A short introduction to functional data

Zhuosong Zhang

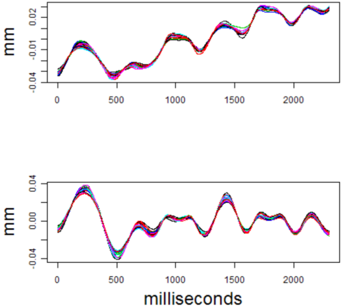
School of Mathematics and Statistics

December 28, 2018

Outline

Some examples

Example: 20 replications, 2401 observations within replications, 2 dimensions:

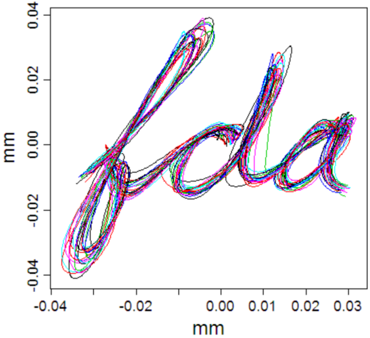


The immediate characteristics:

- High-frequency measurements
- Smooth, but complex
- Repeated observations
- Multiple dimensions

Some examples: Handwriting data

Measures of position of nib of a pen writing "fda". 20 replications, measurements taken at 200 Hz.



What is functional data?

- Müller (2006): Functional data is multivariate data with an ordering on the dimensions.

What are we intereted in?

- Representations of distribution of functions
    - mean
    - variation
    - covariation
- Relationships of functional data to
    - covariates
    - responses
    - other functions
- and so on...

Challenges

- Estimation of functional data from noisy, discrete or missing observations.
- Numerical representation of infinite-dimensional objects.
- Representation of infinite-dimensional objects.
- Description of statistical relationships between infinite dimensional objects.
- And so on...

Outline

1. Functional data
2. Linear regression
3. Partially observed data

Linear regression

- Suppose we have a data $\{(X_1, Y_1), ..., (X_n, Y_n)\}$, where the $X_i(\cdot)$'s are i.i.d. as a random function $X(\cdot)$ with mean function 0, defined on an interval $\mathcal{I}$, and the Y_i's are generated by the regression model:

$$Y_i = \ a \ + \int_{\mathcal{I}} b(t) \ X_i(t) dt + \varepsilon_i.$$

Linear regression

- Suppose we have a data $\{(X_1, Y_1), ..., (X_n, Y_n)\}$, where the $X_i(\cdot)$'s are i.i.d. as a random function $X(\cdot)$ with mean function 0, defined on an interval $\mathcal{I}$, and the Y_i's are generated by the regression model:

$$Y_i = a + \int_{\mathcal{I}} b(t) X_i(t) dt + \varepsilon_i.$$

Linear regression

- Suppose we have a data $\{(X_1, Y_1), ..., (X_n, Y_n)\}$, where the $X_i(\cdot)$'s are i.i.d. as a random function $X(\cdot)$ with mean function 0, defined on an interval $\mathcal{I}$, and the Y_i's are generated by the regression model:

$$Y_i = a + \int_{\mathcal{I}} b(t) X_i(t)dt + \varepsilon_i.$$

- a is a constant: the intercept
- $b(\cdot)$ is a function on $\mathcal{I}$: the slope function
- ε_i i.i.d. random variables, independent of $\{X_i; 1 \le i \le n\}$
  $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{E}\left(\varepsilon_i^2\right) < \infty$.

Linear regression

- Suppose we have a data $\{(X_1, Y_1), ..., (X_n, Y_n)\}$, where the $X_i(\cdot)$'s are i.i.d. as a random function $X(\cdot)$ with mean function 0, defined on an interval $\mathcal{I}$, and the Y_i's are generated by the regression model:

$$Y_i = a + \int_{\mathcal{I}} b(t) X_i(t)dt + \varepsilon_i.$$

- a is a constant: the intercept
- $b(\cdot)$ is a function on $\mathcal{I}$: the slope function
- ε_i i.i.d. random variables, independent of $\{X_i; 1 \leq i \leq n\}$
  $\mathrm{E}\left(\varepsilon_i\right) = 0$ and $\mathrm{E}\left(\varepsilon_i^2\right) < \infty$.
- Question: How to estimate a and b?

