Dynamic Modelling of Sparse Longitudinal Data and Functional Snippets

Yidong Zhou, UC Davis

(joint work with Hans-Georg Müller)

GGB Seminar, April 15, 2025

What is Functional Data

- Functional data arise when the basic observational unit is a function or curve, rather than a scalar or vector.
- Common in longitudinal, biomedical, financial, and engineering applications.
- Example: Growth curves, ECG signals, temperature trajectories, stock price curves.

Types of Functional data

Type	Description	Examples
Dense	Many measurements per subject at well-spaced time points across the entire domain	Wearable device data, EEG, growth studies
Sparse	A small number of measurements per subject, spread over the entire domain	Survey data, clinical visits
Snippet	Few measurements per subject over a narrow sub-interval of the domain	Accelerated longitudinal studies

Dense Functional Data

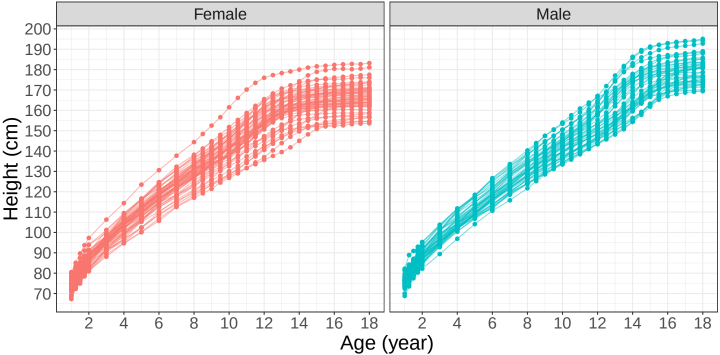


Figure 1: Berkeley Growth Study: Height measurements for 54 females (left) and 39 males (right), each with 31 regularly spaced observations (Tuddenham & Snyder, 1954).

Sparse Functional Data

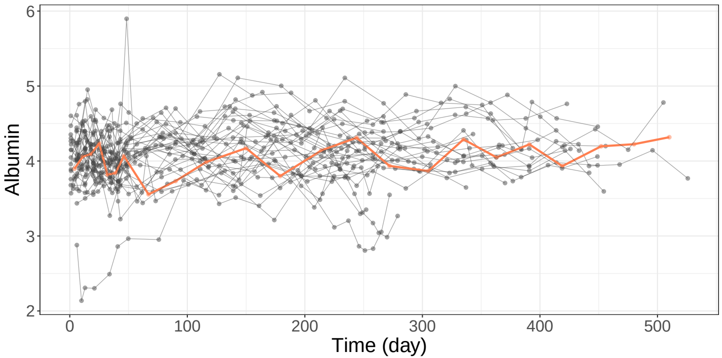


Figure 2: Albumin levels measured for 35 hemodialysis patients, each with 12–18 irregularly spaced observations (Kaysen et al., 2000).

Functional Snippets

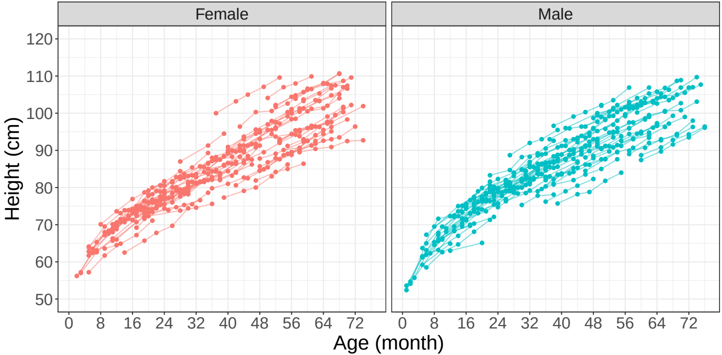


Figure 3: Nepal Growth Study: Height measurements for 87 females (left) and 96 males (right), each with 2–5 observations spaced approximately four months apart, spanning at most 16 months (West Jr et al., 1997).

