dX_t = g(t)dt + \sigma dB_t$ where $\sigma > 0$ and g is a deterministic function of time is a stochastic interest rate model widely used for the pricing of bond options and to model future interest rates. The Ornstein–Uhlenbeck process $dX_t = -\theta X_t dt + \sigma dB_t$ with $\theta > 0, \sigma > 0$ is often used to describe mean-reverting phenomena in the physical sciences, evolutionary biology and finance, where

the coefficient θ characterizes the restoring force towards the mean and σ the degree of volatility around the mean.

3 Estimation

3.1 Simulating sample paths

To estimate sample paths of X_t from functional snippets, given an initial condition, it is instructive to rewrite the SDE in (6) as

$$\lim_{s\to t^+}(X_s - X_t) = \lim_{s\to t^+}\left\{\frac{E(X_s \mid X_t) - E(X_t \mid X_t)}{s-t}(s-t) + \left\{\frac{\mathrm{Var}(X_s \mid X_t) - \mathrm{Var}(X_t \mid X_t)}{s-t}\right\}^{1/2}(B_s - B_t)\right\}$$

with initial condition $X_0 = x_0$. The above formula gives rise to a method to simulate the continuous-time process X_t at a set of discrete time points given an initial condition. Consider a pre-specified equidistant time grid $0 \le t_0 < t_1 < \cdots < t_{K-1} < t_K \le 1$ with the common time spacing Δ. Denoting the initial value of X_t at t_0 by X_0 and the simulation of X_t at t_k by X_k for $k = 1, \ldots, K$, we simulate the continuous-time process X_t at the discrete time points t_k, $k = 1, \ldots, K$, given an initial condition $X_0 = x_0$, by the recursion

$$X_k - X_{k-1} = \frac{E(X_k \mid X_{k-1}) - E(X_{k-1} \mid X_{k-1})}{\Delta}\Delta + \left\{\frac{\mathrm{Var}(X_k \mid X_{k-1}) - \mathrm{Var}(X_{k-1} \mid X_{k-1})}{\Delta}\right\}^{1/2}(B_{t_k} - B_{t_{k-1}}).$$

Observing that $E(X_{k-1} \mid X_{k-1}) = X_{k-1}$, $\mathrm{Var}(X_{k-1} \mid X_{k-1}) = 0$, and $(B_{t_k} - B_{t_{k-1}})/\sqrt{\Delta} \sim N(0, 1)$, the above recursion reduces to

$$X_k = E(X_k \mid X_{k-1}) + \{\mathrm{Var}(X_k \mid X_{k-1})\}^{1/2} W_k, \quad X_0 = x_0, \tag{7}$$

where $W_k \sim N(0, 1)$ are independent for $k = 1, \ldots, K$.

We emphasize that under Gaussianity, the recursion in (7) generates an exact simulation (Glasserman, 2004) of X_t at $t_1, \ldots, t_K$ in the sense that the X_k it produces follows the same distribution of the process X_t at t_k for all $k = 1, \ldots, K$; see Lemma 1 in Section 4. Classical simulation methods for SDEs, such as the Euler–Maruyama method and the Milstein method (Kloeden & Platen, 1999), in general, introduce discretization error at $t_1, \ldots, t_K$, because the increments do not have exactly the right mean and variance. To simulate X_t using recursion (7), there is hence no need to consider increasing numbers of discrete time points K. In practice and particularly for the case of accelerated longitudinal studies, a good rule of thumb is to set the time spacing Δ as the scheduled (as opposed to actual) visit spacing for each subject. The number of discrete time points K to simulate X_t is then determined by the time spacing Δ and the time interval of interest; see Section 6 for the selection of time grids in real data applications.

To estimate sample paths of the process X_t, one needs to iteratively generate a random sample from the Gaussian distribution $N\{E(X_k \mid X_{k-1}), \mathrm{Var}(X_k \mid X_{k-1})\}$ to simulate X_t at t_k for $k = 1, \ldots, K$ as per (7). In practice, both the conditional mean $E(X_k \mid X_{k-1})$ and conditional variance $\mathrm{Var}(X_k \mid X_{k-1})$ are unknown and thus need to be estimated.

3.2 Estimation of conditional mean and conditional variance

Note that the information contained in $X_{k-1} = X_{t_{k-1}}$ is twofold and includes X_{k-1} itself as well as the time index t_{k-1}. One can then formulate the estimation of the conditional mean $E(X_k \mid X_{k-1})$ and conditional variance $\mathrm{Var}(X_k \mid X_{k-1})$ as a regression problem with response X_k and predictor $(X_{k-1}, t_{k-1})^{\mathrm{T}}$.

Recall that each subject is observed at least twice, at time points T_{i1} and T_{i2}, where the corresponding measurements are Y_{i1} and Y_{i2}. Let $Z_i = (Y_{i1}, T_{i1})^{\mathrm{T}}$ and with a slight abuse of notation

Algorithm 1 Estimating sample paths

Input: Training data $\{(Z_i, Y_i)\}_{i=1}^n$, initial condition $Z_0 = (x_0, t_0)^T$, and time discretization $\{t_k, k = 0, \ldots, K\}$.

Output: $(\hat{X}_1, \ldots, \hat{X}_K)^T$.

1 **for** $k = 1, \ldots, K$ **do**

2 Estimate the conditional mean $E(X_k \mid X_{k-1})$ and conditional variance $\mathrm{Var}(X_k \mid X_{k-1})$ by $\hat{m}(\hat{Z}_{k-1})$ and $\hat{v}^2(\hat{Z}_{k-1})$, respectively;

3 Draw a random sample $\hat{X}_k$ from $N\{\hat{m}(Z_{k-1}), \hat{v}^2(Z_{k-1})\}$;

4 $\hat{Z}_k \leftarrow (\hat{X}_k, t_k)^T$;

5 **end**

set $Y_i = Y_{i2}$ for $i = 1, \ldots, n$. Viewing the $\{(Z_i, Y_i)\}_{i=1}^n$ as n i.i.d. realizations of the pair of random variables (Z, Y), consider the regression model

$$Y_i = m(Z_i) + v(Z_i)\epsilon_i, \tag{8}$$

where $m(z) = E(Y \mid Z = z)$ and $v^2(z) = \mathrm{Var}(Y \mid Z = z)$ are respectively the conditional mean function and conditional variance function. The error term ϵ_i satisfies $E(\epsilon_i \mid Z_i) = 0$ and $\mathrm{Var}(\epsilon_i \mid Z_i) = 1$. The estimation of both conditional mean and conditional variance using parametric or non-parametric regression methods has been thoroughly studied. For the estimation of conditional variance we adopt the well-known approach of fitting a regression model for the squared residuals $\{Y_i - \hat{m}(Z_i)\}^2$ as responses and Z_i as predictors (Fan & Yao, 1998); see Section S.2 of the online supplementary material for more details.