- The covariance function:

$$K(u,v) = \mathrm{Cov}\left\{X(u), X(v)\right\} = \sum_{j=1}^{\infty} \lambda_j \phi_j(u)\phi_j(v), \quad u, v \in \mathcal{I},$$

where the λ_j and ϕ_j are eigenvalues and eigenfunctions of K and $\lambda_1 > \lambda_2 > ...$

- Write

$$b(t) = \sum_{j=1}^{\infty} b_j \phi_j(t),$$

where $b_j = \int_{\mathcal{I}} b(t)\phi_j(t)dt$. Let

$$\begin{aligned} g(u) &= \mathrm{Cov}\big\{X(u), Y\big\} \\ &= \int_{\mathcal{I}} K(u,v)b(v)dv \\ &= \sum_{j=1}^{\infty} \lambda_j b_j \phi_j(u). \end{aligned}$$

- The covariance function:

$$K(u,v) = \mathrm{Cov}\left\{X(u), X(v)\right\} = \sum_{j=1}^{\infty} \lambda_j \phi_j(u)\phi_j(v), \quad u, v \in \mathcal{I},$$

  where the λ_j and ϕ_j are eigenvalues and eigenfunctions of K and $\lambda_1 > \lambda_2 > \ldots$

- Write

$$b(t) = \sum_{j=1}^{\infty} b_j \phi_j(t),$$

  where $b_j = \int_{\mathcal{I}} b(t)\phi_j(t)dt$. Let

$$\begin{aligned} g(u) &= \mathrm{Cov}\big\{X(u), Y\big\} \\ &= \int_{\mathcal{I}} K(u,v)b(v)dv \\ &= \sum_{j=1}^{\infty} \lambda_j b_j \phi_j(u). \end{aligned} \quad \Longrightarrow \quad \begin{aligned} g_j &= \int_{\mathcal{I}} g(t)\phi_j(t)dt, \\ & \boxed{b_j = \lambda_j^{-1} g_j.} \end{aligned}$$

This gives the idea to estimate b.

- To construct the estimation of a and b, we need the estimation of K.
- The empirical version of K and g,

$$\hat{K}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\big(X_i(u) - \bar{X}(u)\big)\big(X_i(v) - \bar{X}(v)\big) = \sum_{j=1}^{\infty} \hat{\lambda}_j \hat{\phi}_j(u)\hat{\phi}_j(v),$$

$$\hat{g}(u) = \frac{1}{n}\sum_{i=1}^{n}\big\{X_i(t) - \bar{X}(t)\big\}\big(Y_i - \bar{Y}\big) = \sum_{j=1}^{\infty} \hat{g}_j \hat{\phi}_j(u),$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots$

- To construct the estimation of a and b, we need the estimation of K.
- The empirical version of K and g,

$$\hat{K}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\big(X_i(u)-\bar{X}(u)\big)\big(X_i(v)-\bar{X}(v)\big) = \boxed{\textstyle\sum_{j=1}^{\infty}\hat{\lambda}_j\hat{\phi}_j(u)\hat{\phi}_j(v)},$$

$$\hat{g}(u) = \frac{1}{n}\sum_{i=1}^{n}\big\{X_i(t)-\bar{X}(t)\big\}\big(Y_i-\bar{Y}\big) = \boxed{\textstyle\sum_{j=1}^{\infty}\hat{g}_j\hat{\phi}_j(u)},$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots$

- Cai and Hall (2006): Take $\boxed{\hat{b}(u) = \sum_{j=1}^{K}\hat{b}_j\hat{\phi}_j(u)}$ where $\boxed{\hat{b}_j = \hat{\lambda}_j^{-1}\hat{g}_j}$, as the estimator of b.
- $\hat{a} = \bar{Y} - \int_{\mathcal{I}}\hat{b}(u)\bar{X}(u)du.$