Functional Snippets

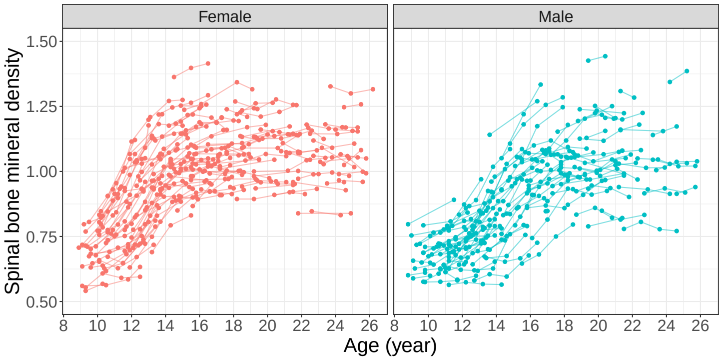


Figure 4: Spinal Bone Mineral Density Study: Bone density measurements for 153 females (left) and 127 males (right), each with 2–4 observations spaced approximately one year apart (Bachrach et al., 1999).

What Are Functional Snippets?

- In accelerated longitudinal studies, subjects are enrolled in the study at a random time and are only tracked for a limited amount of time relative to the domain of interest.

- These designs are common in social and life sciences due to *lower cost, reduced burden, and shorter follow-up per subject*.

- Denote the domain of interest by $\mathcal{T} = [a, b]$. Subject i is only observed over a short interval $[A_i, B_i] \subset \mathcal{T}$, where

$$B_i - A_i \leq \eta(b - a), \quad \text{for some } \eta \in (0, 1).$$

- When η is much smaller than 1, these are functional snippets.

The Core Challenge of Functional Snippets

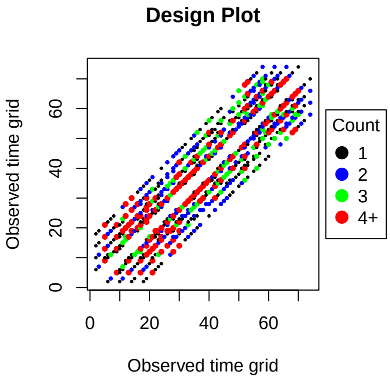
- Assume the observed snippets are generated by an underlying stochastic process $X_t = X(t)$ defined on a compact domain $\mathcal{T}$, which we take without loss of generality to be $[0, 1]$.

- In standard functional data analysis, we typically estimate:
  - Mean function: $\mu(t) = E(X_t)$
  - Covariance function: $\Sigma(s, t) = \mathrm{Cov}(X_s, X_t)$

The Core Challenge of Functional Snippets

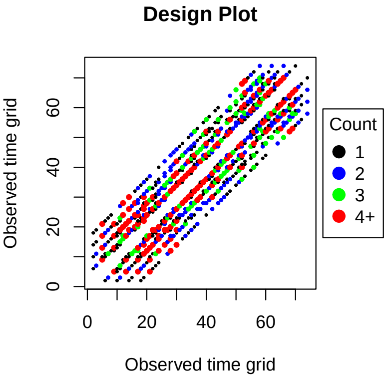
For functional snippets, observations are concentrated near the diagonal in the design plot.

- Design plot for females in the Nepal Growth Study.
- There is no information in the off-diagonal regions.
- Covariance estimation is ill-posed.
- Functional PCA and smoothing-based methods fail.



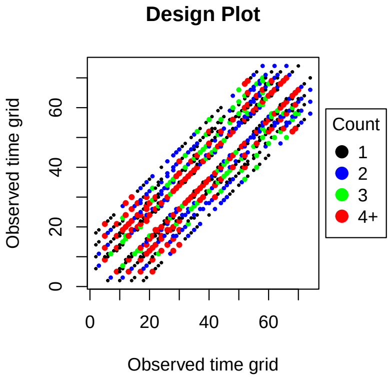
The Core Challenge of Functional Snippets

Covariance completion methods require strong, unverifiable assumptions about the global structure of the covariance function.



The Core Challenge of Functional Snippets

Covariance completion methods require strong, unverifiable assumptions about the global structure of the covariance function.



We need a fundamentally different approach.

A New Perspective: Modelling Dynamics via SDEs

- **Our approach:** Instead of estimating the covariance structure, we model the underlying stochastic process X_t directly as the solution of a data-adaptive stochastic differential equation (SDE).

