Dynamic Modeling of Functional Snippets

Yidong Zhou, Hans-Georg Müller

Department of Statistics, University of California, Davis

Statistics in the Age of AI

Motivation

- In accelerated longitudinal studies, 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.
- Denoting the domain 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.
- Functional snippets: η is much smaller than 1.

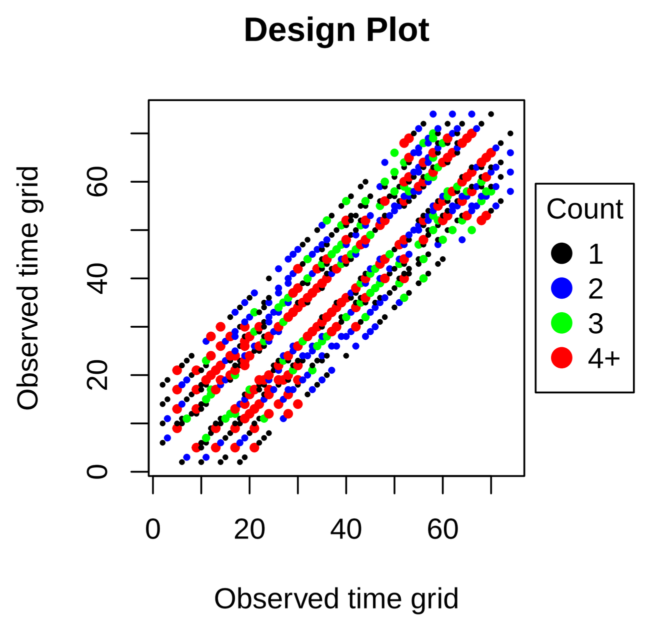


Figure 1. Design plots for females in the Nepal growth study data.

Alternative Formulation of Stochastic Differential Equations

- A typical (Itô) stochastic differential equation (SDE) takes the form

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

  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.
- The drift and diffusion coefficients can be viewed as the instantaneous rate of change in the mean and squared fluctuations of the process given X_t,

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

- Alternative formulation of the SDE,

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

- We then 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

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

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

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

  where $W_k \sim N(0,1)$ are independent for $k = 1, \ldots, K$.
- Under Gaussian assumption on the process X_t, the recursion in (1) generates an exact simulation of X_t at $t_1, \ldots, t_K$.
- To estimate sample paths of the process X_t, one needs to iteratively generate a random sample from $N\{E(X_k|X_{k-1}), \mathrm{Var}(X_k|X_{k-1})\}$ to simulate X_t at t_k for $k = 1, \ldots, K$. In practice, both the conditional mean $E(X_k|X_{k-1})$ and conditional variance $\mathrm{Var}(X_k|X_{k-1})$ are unknown.

Estimating Sample Paths From Functional Snippets

- Consider an underlying stochastic process X_t with mean function $\mu(t) = E(X_t)$, covariance function $\Sigma(s,t) = \mathrm{Cov}(X_s, X_t)$, and n realizations $\{X_{t,1}, \ldots, X_{t,n}\}$.
- We aim to infer stochastic dynamics of X_t from the observed snippets (T_{ij}, Y_{ij}), $i = 1, \ldots, n, j = 1, \ldots, N_i$, where $Y_{ij} = X_{T_{ij},i}$ and $|T_{ij} - T_{ik}| \leq \eta(b-a)$.
- To illustrate the effectiveness of the proposed method for snippets with minimal numbers of observations, we consider the case $N_i = 2$ for simplicity.
- Let $Z_i = (Y_{i1}, T_{i1})'$ and with a slight abuse of notation set $Y_i = Y_{i2}$ for $i = 1, \ldots, n$. Viewing the $\{(Z_i, Y_i)\}_{i=1}^n$ as i.i.d. realizations of the pair of random variables (Z, Y), consider the regression model

$$Y_i = m(Z_i) + v(Z_i)\epsilon_i,$$

  where $m(z) = E(Y|Z=z)$ and $v^2(z) = \mathrm{Var}(Y|Z=z)$ are respectively the conditional mean and conditional variance functions. The error term ϵ_i satisfies $E(\epsilon_i|Z_i) = 0$ and $\mathrm{Var}(\epsilon_i|Z_i) = 1$.
- With estimates of the conditional mean function $\hat{m}(\cdot)$ and the conditional variance function $\hat{v}^2(\cdot)$ in hand, the corresponding recursive procedure to obtain X_t at $t_1, \ldots, t_K$ is

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

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

Algorithm 1: Estimating sample paths of X_t from functional snippets

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

Output: $(\hat{X}_0, \ldots, \hat{X}_K)'$.

```
for k = 1, ..., K do
    Estimate the conditional mean E(X_k|X_{k-1}) and conditional
     variance Var(X_k|X_{k-1}) by m̂(Ẑ_{k-1}) and v̂²(Ẑ_{k-1}), respectively;
    Draw a random sample X̂_k from N{m̂(Z_{k-1}), v̂²(Z_{k-1})};
    Ẑ_k ← (X̂_k, t_k)';
end
```

Theorem 1

If the stochastic process X_t is Gaussian and satisfies certain regularity conditions, then for the estimated sample path of the SDE as defined in (2),

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

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 $\hat{v}^2(\cdot)$.