Outline

Partially observed functional data

- Unfortunately, in many cases, we cannot fully observe the functional data.
- For example, if we have the data $\{(X_i, \mathcal{I}_i); 1 \le i \le n\}$ where $X_i(t)$ is a functional data and $\mathcal{I}_i = [A_i, B_i]$ is a random interval. For each i, we observe $X_i(t), t \in \mathcal{I}_i$.
- We cannot estimate $K(s,t)$ based on the partially observed data.
- Re-construct the functional data...

An example: partially observed fragmentary functional data

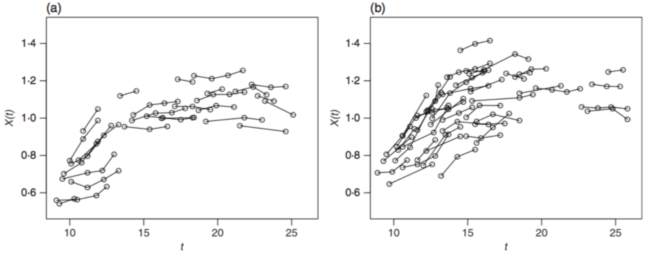


Fig. 1. Curve fragments of growth, measured by the spine bone mineral density, in g cm^{-2}, for females from the (a) Hispanic and (b) Black ethnic groups described in Bachrach et al. (1999).

How to reconstruct the missing data?

- James et al. (2000) and Yao et al. (2005): use a predictor based on the assumption that the data are normally distributed
- Delaigle and Hall (2013): adjoin shifted versions of other observed fragments
- Kraus (2015): construct prediction intervals for principal scores and bands for missing parts of trajectories
- Delaigle and Hall (2016) : suggest an approach based on a combination of Markov chains and non-parametric smoothing techniques

Markov chain method

- Let $\{(X_1, Y_1), ..., (X_n, Y_n)\}$ be i.i.d. data, where X_i is a function supported on $\mathcal{I}_0 = [a, b]$.
- Y_i is a scalar response.
- We only observe X_i on $\mathcal{I}_i = [A_i, B_i] \subset \mathcal{I}_0$.
- The Markov chain model is used based on a discrete version of the process X.

Discretise the functional data

- $\mathcal{I}_0^{\text{disc}} = \{t_1, ..., t_{m_1}\} \subset \mathcal{I}_0$, where $a \leq t_1 < t_2 < \cdots < t_{m_1} \leq b$
- Define $\Gamma = \{z_1, ..., z_{m_2}\}$ where
  $-\infty = z_0 < z_1 < z_2 < \cdots < z_{m_2} < z_{m_2+1} = \infty$
- Define $Z_i(t_j) = z_k$ if $(z_{k-1} + z_k)/2 < X_i(t_j) \leq (z_k + z_{k+1})/2$
- We construct the Markov chain based on the discretised data.

Markov Chain

- Markov property: Let $X_1, X_2, \ldots$ be a sequence of random variables,

$$\mathrm{P}(X_{n+1} = x | X_1 = x_1, \ldots, X_n = x_n) = \mathrm{P}(X_{n+1} = x | X_n = x_n),$$

  where the possible values of X_i form a countable set S, which is called the state space of the chain.

- Transition probability: Let $S = \{s_1, \ldots, s_{m_2}\}$,

$$P = \left(p_{ij}; 1 \leq i, j \leq m_2\right), \text{ where } p_{ij} = \mathrm{P}(X_{n+1} = s_j | X_n = s_i).$$