- **Key idea**: Learn the local dynamics of X_t through the SDE

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

where B_t is Brownian motion and b, σ are drift and diffusion terms.

A New Perspective: Modelling Dynamics via SDEs

- Rather than imposing strong structural assumptions, we learn b and σ **nonparametrically** from the data via conditional moments.

- This SDE-based framework enables the recovery of dynamic distributions at the subject level, even from minimal snippets.

From SDE to Diffusion Process

- Consider the stochastic differential equation (SDE):

$$dX_t = b(t, X_t)\, dt + \sigma(t, X_t)\, dB_t, \quad t \in \mathcal{T}.$$

- **If** $b(t, x)$ and $\sigma(t, x)$ satisfy two regularity conditions:
    - Lipschitz condition:

    $$|b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leq C|x - y|.$$

    - Linear growth condition:

    $$|b(t, x)| + |\sigma(t, x)| \leq C(1 + |x|).$$

- Then the SDE has a **unique solution**, and X_t is a diffusion process (a continuous-time stochastic process with continuous sample paths).

Characterizing Drift and Diffusion

- For a diffusion process X_t, the drift and diffusion coefficients can be interpreted as instantaneous rates of change:
- **Drift**: instantaneous change in the conditional mean.

$$b(t, x) = \lim_{s \to t^+} \frac{\mathbb{E}(X_s - X_t | X_t = x)}{s - t}$$

- **Diffusion**: instantaneous change in the conditional variance.

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

- These local characterizations make it possible to learn the local dynamics of X_t from data.

Reformulating the SDE

To recover paths of X_t from snippets, we rewrite the SDE by plugging in the alternative characterization of drift and diffusion coefficients:

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

Reformulating the SDE

To recover paths of X_t from snippets, we rewrite the SDE by plugging in the alternative characterization of drift and diffusion coefficients:

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

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.

Discretizing the SDE

- Given a time grid $0 = t_0 < t_1 < \cdots < t_K = 1$ with spacing Δ, we approximate the process recursively:

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

where $X_k = X_{t_k}$ for $k = 0, \ldots, K$.

Discretizing the SDE

- Given a time grid $0 = t_0 < t_1 < \cdots < t_K = 1$ with spacing Δ, we approximate the process recursively:

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

where $X_k = X_{t_k}$ for $k = 0, \ldots, K$.

- Using properties of conditional expectation and Brownian increments:

$$\mathbb{E}(X_{k-1} | X_{k-1}) = X_{k-1}, \quad (B_{t_k} - B_{t_{k-1}})/\sqrt{\Delta} \sim N(0, 1).$$

Discretizing the SDE

- Given a time grid $0 = t_0 < t_1 < \cdots < t_K = 1$ with spacing Δ, we approximate the process recursively:

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

where $X_k = X_{t_k}$ for $k = 0, \dots, K$.

- Using properties of conditional expectation and Brownian increments:

$$\mathbb{E}(X_{k-1}|X_{k-1}) = X_{k-1}, \quad (B_{t_k} - B_{t_{k-1}})/\sqrt{\Delta} \sim N(0,1).$$

- The recursion simplifies to:

$$X_k = \mathbb{E}(X_k|X_{k-1}) + \{\mathrm{Var}(X_k|X_{k-1})\}^{1/2} W_k, \quad W_k \sim N(0,1).$$

This defines the evolution of X_k from X_{k-1} using conditional mean and conditional variance, which provides a practical simulation strategy to reconstruct paths from snippets.

- To reconstruct paths of X_t, one needs to iteratively generate a sample from the Gaussian distribution $N\{E(X_k|X_{k-1}), \mathrm{Var}(X_k|X_{k-1})\}$.

- In practice, both the conditional mean $E(X_k|X_{k-1})$ and the conditional variance $\mathrm{Var}(X_k|X_{k-1})$ are unknown and need to be estimated from data.

Estimating Conditional Mean and Variance

- $X_{k-1} = X_{t_{k-1}}$ contains two pieces of information: measurement X_{k-1} and observation time t_{k-1}.