Based on the regression model (8), the recursion for simulating sample paths in (7) simplifies to

$$X_k = m(Z_{k-1}) + v(Z_{k-1})W_k, \quad X_0 = x_0, \tag{9}$$

where $Z_{k-1} = (X_{k-1}, t_{k-1})^T$ for $k = 1, \ldots, K$. With estimates of the conditional mean function $\hat{m}(\cdot)$ and conditional variance function $\hat{v}^2(\cdot)$ in hand, we estimate the sample path of the underlying process X_t at $t_1, \ldots, t_K$, given an initial condition $X_0 = x_0$, by the following recursive procedure,

$$\begin{aligned} \hat{X}_1 &= \hat{m}(Z_0) + \hat{v}(Z_0)W_1, \\ \hat{X}_k &= \hat{m}(\hat{Z}_{k-1}) + \hat{v}(\hat{Z}_{k-1})W_k, \quad k = 2, \ldots, K, \end{aligned} \tag{10}$$

where $Z_0 = (x_0, t_0)^T$ and $\hat{Z}_{k-1} = (\hat{X}_{k-1}, t_{k-1})^T$ for $k = 2, \ldots, K$; see Algorithm 1.

If one has no prior knowledge about the conditional mean and conditional variance structure, which is often the case in real data applications, it may be preferable to adopt non-parametric approaches that are more flexible than say multiple linear regression, while incurring a lower rate of convergence.

4 Theoretical results

We establish existence and uniqueness of the solution to the proposed SDE as per (6) and the rate of convergence for the estimated sample paths. Existence and uniqueness follows from the Gaussianity of the process X_t; for every finite set of time points $t_1, \ldots, t_k$ in $\mathcal{T}$, $(X_{t_1}, \ldots, X_{t_k})^T$ are jointly Gaussian. The key step is to express the conditional mean $E(X_s \mid X_t)$ and conditional variance $\mathrm{Var}(X_s \mid X_t)$ in terms of the mean and covariance functions of X_t, whence drift and diffusion coefficients in (6) are seen to satisfy the Lipschitz and linear growth conditions as per (2) and (3). Specifically,

$$E(X_s \mid X_t) = \mu(s) + \Sigma(s, t)\Sigma^{-1}(t, t)\{X_t - \mu(t)\}, \tag{11}$$

$$\mathrm{Var}(X_s \mid X_t) = \Sigma(s, s) - \Sigma(s, t)\Sigma^{-1}(t, t)\Sigma(t, s). \tag{12}$$

If X_t is non-Gaussian, as long as a unique solution exists, the rate of convergence for the estimated sample path can be similarly derived by assuming Lipschitz continuity for the conditional mean function $m(\cdot)$ and conditional variance function $v^2(\cdot)$; see Lemma 2.

To show that the drift and diffusion coefficients in (6) satisfy the Lipschitz and linear growth conditions as per (2) and (3), we require the following conditions.

(A1) The mean function $\mu(t) = E(X_t)$ is continuously differentiable on $\mathcal{T}$.
(A2) The covariance function $\Sigma(s, t) = \mathrm{Cov}(X_s, X_t)$ is continuously differentiable in the lower triangular region $\{(s, t) : s \geq t, s, t \in \mathcal{T}\}$. Equivalently, the two partial derivative functions

$$\Sigma'_s(s, t) = \frac{\partial \Sigma(s, t)}{\partial s}, \quad \Sigma'_t(s, t) = \frac{\partial \Sigma(s, t)}{\partial t}$$

exist and are continuous for every $s, t \in \mathcal{T}$ and $s \geq t$.

Conditions (A1) and (A2) are regularity conditions on the process X_t, where the latter implies that $\Sigma(s, t)$ is continuously differentiable in the upper triangular region $\{(s, t) : s \leq t, s, t \in \mathcal{T}\}$ but may not be differentiable across the diagonal $s = t$, as for example is the case for Brownian motion. It is easy to verify that all examples of processes in Section 2.2 satisfy Conditions (A1) and (A2); see Section S.1 of the online supplementary material.

Theorem 1 If the stochastic process X_t is Gaussian, satisfies Conditions (A1), (A2), and the initial value x_0 is a random variable independent of the σ-algebra $\mathcal{F}_\infty$ generated by $\{B_s, s \geq 0\}$ with $E(x_0^2) < \infty$, then the stochastic differential equation (6) has a pathwise unique strong solution

$$X_t = x_0 + \int_0^t \frac{\partial}{\partial r} E(X_r \mid X_s)\Big|_{r=s} ds + \int_0^t \left\{ \frac{\partial}{\partial r} \mathrm{Var}(X_r \mid X_s)\Big|_{r=s} \right\}^{1/2} dB_s, \quad t \in \mathcal{T}$$

with the property that

$$X_t \text{ is adapted to the filtration } \mathcal{F}_t^{x_0} \text{ generated by } x_0 \text{ and } \{B_s, s \in [0, t]\} \tag{13}$$

and

$$\sup_{t \in \mathcal{T}} E(X_t^2) < \infty. \tag{14}$$

All proofs are given in Section S.3 of the online supplementary material. The uniqueness of the solution means that if X_t and Y_t are two processes satisfying (6), (13), and (14) then

$$X_t = Y_t \quad \text{for all } t \in \mathcal{T} \quad \text{a.s.}$$

The solution X_t in Theorem 1 is a strong solution because the version B_t of Brownian motion is given in advance and the solution X_t is $\mathcal{F}_t^{x_0}$-adapted. The Gaussianity implies that X_t must be governed by a narrow-sense linear SDE (Kloeden & Platen, 1999), where the drift coefficient is $b(t, X_t) = a(t)X_t + c(t)$ and the diffusion coefficient is additive, i.e. $\sigma(t, X_t) = \sigma(t)$. The drift and diffusion coefficients in (6) under Gaussianity are

$$b(t, X_t) = \mu'(t) + \Sigma'_s(s, t)\big|_{s=t} \Sigma^{-1}(t, t)\{X_t - \mu(t)\},$$
$$\sigma(t, X_t) = \left\{ \Sigma'(t, t) - 2\Sigma'_s(s, t)\big|_{s=t} \right\}^{1/2},$$

indicating the SDE (6) is narrow-sense linear. The general solution of a linear SDE can be found explicitly. Specifically, if X_t is Gaussian, as a solution of (6) it is of the form

$$X_t = \Phi(t)\Big\{x_0 + \textstyle\int_0^t c(s)\Phi^{-1}(s)ds + \int_0^t \sigma(s)\Phi^{-1}(s)dB_s\Big\},$$

with $a(t) = \Sigma_s'(s,t)|_{s=t}\Sigma^{-1}(t,t)$, $c(t) = \mu'(t) - \Sigma_s'(s,t)|_{s=t}\Sigma^{-1}(t,t)\mu(t)$ and $\Phi(t) = e^{\int_0^t a(s)ds}$.

An important feature of the recursion in (7) is that it generates an exact simulation of X_t at $t_1, \ldots, t_K$ (Glasserman, 2004) if X_t is a Gaussian process.