Remark

- *If X_t is non-Gaussian, the rate of convergence for the estimated sample path can be similarly derived by assuming Lipschitz continuity for $m(\cdot)$ and $v^2(\cdot)$.*
- *Theorem 1 also applies to $\hat{X}_k$ for any k, thereby establishing pathwise strong convergence of the estimated sample path to the true process.*
- *$\alpha_n = \beta_n = n^{-1/2}$ for multiple linear regression and $n^{-1/3}$ for local linear regression.*

Finite Sample Performance

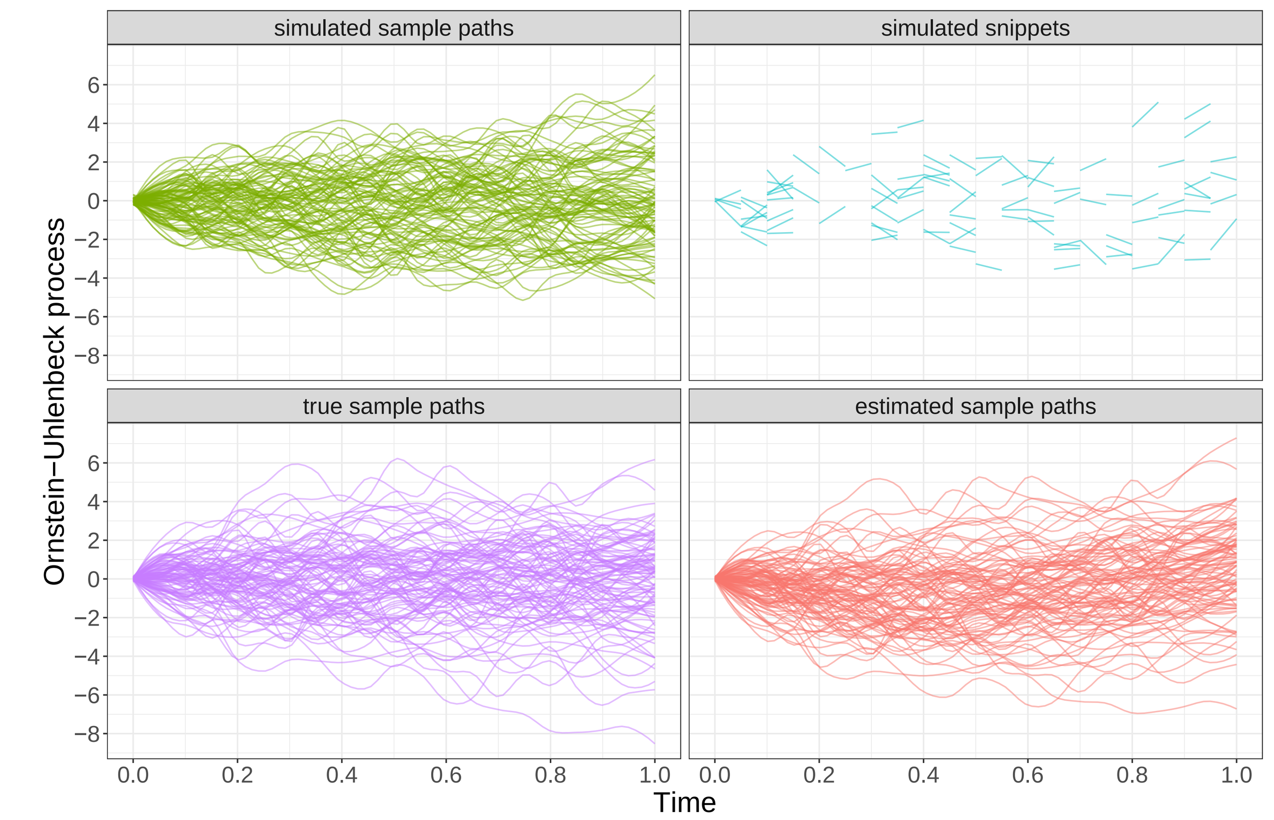


Figure 2. $M = 100$ simulated sample paths (top left), simulated snippets (top right), true sample paths (bottom left), and estimated sample paths (bottom right) for the Ornstein-Uhlenbeck (O-U) process. The sample size is $n = 100$ and the noise level is $\nu = 0.1$.