- One can then formulate the estimation of $E(X_k|X_{k-1})$ and $\mathrm{Var}(X_k|X_{k-1})$ as a regression problem with response X_k and predictor $(X_{k-1}, t_{k-1})'$.

Constructing the Regression Dataset

- Each subject is observed at least twice: say at T_{i1} and T_{i2}, with observations Y_{i1} and Y_{i2}.

- Define $Z_i = (Y_{i1}, T_{i1})'$ and $Y_i = Y_{i2}$ for $i = 1, \ldots, n$.

- Consider the regression model:

$$Y_i = m(Z_i) + v(Z_i)\,\epsilon_i, \quad \epsilon_i \sim N(0, 1).$$

- The conditional mean $m(\cdot)$ can be estimated using standard regression techniques such as multiple linear regression or local linear regression.

- For conditional variance $v^2(\cdot)$, we fit a regression model to the squared residuals:

$$\{Y_i - \hat{m}(Z_i)\}^2 \sim Z_i.$$

Final Estimation

With the estimated conditional mean and variance $\hat{m}(\cdot)$ and $\hat{v}^2(\cdot)$, we reconstruct the sample path of X_t recursively, starting from an initial condition $X_0 = x_0$:

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

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

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: $(\hat{X}_1, \ldots, \hat{X}_K)'$.

```
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 ~ N{m̂(Ẑ_{k-1}), v̂²(Ẑ_{k-1})};
    Set Ẑ_k ← (X̂_k, t_k)';
end
```

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

- Estimate the conditional mean $E(X_k|X_{k-1})$ and conditional variance $\mathrm{Var}(X_k|X_{k-1})$ by $\hat{m}(\hat{Z}_{k-1})$ and $\hat{v}^2(\hat{Z}_{k-1})$, respectively;
- Draw a random sample $\hat{X}_k \sim N\{\hat{m}(\hat{Z}_{k-1}), \hat{v}^2(\hat{Z}_{k-1})\}$;
- Set $\hat{Z}_k \leftarrow (\hat{X}_k, t_k)'$;

end

Theoretical Foundations: Existence and Uniqueness

- Under regularity conditions and Gaussianity, the alternative SDE formulation we use admits a pathwise unique strong solution.

- *Takeaway:* Our data-driven SDE is not just a heuristic — it corresponds to a well-defined stochastic process.

Main Theoretical Result: Pathwise Convergence

Theorem

Assume regularity conditions and Gaussianity. Then the estimated sample path obtained from Algorithm 1 satisfies

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

where α_n and β_n are the convergence rates for the estimated conditional mean function $\hat{m}(\cdot)$ and variance function $\hat{v}^2(\cdot)$, respectively.

This result ensures that both the mean and variance of $\hat{X}_k$ consistently approximate those of the true process X_k. Moreover, the convergence holds uniformly over $k = 1, \ldots, K$, thereby establishing the pathwise consistency of the estimated sample path.

Corollary: Distributional Convergence

Corollary

Under the same assumptions as the theorem, the distribution of the estimated process $\hat{X}_K$ converges to that of the true process X_K in Wasserstein distance:

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

where $\mathcal{L}(X_K)$ denotes the law of X_K, and d_W is the Wasserstein distance.

This result means our method not only reconstructs individual paths accurately, but also recovers the correct population-level distribution of outcomes.

Understanding the Convergence Rates

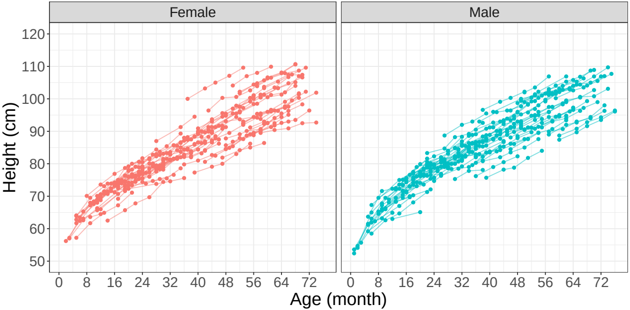
- The theoretical accuracy of our reconstructed path depends on two key quantities: α_n and β_n.
- These rates depend on the choice of regression method used to estimate the conditional mean and variance.