Lemma 1 If the stochastic process X_t is Gaussian, then the recursion in (7) generates an exact simulation of the stochastic process X_t at $t_1, \ldots, t_K$ in the sense that the distribution of the $X_1, \ldots, X_K$ it produces is precisely that of the continuous-time process X_t at time points $t_1, \ldots, t_K$.

Lemma 1 ensures that under Gaussianity the discretization error does not affect the rate of convergence of estimated sample paths. To study the asymptotics of estimated sample paths (7), we investigate the rate of convergence of $\hat{X}_K$; the same rate then applies for $\hat{X}_k$ for any $k \le K$. The proof relies on a recursion for $|\hat{X}_k - X_k|$ for increasing k, using the Lipschitz continuity of the conditional mean function $m(\cdot)$ and of the conditional variance function $v^2(\cdot)$. We require the following conditions regarding the variance function $\Sigma(t,t)$.

(B1) The variance function $\Sigma(t,t)$ is strictly positive on the half-open interval $(0, 1]$.

Condition (B1) can be expected to be satisfied in real data applications; all example processes discussed in Section 2.2 satisfy this condition, see Section S.1 of the online supplementary material. Under Gaussianity, the conditional mean and conditional variance in recursion (9) becomes

$$m(Z_{k-1}) = \mu(t_k) + \Sigma(t_k, t_{k-1})\Sigma^{-1}(t_{k-1}, t_{k-1})\{X_{k-1} - \mu(t_{k-1}))\},$$
$$v^2(Z_{k-1}) = \Sigma(t_k, t_k) - \Sigma(t_k, t_{k-1})\Sigma^{-1}(t_{k-1}, t_{k-1})\Sigma(t_{k-1}, t_k),$$

where $Z_{k-1} = (X_{k-1}, t_{k-1})^{\mathrm{T}}$ and $t_k = t_{k-1} + \Delta$ denotes the discrete time points used to simulate the sample path of the underlying process X_t.

Lemma 2 If the stochastic process X_t is Gaussian and satisfies (A1), (A2), and (B1), then for $k = 2, \ldots, K$ the conditional mean and conditional variance in recursion (9) satisfy

$$|m(\hat{Z}_{k-1}) - m(Z_{k-1})| \le L|\hat{X}_{k-1} - X_{k-1}|,$$
$$|v(\hat{Z}_{k-1}) - v(Z_{k-1})| = 0,$$

where $L = \max_{t\in\{t_1,\ldots,t_{K-1}\}} |\Sigma(t+\Delta, t)\Sigma^{-1}(t,t)|$ and $Z_{k-1} = (X_{k-1}, t_{k-1})^{\mathrm{T}}$, $\hat{Z}_{k-1} = (\hat{X}_{k-1}, t_{k-1})^{\mathrm{T}}$.

Lemma 2 imposes a limit on the growth of the sequence $|\hat{X}_k - X_k|$, whence one can bound $|\hat{X}_K - X_K|$ by recursion. Lemma 2 holds for all example processes discussed in Section 2.2 with Lipschitz constant $L = 1$; see Section S.1 of the online supplementary material.

To obtain the rate of convergence for the estimated sample path, one also needs to examine the asymptotic behaviour of the conditional mean function estimate $\hat{m}(\cdot)$ and the conditional variance function estimate $\hat{v}^2(\cdot)$. Assume one has results for any fixed $z \in \mathbb{R}\times\mathcal{T}$ of the type

$$[E\{|\hat{m}(z) - m(z)|^2\}]^{1/2} = O(\alpha_n), \quad [E\{|\hat{v}^2(z) - v^2(z)|^2\}]^{1/2} = O(\beta_n). \tag{15}$$

Adopting the residual-based estimator as described in Section S.2 of the online supplementary material to estimate the conditional variance function $v^2(\cdot)$, it is well-known that the estimation

of the conditional mean function $m(\cdot)$ has no influence on the estimation of $v^2(\cdot)$ (Fan & Yao, 1998). Then $\beta_n = \alpha_n$ if the same regression method is used to estimate $m(\cdot)$ and $v^2(\cdot)$. When multiple linear regression applies, $\alpha_n = \beta_n = n^{-1/2}$, while $\alpha_n = \beta_n = n^{-1/3}$ for local linear regression.

Theorem 2 If the stochastic process X_t is Gaussian and satisfies (A1), (A2), and (B1), then for the estimated sample path of the SDE (6) as defined in (10),

$$\{E(|\hat{X}_K - X_K|^2)\}^{1/2} = O(\alpha_n + \beta_n),$$

where α_n and β_n are the rates of convergence for the conditional mean function estimate $\hat{m}(\cdot)$ and conditional variance function estimate $\hat{v}^2(\cdot)$ as per (15).

Theorem 2 implies that $\hat{X}_K$ converges to X_K in the sense that both mean and variance converge to their respective targets, i.e.

$$|E(\hat{X}_K) - E(X_K)| = O(\alpha_n + \beta_n), \quad |\mathrm{Var}(\hat{X}_K) - \mathrm{Var}(X_K)| = O(\alpha_n^2 + \beta_n^2).$$

Note that this convergence holds uniformly over k, thereby establishing the pathwise convergence of the estimated sample path to the true process.

Writing $\mathcal{L}(X_K)$, $\mathcal{L}(\hat{X}_K)$ for the distributions of X_K and of the corresponding estimator $\hat{X}_K$, respectively, we aim to quantify the discrepancy between $\mathcal{L}(\hat{X}_K)$ and $\mathcal{L}(X_K)$ as a measure of the performance of the estimator. The strong convergence results obtained in Theorem 2 can be used to obtain the rate of convergence of the 2-Wasserstein distance (Villani, 2009) $d_W\{\mathcal{L}(\hat{X}_K), \mathcal{L}(X_K)\}$, where the 2-Wasserstein distance between two probability measures ν_1, ν_2 on $\mathbb{R}$ is $d_W^2(\nu_1, \nu_2) = \int_0^1 \{F_1^{-1}(p) - F_2^{-1}(p)\}^2 dp$, with F_1^{-1} and F_2^{-1} denoting the quantile functions of ν_1, ν_2, respectively. If ν_1 and ν_2 are one-dimensional Gaussians with means and variances (m_1, σ_1^2) and (m_2, σ_2^2) then $d_W^2(\nu_1, \nu_2) = (m_1 - m_2)^2 + (\sigma_1 - \sigma_2)^2$. For the Wasserstein rate of convergence we obtain

Corollary 1 Under the conditions of Theorem 2, the distribution of the estimated sample path as per (10) satisfies

$$d_W\{\mathcal{L}(\hat{X}_K), \mathcal{L}(X_K)\} = O(\alpha_n + \beta_n),$$

where α_n and β_n are the rates of convergence for the conditional mean function estimate $\hat{m}(\cdot)$ and conditional variance function estimate $\hat{v}^2(\cdot)$ as per (15).

