Reconstruction of incomplete functional data when missingness is not at random

MIPCC 2017

Debajit Dutta

Principal Supervisor: Prof. Aurore Delaigle

School of Mathematics and Statistics

University of Melbourne

What is incomplete functional data?

- The data consists of a series of independent and identically distributed curves $X_i, 1 \leq i \leq n$ each defined over a common support, a common interval denoted by $\mathcal{I}_0 = [a, b]$.
- For each i, X_i is observed in an interval $\mathcal{I}_i = [A_i, B_i] \subseteq \mathcal{I}_0$.

What is incomplete functional data? (contd.)

The following figure shows the type of incomplete functional data we are interested in:

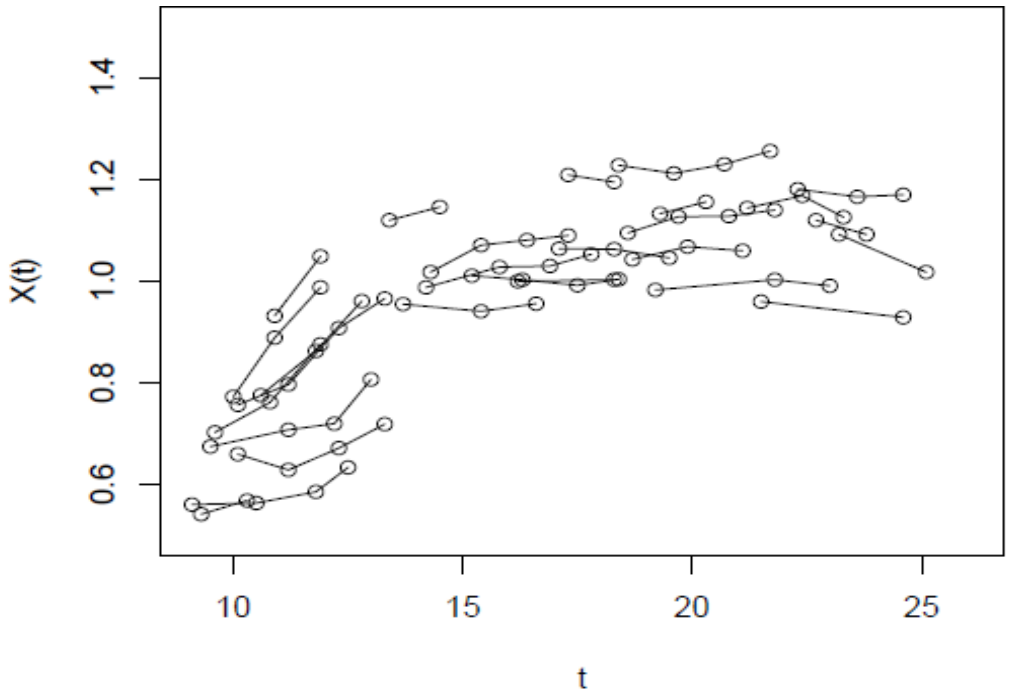

Figure: Curve fragments of growth, measured by the spine bone mineral density, in g=cm2, for females from the Hispanic ethnic group (right) described in Bachrach et al. (1999) (Delaigle and Hall, 2016).

What is incomplete functional data? (contd.)

Problems associated with fragmentary functional data with "little" overlap:

- Difficult to estimate the mean function i.e. $\mu(t) = \mathrm{E}X(t)$ for some t.
- Difficult to estimate co-variance function i.e. $\mathrm{K}(s,t) = \mathrm{Cov}\{X(s), X(t)\}$ at some (s,t).
- The above points makes it hard to apply techniques like classification and clustering for analyzing the data.

Reconstruction of incomplete curves with almost no overlap: existing methods

Delaigle and Hall (2013):

- The incomplete fragments are extended beyond its endpoints by adjoining fragments obtained from those which are observed.
- The method of choosing the observed fragment is either by randomisation or by using the fragments which are of similar shape as the incomplete curve.
- This method does not work in case more than one fragment per curve is observed and it forces each reconstructed curve to have the shape of an observed fragment.

Reconstruction of incomplete curves with almost no overlap: existing methods (contd.)

Delaigle and Hall (2016):

- Approach based on discretization with respect to both space and time.
- The time axis is divided into a horizontal grid of points $\{t_j\}_{j=1}^{m_1}$.
- Similarly the space domain is divided into a vertical grid, $\{z_k\}_{k=1}^{m_2}$ in the space domain.
- The spatial discretization is constructed using the following criterion:

$$Z_i(t_j) = z_k \quad \text{if} \quad \frac{z_{k-1} + z_k}{2} < X_i(t_j) \leq \frac{z_k + z_{k+1}}{2} \ .$$

- For convenience, it is assumed that t_j's and z_k's are equidistant although this may not be necessarily true.

Reconstruction of incomplete curves with almost no overlap: existing methods (contd.)

Delaigle and Hall (2016) (contd.):

- The transition of of $Z_i(t_j)$ to $Z_i(t_{j+1})$ or from $Z_i(t_{j+1})$ to $Z_i(t_j)$ is modelled using Markov chain.
- The missing values are imputed by conditional means [conditioned on the values of Z_i observed at the end-points i.e. $Z_i(A_i)$ or $Z_i(B_i)$ where $[A_i, B_i]$ is the interval where Z_i is observed] calculated using the transition probabilities.

Side by side comparison of the true curves (left) with the reconstructed parts using Delaigle and Hall's 2013 (middle) and 2016 method (right).