Come back to our model

- To reconstruct our data, we want to calculate $\mathrm{E}\left(Z_i(t) \,\middle|\, Z_i(s), s \in \mathcal{I}_i^{\text{disc}}\right)$.
- By the Markov property, we have for $H \subset \Gamma$,

$$\begin{aligned} &\mathrm{P}(Z_i(t) \in H | Z_i(s), s \in \mathcal{I}_i^{\text{disc}}) \\ &= \begin{cases} \mathrm{P}(Z_i(t) \in H | Z_i(A_i)) & \text{if } t < A_i, \\ \mathbb{1}(Z_i(t) \in H) & \text{if } A_i \leq t \leq B_i, \\ \mathrm{P}(Z_i(t) \in H | Z_i(B_i)) & \text{if } t > B_i. \end{cases} \end{aligned}$$

- We need to estimate

$$\begin{aligned} p(t_j, z_{k_1}, z_{k_2}) &= \mathrm{P}(Z_{j+1} = z_{k_2} | Z(t_j) = z_{k_1}), \\ q(t_{j+1}, z_{k_1}, z_{k_2}) &= \mathrm{P}(Z_j = z_{k_2} | Z(t_{j+1}) = z_{k_1}). \end{aligned}$$

Estimation of the transition probability

- Suppose (A_i, B_i) is independent of X_i (or Z_i).
- Estimation of $p(t_j, z, z')$: Take

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

and

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

where $\hat{A}$ denotes a smoothed version of N.

- Similarly we can estimate the transition probability q.

Estimating the missing parts of the curves

- We have

$$\hat{Z}(t) = \hat{\mathrm{E}}\left(Z(t) \,|\, Z(A)\right) = \sum_{l=1}^{m_2}\Big(\sum_{\text{paths to } z_l}\prod_{k=1}^{r-1}\hat{q}(t_{j-k+1}, z_{j_k}, z_{j_{k+1}})\Big) z_l, \ \text{if } a \leq t < A,$$

$$\hat{Z}(t) = \hat{\mathrm{E}}\left(Z(t) \,|\, Z(B)\right) = \sum_{l=1}^{m_2}\Big(\sum_{\text{paths to } z_l}\prod_{k=1}^{r-1}\hat{p}(t_{j-k+1}, z_{j_k}, z_{j_{k+1}})\Big) z_l, \ \text{if } B < t \leq b,$$

where the summation $\sum_{\text{paths to } z_l}$ is over all all paths $z^0 = z_{j_1} \to z_{j_2} \to \cdots \to z_{j_r} = z_l$ that leads from state z^0 to z_l in just $r-1$ steps, with z^0 denoting $Z(A)$ or $Z(B)$ in the cases q and p.

An example

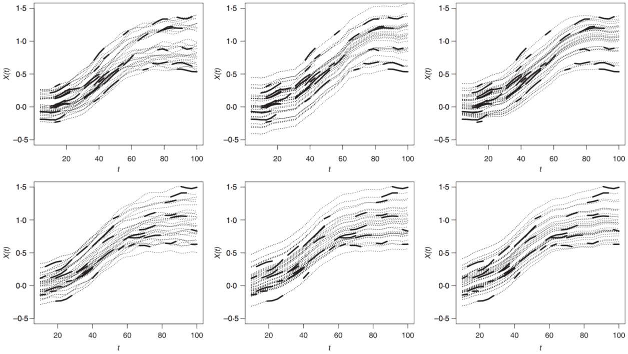


Fig. 3. Reconstruction of $n = 30$ curves for two samples, one in each row of panels, from model (i): true curves (left) and reconstructions using the method of Delaigle & Hall (2013) (middle) or our new approach (right); the observed fragments are shown in bold.