So far, we have assumed that snippets are observed without measurement errors, which applies to situations such as longitudinal growth curves, where anthropometric measurements are often considered to be error-free. Applications to growth curves are highlighted in Section 6.1 and Section S.5.1 of the online supplementary material using growth curve data for the Nepal and Berkeley growth studies. The presence of measurement errors will lead to an errors-in-variables scenario (Griliches & Hausman, 1986), which will be discussed in Section S.4 of the online supplementary material. including theoretical analysis that characterizes the impact of measurement errors on the asymptotic behaviour of the estimated sample paths. In Section 5.2, we demonstrate that the proposed approach is quite robust in the presence of measurement errors. If one nevertheless would like to further address bias caused by measurement errors, this will require adopting some of the available measurement error correction techniques (Carroll et al., 2006; Cook & Stefanski, 1994).

5 Finite sample performance

5.1 Implementation details

The proposed dynamic modelling approach is straightforward to implement, as outlined in Algorithm 1. The regression model (8) involves only a two-dimensional predictor, resulting in a time complexity of $O(n)$ for training. Consequently, Algorithm 1 also runs in $O(n)$ time, given

that calculating the conditional mean and conditional variance takes two steps and K is fixed. For generating M sample paths, the time complexity is $O(Mn)$, making it linear with respect to the sample size. This computational efficiency makes the proposed approach highly suitable for large datasets. The algorithm has been implemented in R and is available on GitHub at https://github.com/yidongzhou/Dynamic-Modeling-of-Functional-Snippets.

Furthermore, the dynamic modelling approach inherently provides uncertainty quantification for the estimated sample paths. Practically, one can repeat the recursive process described in (10) M times for a sufficiently large M, such as $M = 1{,}000$. With these M simulated sample paths in hand, an empirical $1 - \alpha$ (pointwise) confidence band for the underlying process can be calculated. This method is demonstrated in Section 6.1 for identifying developmental delays in children's growth and is validated through simulations in Section S.5.5 of the online supplementary material.

Note that non-parametric regression models, such as local linear regression, rely on two bandwidths h_1 and h_2 for estimating the conditional mean and conditional variance, respectively, as defined in (8). While h_1 can be selected via cross-validation, the bias in the squared residuals $\{Y_i - \hat{m}(Z_i)\}^2$ makes cross-validation infeasible for choosing h_2. In our implementation, we choose $h_2 = h_1$, where we select h_1 by cross-validation for conditional mean estimation, minimizing $\mathrm{CV}(h) = \sum_{l=1}^{n} \{Y_i - \hat{m}_h^{(-l)}(Z_l)\}^2$. Here $\hat{m}_h^{(-l)}(\cdot)$ denotes the local linear regression estimate using bandwidth h based on the reduced sample $\{(Z_i, Y_i)\}_{i \neq l}$; users can choose to substitute alternative values for h_1, h_2.

5.2 Simulation studies

We demonstrate the utility of the proposed approach in recovering underlying dynamics from functional snippets across various scenarios. Existing work based exclusively on covariance completion is not directly comparable, as one of the advantages of the proposed approach is that it entirely bypasses covariance estimation and does not rely on functional principal component analysis. For comparative purposes, we estimate the covariance function using the covariance completion approach of Lin and Wang (2022), denoted as LW, via the mcfda package available at https://github.com/linulysses/mcfda. Subsequently, assuming Gaussianity, we derive the conditional mean and conditional variance using the estimated mean and covariance functions. Finally, we apply the recursive procedure outlined in (10) to reconstruct the underlying stochastic process.

We generate functional snippets from the Ho–Lee model and the Ornstein–Uhlenbeck process as discussed in Section 2.2, respectively. To obtain functional snippets, we first simulate the sample path of $X_{t,i}$ at a regular time grid $\{t_k\}_{k=0}^{K}$ with $t_k = k\delta$ and $K\delta = 1$ for each $i = 1, \ldots, n$. Denoting the simulated values for the n processes by $\{(X_{t_0,i}, X_{t_1,i}, \ldots, X_{t_K,i})^{\mathrm{T}}\}_{i=1}^{n}$, functional snippets are generated as $\{(X_{T_i,i}, X_{T_i+\delta,i})^{\mathrm{T}}\}_{i=1}^{n}$ where for each i, T_i is a time point randomly selected from the time grid $\{t_k\}_{k=0}^{K-1}$. Since both the Ho–Lee model and the Ornstein–Uhlenbeck process are narrow-sense linear SDEs, exact methods for simulating their paths are available by examining their explicit solutions (Glasserman, 2004). Specifically, for the Ho–Lee model $dX_t = g(t)dt + \sigma dB_t$, a simple recursive procedure for simulating values at $\{t_k\}_{k=0}^{K}$ is

$$X_{k+1} = X_k + \int_{t_k}^{t_{k+1}} g(s)ds + \sigma(t_{k+1} - t_k)^{1/2} W_k, \tag{16}$$

where $W_k \sim N(0, 1)$ are independent for all k and $X_0 = x_0$. Similarly for the Ornstein–Uhlenbeck process $dX_t = -\theta X_t dt + \sigma dB_t$, one can set

$$X_{k+1} = e^{-\theta(t_{k+1}-t_k)} X_k + \left\{ \frac{\sigma^2}{2\theta} (1 - e^{-2\theta(t_{k+1}-t_k)}) \right\}^{1/2} W_k. \tag{17}$$

The above procedures are exact in the sense that the joint distribution of the simulated values coincides with the joint distribution of the corresponding continuous-time process on the simulation grid. To investigate the effect of noise, we add independent errors to the generated functional snippets $\{(X_{T_i,i}, X_{T_i+\delta,i})^{\mathrm{T}}\}_{i=1}^{n}$. Specifically, we consider the contaminated functional snippets $\{(Y_{i1}, Y_{i2})^{\mathrm{T}}\}_{i=1}^{n}$, where $Y_{i1} = X_{T_i,i} + \varepsilon_{i1}$ and $Y_{i2} = X_{T_i+\delta,i} + \varepsilon_{i2}$ with $\varepsilon_{ij} \sim N(0, \nu^2)$ independently.