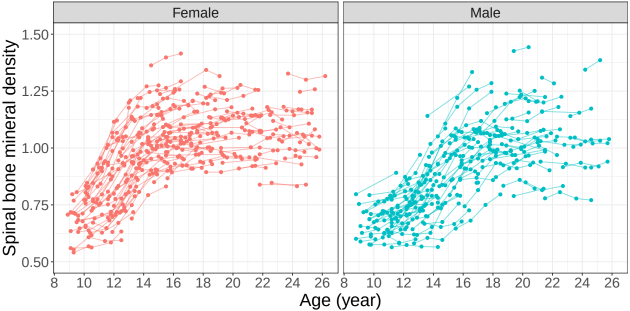
$$\left[E\{|\hat{m}(z) - m(z)|^2\}\right]^{1/2} = O(\alpha_n), \quad \left[E\{|\hat{v}^2(z) - v^2(z)|^2\}\right]^{1/2} = O(\beta_n).$$

- If the same regression method is used for both $\hat{m}$ and $\hat{v}^2$, then $\beta_n = \alpha_n$. Typical examples include
  - Multiple linear regression: $\alpha_n = \beta_n = n^{-1/2}$.
  - Local linear regression: $\alpha_n = \beta_n = n^{-1/3}$.

Overview of Real Data Applications

- We apply the proposed method to two longitudinal datasets: Nepal Growth Study and Spinal Bone Mineral Density Study.
- Both datasets feature:
  - Irregular and sparse measurements across individuals.
  - Short longitudinal windows per subject.
  - No full-trajectory coverage across individuals.
- These properties make them well-suited for evaluating the proposed SDE-based modelling of functional snippets.


Female
Male
Height (cm)
Age (month)


Female
Male
Spinal bone mineral density
Age (year)

Summary of Real Data Applications

Nepal Growth Study

- $n = 183$ (87 females, 96 males).
- 2–5 height measurements per child over a short window of approximately 16 months.

Spinal Bone Mineral Density Study

- $n = 280$ (153 females, 127 males).
- 2–4 bone mineral density measurements per subject, taken annually.

Modelling Setup

- We apply the proposed method separately to male and female subjects.

- Conditional mean $\hat{m}(\cdot)$ and variance $\hat{v}^2(\cdot)$ are estimated using local linear regression.

Nepal Growth Study: Growth Monitoring

- Beyond recovering population trends, the proposed method enables individualized growth monitoring — predicting a child's future development from minimal early data.

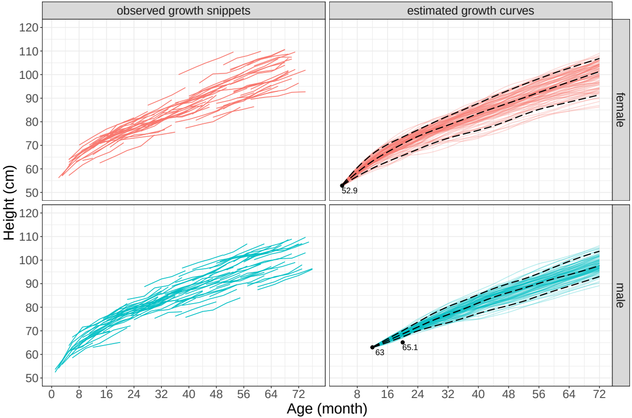
- As new measurements become available, they can be compared against the predicted growth trajectory to screen for developmental deviations.

Nepal Growth Study: Growth Monitoring

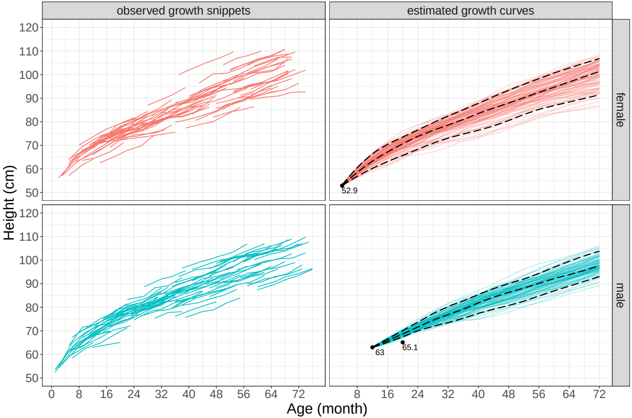
- We illustrate this using two children not included in model fitting:
  - Selected female: only one height measurement at 4 months: 52.9 cm.
  - Selected male: two measurements at 12 and 20 months: 63 cm and 65.1 cm.