Sample size \ Noise level	Ho-Lee model			O-U process		
	0	0.01	0.1	0	0.01	0.1
100	0.56	0.56	0.58	1.28	1.31	1.33
200	0.39	0.40	0.41	0.89	0.89	0.91
500	0.24	0.23	0.27	0.56	0.53	0.54
1000	0.17	0.17	0.21	0.37	0.37	0.38
2000	0.12	0.12	0.17	0.25	0.26	0.26
5000	0.07	0.07	0.14	0.16	0.16	0.17

Table 1. Average root-mean-square error for different sample sizes and noise levels.

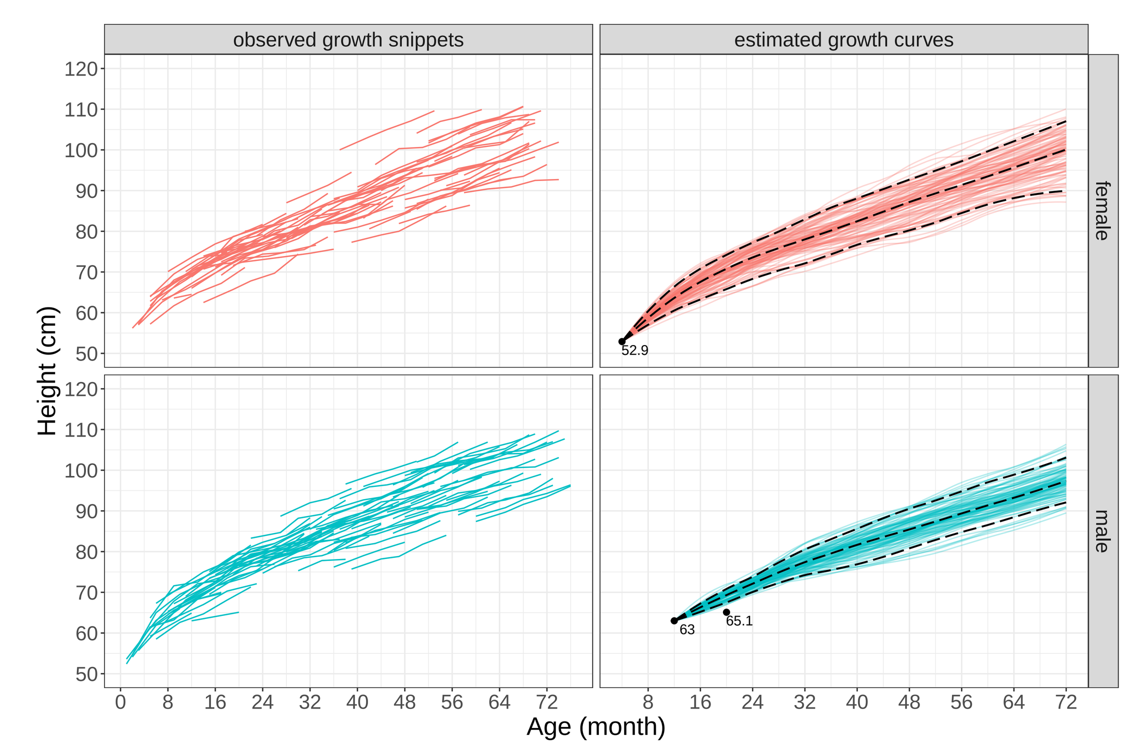
- $\mathrm{ARMSE} = Q^{-1}\sum_{q=1}^{Q} \mathrm{RMSE}_q$, where

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

- The ARMSE decreases with increasing sample size, while the presence of noise does not impact the results much.

Nepal Growth Study Data

This data set contains height measurements for 107 males and 93 females from rural Nepal taken at five adjacent time points from birth to 76 months, spaced approximately four months apart.



- The starting height X_0 is chosen as the initial height measurement.
- Compared to the estimated growth patterns, the recent height measurement for the selected male falls below the 5% percentile curve, suggesting potential developmental delay and the need for additional monitoring.

Figure 3. Observed growth snippets (left) and estimated growth curves (right) for the Nepal growth study data. The black dashed curves indicate 5%, 50%, and 95% percentiles. Height measurements for the selected female and male are also highlighted.

Relevant Literature

- Zhou, Y., & Müller, H. G. (2023). Dynamic Modeling of Sparse Longitudinal Data and Functional Snippets With Stochastic Differential Equations. arXiv:2306.10221.