We examined the performance of the proposed approach across sample sizes $n = 50, 200, 1{,}000$ and noise levels $v = 0, 0.01, 0.1$. For each combination of sample size and noise level, the simulation was repeated $Q = 500$ times. The time interval was chosen as $[0, 1]$ and the time spacing was $\delta = 0.05$. In each simulation, the recursive procedure as per (10) was performed $M = 1{,}000$ times using the contaminated functional snippets $\{(Y_{i1}, Y_{i2})^{\mathrm{T}}\}_{i=1}^{n}$ with the initial condition $Z_0 = (0, 0)^{\mathrm{T}}$, from which $M = 1{,}000$ estimated sample paths evaluated at the time grid $\{t_k\}_{k=1}^{K}$ were obtained. We write $\{(\hat{X}_{t_1,l}, \ldots, \hat{X}_{t_K,l})^{\mathrm{T}}\}_{l=1}^{M}$ for the M estimated sample paths. For each l, the corresponding true sample path $(X_{t_1,l}, \ldots, X_{t_K,l})^{\mathrm{T}}$ was obtained using the recursive procedure in (16) or (17) with the same initial value and W_k. For each run of a particular sample size and noise level, the quality of the estimation was quantified by the root-mean-square error,

$$\mathrm{RMSE} = \left\{ \frac{1}{M} \sum_{l=1}^{M} (\hat{X}_{t_K,l} - X_{t_K,l})^2 \right\}^{1/2}.$$

We chose $g(t) = \cos(t)$, $\theta = \sigma = 1$ and the initial condition $X_0 = 0$ for the Ho–Lee model and the Ornstein–Uhlenbeck process, respectively and used multiple linear regression to estimate the conditional mean and conditional variance for both cases. The mean and standard deviation of RMSE across $Q = 500$ runs for various sample sizes and noise levels are summarized in Table 1. We observe that the mean RMSE of the proposed approach diminishes as the sample size increases, while the presence of noise has only a minor effect. In contrast, the mean RMSE of the covariance completion method is substantial even with a sample size of 1,000. This discrepancy may stem from the exceptionally sparse nature of this simulation scenario, where each process is observed within a narrow window of length 0.05, contrasting sharply with the broader interval of interest, which is $[0, 1]$. Consequently, the available information may be too limited for covariance completion methods to accurately reconstruct the entire covariance surface. As shown in Section S.5.3 of the online supplementary material, the covariance completion approach performs better but is still inferior to the proposed approach when more measurements are available.

To further illustrate the performance of the proposed dynamic modelling approach, we visualize the simulation results for the Ornstein–Uhlenbeck process with sample size $n = 200$ and noise level $v = 0.1$ in Figure 1, where $M = 100$ estimated sample paths are considered, along with the corresponding true sample paths. It is evident that the estimated sample paths recover the underlying stochastic dynamics from very sparse data, demonstrating that the proposed approach performs well.

Table 1. Mean and standard deviation (in parentheses) of root-mean-square errors across 500 runs for the Ho–Lee model and the Ornstein–Uhlenbeck process

Sample size	Noise level					
	DM			LW		
	0	0.01	0.1	0	0.01	0.1
Ho–Lee model						
50	0.92 (0.87)	0.93 (0.93)	1.21 (1.71)	1.08 (0.37)	1.07 (0.34)	1.11 (0.45)
200	0.39 (0.23)	0.38 (0.22)	0.54 (0.38)	0.91 (0.34)	0.92 (1.15)	0.89 (0.27)
1,000	0.17 (0.09)	0.17 (0.09)	0.27 (0.13)	0.89 (0.24)	0.89 (0.24)	0.88 (0.25)
Ornstein–Uhlenbeck process						
50	0.63 (0.65)	0.62 (0.84)	0.71 (1.81)	0.72 (0.23)	0.71 (0.2)	0.74 (0.28)
200	0.25 (0.15)	0.26 (0.16)	0.27 (0.13)	0.67 (0.17)	0.66 (0.21)	0.67 (0.23)
1,000	0.11 (0.06)	0.11 (0.06)	0.15 (0.05)	0.71 (0.09)	0.70 (0.10)	0.70 (0.11)

Note. Here, DM is the proposed dynamic modelling approach and LW the covariance completion approach of Lin and Wang (2022).

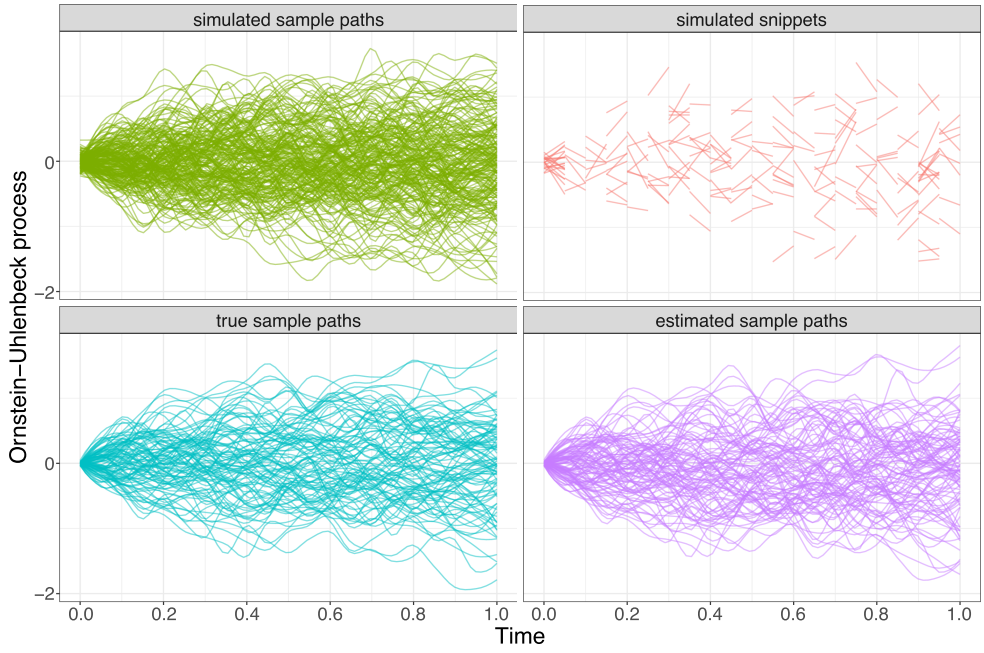

Figure 1. An illustration of the proposed dynamic modelling approach for a simulated Ornstein–Uhlenbeck process with sample size $n = 200$ and noise level $\nu = 0.1$. *Simulated sample paths*: fully observed sample paths contaminated with noise (upper left panel); *simulated snippets*: two consecutive measurements randomly extracted from each simulated sample path (upper right panel); *true underlying sample paths:* true underlying sample paths without noise (lower left panel); *estimated sample paths*: estimated sample paths obtained using the proposed dynamic modelling approach with the simulated snippets as input (lower right panel).

Further simulations are provided in Section S.5 of the online supplementary material, where we emulate the Berkeley growth study data, assess the resilience of the proposed method to departures from Gaussianity and explore denser scenarios with $N_i = 5$ measurements per subject. While the primary objective of the proposed approach is to reconstruct the dynamic distribution of the underlying process, we also investigate its capability for estimating the mean and variance function in Section S.5.6 of the online supplementary material.

6 Data applications

6.1 Nepal growth study data

Screening children's development status and monitoring height growth is essential for paediatric public health (K. Chen & Müller, 2012) and due to limited resources often must be based on incomplete data. We demonstrate the potential of the proposed dynamic modelling approach to characterize underlying growth patterns and reveal specific growth trends with snippet data from a Nepal growth study (West et al., 1997). This data set contains height measurements for 2,258 children from rural Nepal taken at five adjacent times points from birth to 76 months, spaced approximately four months apart. To facilitate the exploration of these data, we use the first 1,000 records, containing measurements for 107 males and 93 females. Due to missing data, the actual number of measurements per child ranges between 1 and 5. Children with at least two subsequent measurements are included in the analysis, while the rest are used for model validation. We applied the proposed method to females and males separately since female and male growth trends differ significantly, with females reaching puberty earlier than males.