- For each child, we simulate 100 sample paths using the recursive procedure in Algorithm 1 and construct 5%, 50%, and 95% percentile growth curves.

Nepal Growth Study: Key Findings

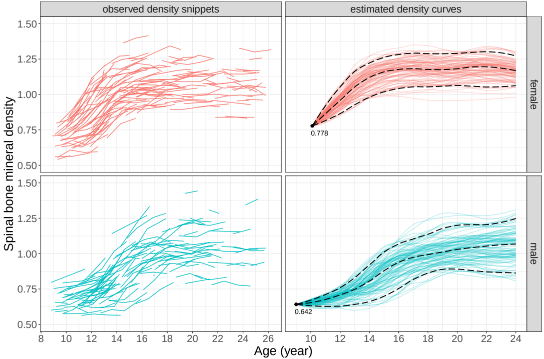


Nepal Growth Study: Key Findings

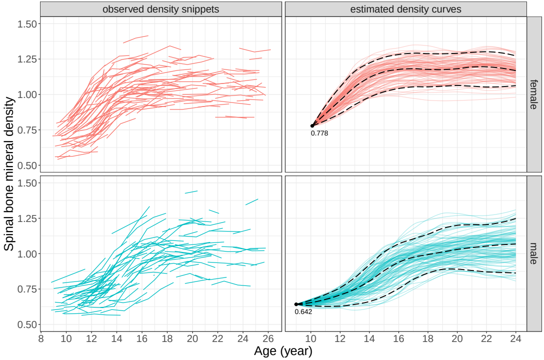


For the selected male, the new observed height at 20 months (65.1 cm) falls below the 5% percentile, potentially indicating growth delay and prompting clinical follow-up.

Spinal Bone Mineral Density Study



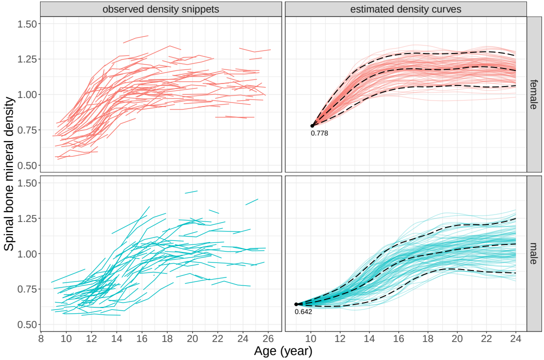
Spinal Bone Mineral Density Study



The reconstructed curves reflect known physiological trends:

- Female plateaus around age 16.
- Male plateaus later, around age 18.

Spinal Bone Mineral Density Study



The reconstructed curves reflect known physiological trends:

- Female plateaus around age 16.
- Male plateaus later, around age 18.

The model reconstructs realistic subject-specific trajectories despite data sparsity, effectively capturing growth trends and uncertainty.

Key Takeaways

- We proposed a dynamic modelling framework for functional snippets via data-adaptive SDEs.
- Our approach bypasses covariance estimation and enables subject-level path reconstruction.
- Theoretical guarantees establish pathwise consistency of the reconstructed trajectories.
- Applications to growth and bone density data demonstrate the method's flexibility and clinical utility, especially for early screening and prediction.

References

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*(12), 4702–4712.

Kaysen, G. A., Dubin, J. A., Müller, H.-G., Rosales, L. M., & Levin, N. W. (2000). The acute-phase response varies with time and predicts serum albumin levels in hemodialysis patients. *Kidney International*, *58*, 346–352.

Tuddenham, R. D., & Snyder, M. M. (1954). Physical growth of California boys and girls from birth to eighteen years. *University of California Publications in Child Development*, *1*(2), 183–364.

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*(10), 1957–1965.

Questions?