- Let $\mathbf{Z}^O = \left\{Z(t), t \in \mathcal{I}^{\text{disc}}\right\}$, define

$$v_1(t|\mathbf{Z}^O) = \mathrm{E}\left(Z(t)\,\middle|\,\mathbf{Z}^O\right) = \begin{cases} Z(t), & A \le t \le B, \\ \mathrm{E}\left(Z(t) \mid Z(A)\right), & a \le t < A, \\ \mathrm{E}\left(Z(t) \mid Z(B)\right), & B < t \le B. \end{cases}$$

and

$$v_2(t,u|\mathbf{Z}^O) = \begin{cases} Z(t)Z(s), & t,u \in [A,B], \\ Z(t)v_1(u|\mathbf{Z}^O), & t \in [A,B], u \notin [A,B], \\ Z(u)v_1(t|\mathbf{Z}^O), & t \notin [A,B], u \in [A,B], \\ \mathrm{E}\left(Z(t)Z(u) \mid Z(A)\right), & a \le t,u < A, \\ \mathrm{E}\left(Z(t)Z(u) \mid Z(B)\right), & B < t,u \le b, \\ v_1(t|\mathbf{Z}^O)v_1(u|\mathbf{Z}^O), & a \le t < A \le B < u \le b. \end{cases}$$

Estimation of the covariance matrix

- The covariance estimator:

$$\hat{K}(t_k, t_l) = \hat{\mu}_2(t_k, t_l) - \hat{\mu}_1(t_k)\hat{\mu}_1(t_l),$$

where

$$\hat{\mu}_2(t, s) = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_2(s, t|\mathbf{Z}_i^O), \quad \hat{\mu}_1(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_1(s|\mathbf{Z}_i^O).$$

Estimation of the covariance matrix

- The covariance estimator:

$$\hat{K}(t_k, t_l) = \hat{\mu}_2(t_k, t_l) - \hat{\mu}_1(t_k)\hat{\mu}_1(t_l),$$

where

$$\hat{\mu}_2(t, s) = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_2(s, t|\mathbf{Z}_i^O), \quad \hat{\mu}_1(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_1(s|\mathbf{Z}_i^O).$$

- After that, we can do something more...

Linear regression for partially observed segment data

- For functional linear model,

$$Y = a + \int_{\mathcal{I}_0} bX + \varepsilon, \quad \mathrm{E}\left(\varepsilon \,\middle|\, X(t), t \in \mathcal{I}_0\right) = 0.$$

- Since we observe only fragments $X_i(t), t \in \mathcal{I}_i \subset \mathcal{I}_0$, we suggests the following model

$$Y = a + \sum_{j=1}^{m_1} b_j Z(t_j) + \varepsilon, \quad \mathrm{E}\left(\varepsilon \,\middle|\, Z(t), t \in \mathcal{I}_0^{\mathrm{disc}}\right) = 0,$$

where Z is the discrete process.

Estimation of a and b_j

- Instead of minimizing

$$\sum_{i=1}^{n}\left(Y_i - a - \sum_{j=1}^{m_1} b_j Z_i(t_j)\right)^2,$$

- we suggests:

$$(a, b_1, ..., b_{m_1}) = \underset{a,b_1,...,b_{m_1}}{\arg\min} \sum_{i=1}^{n}\left(Y_i - a - \sum_{j=1}^{m_1} b_j \hat{v}_1(t_j|Z_i, \mathcal{I}_i)\right)^2.$$

An example

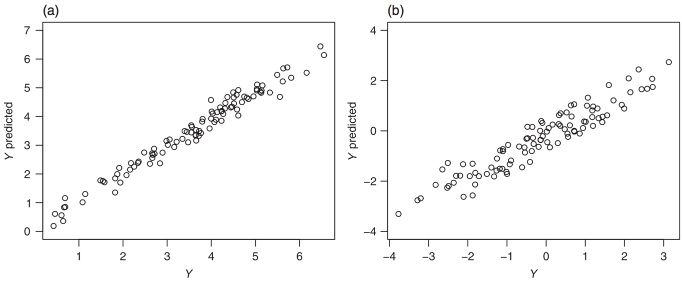


Fig. 5. Scatterplots of pairs $(Y_{\mathrm{NEW},i}, \hat{Y}_{\mathrm{NEW},i})$ for $i = 1, \ldots, 100$, when $\hat{Y}_{\mathrm{NEW},i}$ is the predictor proposed in this paper, computed from data observed on two fragments generated from (a) model (i) or (b) model (ii), with $n = 50$.

Thank you!