So far the number of measurements per subject N_i has been assumed to be 2 for simplicity. For denser scenarios where $N_i > 2$, one could divide the N_i measurements into $N_i - 1$ pairs of contiguous measurements for each i and combine these pairs into a new sample for conditional mean and conditional variance estimation. This is useful to augment the sample size n especially if the sample

size n is relatively small, which is often the case in practice. We employ this strategy to make full use of the Nepal growth study data as well as the spinal bone mineral density data in the next subsection.

While in Section S.5.1 of the online supplementary material, we demonstrate the efficacy of the proposed dynamic modelling approach to recover the underlying growth dynamics from snippet data using Berkeley growth study data, we highlight here another important application of the proposed approach—growth monitoring. Given a child's initial development status, the proposed approach dynamically predicts child-specific growth patterns far into the future. As a child grows older and fresh measurements become available, one can screen the child's development by comparing newly available measurements with the predicted growth. We demonstrate this with a randomly selected female and male who have no contiguous measurements and hence are not included in the model fitting. Specifically, the selected female was measured only once at 4 months, while the male was measured at 12 and 20 months.

To obtain future growth patterns for these two children, the recursive procedure in (10) was implemented 100 times using the growth snippets with at least two measurements in a row to obtain 100 estimated growth curves, where local linear regression was adopted to estimate the conditional mean and conditional variance. The starting time is $t_0 = 4$ months old for the selected female and $t_0 = 12$ months old for the selected male, where the time spacing was set at $\Delta = 4$ months, corresponding to the intended measurement spacing of the Nepal growth study. The starting height X_0 is chosen as the initial height measurement, i.e. 52.9 cm and 63 cm for the selected female and male, respectively.

The estimated growth curves and the corresponding 5%, 50%, 95% percentile curves for these two individuals are shown in the right panels of Figure 2, while the observed snippet data for the Nepal growth study are in the left panels. Although the available information is very limited due to the snippet nature of the data, the proposed approach is capable of capturing relevant dynamics from the observed growth snippets and revealing future growth trends of the selected female and male children. For the selected male child, one additional height measurement is available at a later age (20 months). The newly available height measurement (65.1 cm) falls below the

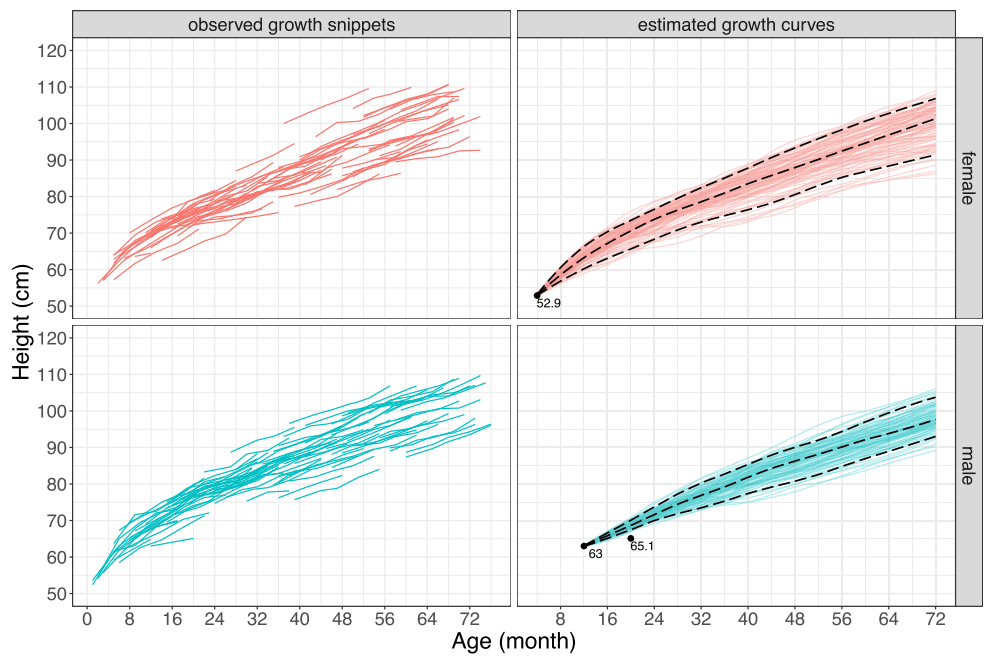

Figure 2. Observed growth snippets for females (upper left panel) and males (lower left panel), depicting the available data for the Nepal growth study, as well as predicted growth curves for a selected female (upper right panel) and a selected male child (lower right panel). The black dashed curves in the right panels indicate predicted 5%, 50%, and 95% percentiles and the available height measurements for the selected female and male are highlighted.

predicted 5% percentile curve, indicating that this child may be developmentally delayed and should be flagged for further follow-up.

6.2 Spinal bone mineral density data

In this study (Bachrach et al., 1999), 423 healthy individuals were longitudinally assessed for their spinal bone mineral density. These assessments were scheduled annually over four consecutive years. However, deviations from the planned visit schedules resulted in varying numbers of measurements available per individual, ranging from 1 to 4, and also varying time intervals between measurements. Such irregular and sparse observations have posed significant challenges in functional data analysis and garnered much attention in the field (Delaigle & Hall, 2016; Delaigle et al., 2021; James & Hastie, 2001; Lin & Wang, 2022; Lin et al., 2021). We included 153 females and 127 males with ages ranging from 8.8 to 26.2 years and featuring at least 2 measurements for model fitting, while the remaining subjects with only one measurement were used for model validation.

To infer individual-specific stochastic dynamics of spinal bone mineral density from the irregularly observed bone density snippets, we again randomly selected one female and one male for whom measurements were available at ages 10 and 9 years, respectively. The recursive procedure in (10) was run 100 times, resulting in 100 estimated bone density curves. Conditional mean and variance were obtained with local linear regression with cross-validation bandwidth selection. The starting age was chosen as $t_0 = 10$ years for the selected female and $t_0 = 9$ years for the selected male, with an end time of $t_K = 24$ years and 1-year time increments, corresponding to the scheduled measurement spacing of the data. The starting values of bone mineral density are 0.778 and 0.642 for the selected female and male, respectively, corresponding to their initial bone density measurements.

Figure 3 depicts the observed snippets and estimated bone density curves, along with 5%, 50%, 95% percentile curves, demonstrating that the proposed dynamic modelling approach is capable of handling the irregularity inherent in these data; see the right panel of Figure 3.

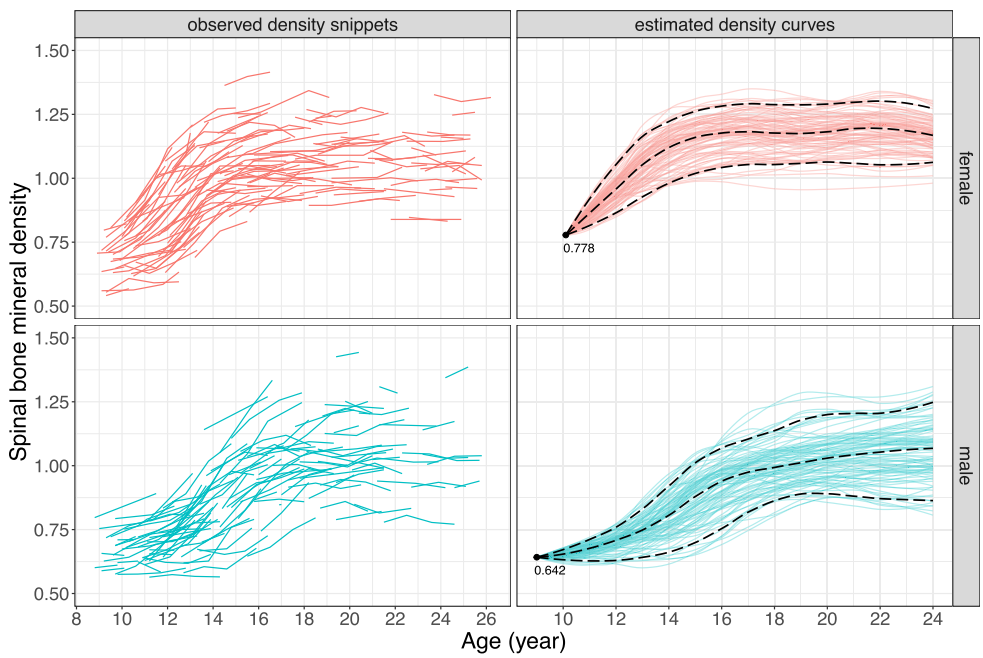

Figure 3. Observed snippets of mineral bone density for females (upper left panel) and males (lower left panel), as well as predicted bone density curves for a randomly selected female (upper right panel) and male (lower right panel), where the black dashed curves indicate 5%, 50%, and 95% percentiles. The available bone density measurements for the selected female and male are also highlighted.

Comparing the predicted bone density curves for the selected female and male, we find that for the female these reach a plateau at around 16 years, while for the male they level off at around 18 years. This finding is in agreement with the literature (Bachrach et al., 1999). Additionally, we applied the covariance completion approach for these data and present the findings in Section S.6 of the online supplementary material.

7 Discussion

In this article, we propose a flexible and robust approach to recover the dynamic distribution from functional snippets using SDE. The proposed framework circumvents the challenge of estimating covariance surfaces in the presence of missing data in the off-diagonal regions, leading to a consistent reconstruction of sample paths from observed snippets. Both theoretical analysis and numerical simulations support the effectiveness and utility of the proposed SDE approach.

Differential equations are extensively used across various scientific fields, including engineering, physics, and biomedical sciences. A significant portion of the literature on differential equations focuses on parameter estimation (Liang & Wu, 2008), with applications in time series (S. Chen et al., 2017) and functional data analysis (Denis et al., 2021). Another research avenue involves neural differential equations, where differential equations enhance the performance of neural networks (Yadav et al., 2015). Notable examples include neural ordinary differential equations (R. T. Chen et al., 2018) and neural SDEs (Jia & Benson, 2019; Oh et al., 2024). Additionally, SDEs are applied in generative modelling, such as score-based diffusion models (Song et al., 2020).

Complementing the existing literature, this article uses SDEs as a powerful tool to model functional snippets and to recover the underlying dynamics, addressing the challenge of minimal data availability for individual trajectories. The proposed tools also make it possible to assess forward dynamics by projecting trajectories into the future when only minimal snippet information or just one measurement is available for a given subject.

Acknowledgments

We thank the anonymous reviewers, the associate editor, and the editors for their constructive comments.

Conflicts of interest: None declared.

Funding

This research was partially supported by National Science Foundation (DMS-2014626) and National Institutes of Health (ECHO).

Data availability

The data supporting the findings of this study are available at https://github.com/yidongzhou/Dynamic-Modeling-of-Functional-Snippets. This repository also includes the source code and detailed instructions for implementing the proposed approach.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Abramson I., & Müller H.-G. (1994). Estimating direction fields in autonomous equation models, with an application to system identification from cross-sectional data. *Biometrika*, *81*(4), 663–672. https://doi.org/10.1093/biomet/81.4.663

Bachrach L. K., Hastie T., Wang M. -C., Narasimhan B., & Marcus R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black, and Caucasian youth: A longitudinal study. *Journal of Clinical Endocrinology & Metabolism*, *84*, 4702–4712. https://doi.org/10.1210/jcem.84.12.6182

Carroll R. J., Ruppert D., Stefanski L. A., & Crainiceanu C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. 2nd ed., Monographs on statistics and applied probability (*105*). Chapman & Hall/CRC.

Castro P. E., Lawton W. H., & Sylvestre E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *28*(4), 329–337. https://doi.org/10.2307/1268982

Chen K., & Lei J. (2015). Localized functional principal component analysis. *Journal of the American Statistical Association*, *110*(511), 1266–1275. https://doi.org/10.1080/01621459.2015.1016225

Chen K., & Müller H.-G. (2012). Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society: Series B*, *74*(1), 67–89. https://doi.org/10.1111/j.1467-9868.2011.01008.x

Chen R. T., Rubanova Y., Bettencourt J., & Duvenaud D. K. (2018). Neural ordinary differential equations. In *Advances in Neural Information Processing Systems* (Vol. 31).

Chen S., Shojaie A., & Witten D. M. (2017). Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association*, *112*(520), 1697–1707. https://doi.org/10.1080/01621459.2016.1229197

Comte F., & Genon-Catalot V. (2020). Nonparametric drift estimation for i.i.d. paths of stochastic differential equations. *Annals of Statistics*, *48*(6), 3336–3365. https://doi.org/10.1214/19-AOS1933

Cook J. R., & Stefanski L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, *89*, 1314–1328. https://doi.org/10.1080/01621459.1994.10476871

Dawson M., & Müller H.-G. (2018). Dynamic modeling of conditional quantile trajectories, with application to longitudinal snippet data. *Journal of the American Statistical Association*, *113*, 1612–1624. https://doi.org/10.1080/01621459.2017.1356321

Delaigle A., & Hall P. (2016). Approximating fragmented functional data by segments of Markov chains. *Biometrika*, *103*, 779–799. https://doi.org/10.1093/biomet/asw040

Delaigle A., Hall P., Huang W., & Kneip A. (2021). Estimating the covariance of fragmented and other related types of functional data. *Journal of the American Statistical Association*, *116*, 1383–1401. https://doi.org/10.1080/01621459.2020.1723597

Denis C., Dion-Blanc C., & Martinez M. (2021). A ridge estimator of the drift from discrete repeated observations of the solution of a stochastic differential equation. *Bernoulli*, *27*, 2675–2713. https://doi.org/10.3150/21-BEJ1327

Descary M.-H., & Panaretos V. M. (2019). Recovering covariance from functional fragments. *Biometrika*, *106*, 145–160. https://doi.org/10.1093/biomet/asy055

Fan J., & Yao Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, *85*, 645–660. https://doi.org/10.1093/biomet/85.3.645

Galbraith S., Bowden J., & Mander A. (2017). Accelerated longitudinal designs: An overview of modelling, power, costs and handling missing data. *Statistical Methods in Medical Research*, *26*, 374–398. https://doi.org/10.1177/0962280214547150

Glasserman P. (2004). *Monte Carlo methods in financial engineering*. Stochastic modelling and applied probability (Vol. *53*). Springer New York.

Griliches Z., & Hausman J. A. (1986). Errors in variables in panel data. *Journal of Econometrics*, *31*, 93–118. https://doi.org/10.1016/0304-4076(86)90058-8

Hall P., & Horowitz J. L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics*, *35*, 70–91. https://doi.org/10.1214/009053606000000957

Hall P., & Hosseini-Nasab M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B*, *68*(1), 109–126. https://doi.org/10.1111/j.1467-9868.2005.00535.x

Hall P., Müller H.-G., & Wang J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, *34*, 1493–1517. https://doi.org/10.1214/009053606000000272

He G., Müller H.-G., & Wang J.-L. (2000). Extending correlation and regression from multivariate to functional data. In M. L. Puri (Ed.), *Asymptotics in Statistics and Probability* (pp. 301–315). VSP International Science Publishers.

Ho T. S., & Lee S. -B. (1986). Term structure movements and pricing interest rate contingent claims. *Journal of Finance*, *41*(5), 1011–1029. https://doi.org/10.1111/jofi.1986.41.issue-5

Hsing T., & Eubank R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley series in probability and statistics (Vol. *997*). John Wiley & Sons.

James G. M., & Hastie T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *63*(3), 533–550. https://doi.org/10.1111/1467-9868.00297

Jia J., & Benson A. R. (2019). Neural jump stochastic differential equations. In *Advances in Neural Information Processing Systems* (Vol. 32).

Kleffe J. (1973). Principal components of random variables with values in a separable Hilbert space. *Mathematische Operationsforschung und Statistik*, *4*(5), 391–406. https://doi.org/10.1080/02331887308801137

Kloeden P. E., & Platen E. (1999). *Numerical solution of stochastic differential equations*. Stochastic modelling and applied probability (Vol. *23*). Springer Berlin.

Kneip A., & Liebl D. (2020). On the optimal reconstruction of partially observed functional data. *Annals of Statistics*, *48*, 1692–1717. https://doi.org/10.1214/19-AOS1864

Kraus D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *77*, 777–801. https://doi.org/10.1111/rssb.12087

Li Y., & Hsing T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*, *38*, 3321–3351. https://doi.org/10.1214/10-AOS813

Liang H., & Wu H. (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, *103*, 1570–1583. https://doi.org/10.1198/016214508000000797

Liebl D., & Rameseder S. (2019). Partially observed functional data: The case of systematically missing parts. *Computational Statistics & Data Analysis*, *131*, 104–115. https://doi.org/10.1016/j.csda.2018.08.011

Lin Z., & Wang J.-L. (2022). Mean and covariance estimation for functional snippets. *Journal of the American Statistical Association*, *117*, 348–360. https://doi.org/10.1080/01621459.2020.1777138

Lin Z., Wang J.-L., & Zhong Q. (2021). Basis expansions for functional snippets. *Biometrika*, *108*, 709–726. https://doi.org/10.1093/biomet/asaa088

Mohammadi N., Santoro L. V., & Panaretos V. M. (2023). Nonparametric estimation for SDE with sparsely sampled paths: An FDA perspective. *Stochastic Processes and their Applications*, *167*, 104239. https://doi.org/10.1016/j.spa.2023.104239

Müller H.-G., & Yao F. (2010). Empirical dynamics for longitudinal data. *Annals of Statistics*, *38*, 3458–3486. https://doi.org/10.1214/09-AOS786

Oh Y., Lim D., & Kim S. (2024). Stable neural stochastic differential equations in analyzing irregular time series data. In *International Conference on Learning Representations*.

Øksendal B. (2003). *Stochastic differential equations: An introduction with applications*. 6th ed., Vol. Universitext (*1*). Springer Berlin.

Panik M. J. (2017). *Stochastic differential equations: An introduction with applications in population dynamics modeling*. John Wiley & Sons.

Pavliotis G. A. (2014). *Stochastic processes and applications: Diffusion processes, the Fokker-Planck and Langevin equations*. Texts in applied mathematics (Vol. *60*). Springer New York.

Ramsay J. O., & Silverman B. W. (2005). *Functional data analysis*. 2nd ed., Vol. Springer series in statistics (*426*). Springer New York.

Rice J. A., & Silverman B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B*, *53*, 233–243. https://doi.org/10.1111/j.2517-6161.1991.tb01821.x

Song Y., Sohl-Dickstein J., Kingma D. P., Kumar A., Ermon S., & Poole B. (2020). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Uhlenbeck G. E., & Ornstein L. S. (1930). On the theory of the Brownian motion. *Physical Review*, *36*(5), 823–841. https://doi.org/10.1103/PhysRev.36.823

Verzelen N., Tao W., & Müller H.-G. (2012). Inferring stochastic dynamics from functional data. *Biometrika*, *99*(3), 533–550. https://doi.org/10.1093/biomet/ass015

Villani C (2009). *Optimal transport: Old and new*. Grundlehren der mathematischen Wissenschaften (Vol. *338*). Springer Berlin.

Vittinghoff E., Malani H. M., & Jewell N. P. (1994). Estimating patterns of CD4 lymphocyte decline using data from a prevalent cohort of HIV infected individuals. *Statistics in Medicine*, *13*(11), 1101–1118. https://doi.org/10.1002/sim.v13:11

Wang J.-L., Chiou J.-M., & Müller H.-G. (2016). Functional data analysis. *Annual Review of Statistics and its Application*, *3*, 257–295. https://doi.org/10.1146/statistics.2016.3.issue-1

West Jr K. P., LeClerq S. C., Shrestha S. R., Wu L. S. -F., Pradhan E. K., Khatry S. K., Katz J., Adhikari R., & Sommer A. (1997). Effects of vitamin A on growth of vitamin A-deficient children: Field studies in Nepal. *Journal of Nutrition*, *127*, 1957–1965. https://doi.org/10.1093/jn/127.10.1957

Yadav N., Yadav A., & Kumar M. (2015). *An introduction to neural network methods for differential equations*. SpringerBriefs in applied sciences and technology (Vol. *1*). Springer Dordrecht.

Yao F., Müller H.-G., & Wang J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, *100*, 577–590. https://doi.org/10.1198/016214504000001745

Zhang X., & Wang J.-L. (2016). From sparse to dense functional data and beyond. *Annals of Statistics*, *44*, 2281–2321. https://doi.org/10.1214/16-aos1446