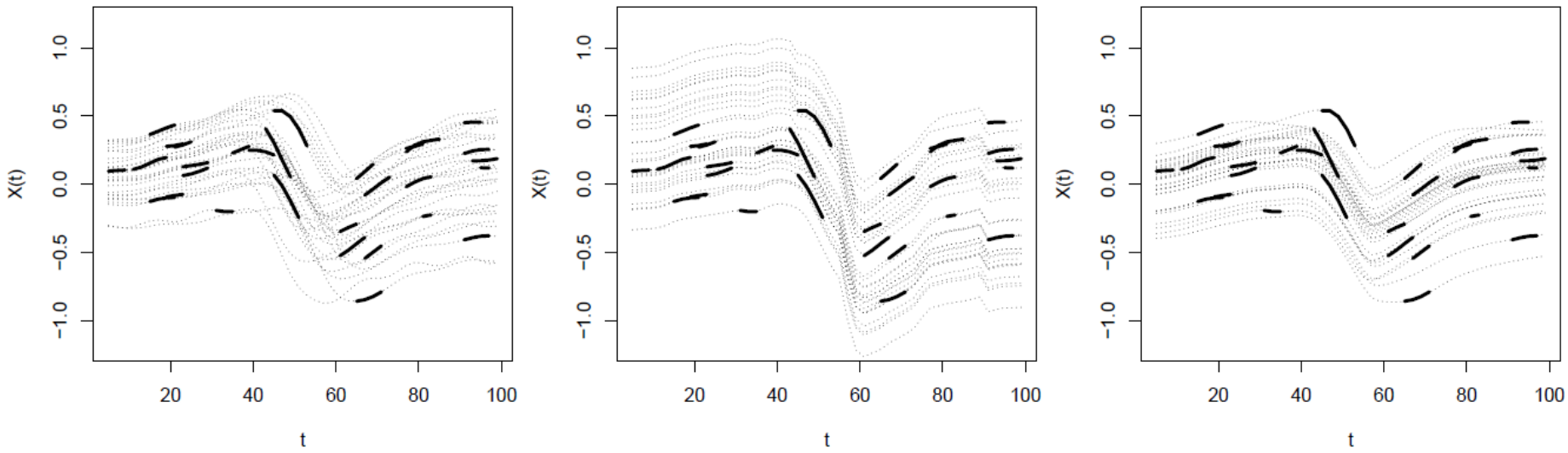

Figure: Dotted lines represent the true curves (left) and reconstructed parts (middle and right). Observed fragments are shown in bold (simulation from model (ii), Delaigle and Hall, 2016).

Objective of the present study

- The present work is an extension to Delaigle and Hall (2016) where the transition probability estimation was carried out using the assumption that the random functions X_i and the intervals $[A_i, B_i]$ are independent.
- The aim is to construct consistent estimators of the transition probabilities when the above independence may not hold.
- We have assumed for the time being that each curve fragment of X_i is observed at the origin of measurement.

Assumptions in the present settings

Specifically, we shall use the following modified set of assumptions:

- The X_i's are independent and identically distributed.
- The B_i's are independent and identically distributed.
- For any $t_j \in \{t_1, \ldots, t_{m_1-1}\}$ in the time grid, $\Pr(B_i > t_j) > 0$.

Methodology

- Similar to Delaigle and Hall (2016) a naive estimator of $p(t_j, z, z') = \Pr\{Z_i(t_{j+1}) = z' | Z_i(t_j) = z\}$ is obtained as:

$$\hat{p}(t_j, z, z') = \frac{N(t_j, z, z')}{N(t_j, z)}\,. \tag{1}$$

- Here $N(t_j, z, z') = \sum_{i=1}^{n} I\{Z_i(t_j) = z, Z_i(t_{j+1}) = z', t_j < B_i\}$ and $N(t_j, z) = \sum_{z'=z_1}^{z_{m_2}} N(t_j, z, z')$, $\forall\ t_j, z, z'$.
- The following estimating equation will lead to the same estimates as in (1):

$$\sum_{i=1}^{n} \Big[I\{Z_i(t_j) = z, Z_i(t_{j+1}) = z', t_j < B_i\} - p(t_j, z, z') I\{Z_i(t_j) = z, t_j < B_i\} \Big] = 0\,. \tag{2}$$

Methodology (contd.)

- In line with the methodology described in Molenberghs et al. (2015) for inverse probability methods we need to consider an unbiased estimating equation.
- We first consider a parametric model for the probability $\Pr\{B_i > t_j | \boldsymbol{Z}_i\}$. We denote this by $\pi\{t_j, \boldsymbol{Z}_i; \boldsymbol{\zeta}\}$ where $\boldsymbol{\zeta}$ is the parameter.
- Assuming $\boldsymbol{\zeta}$ to be known, we modify (4) in the following way:

$$\sum_{i=1}^{n} \pi^{-1}\{t_j, \boldsymbol{Z}_i; \boldsymbol{\zeta}\} \Big[I\{Z_i(t_j) = z, Z_i(t_{j+1}) = z', t_j < B_i\} - p(t_j, z, z') I\{Z_i(t_j) = z, t_j < B_i\} \Big] = 0\,. \tag{3}$$

Methodology (contd.)

- Let $\hat{\zeta}$ denote a consistent estimator of ζ.
- We propose estimating $p(t_j, z, z)$ using the quantities obtained as a solutions to (3) with $\hat{\zeta}$ in place of ζ:

$$\hat{p}(t_j, z, z'; \hat{\zeta}) = \breve{N}(t_j, z, z'; \hat{\zeta}) / \breve{N}(t_j, z; \hat{\zeta}) \,. \tag{4}$$

- Here $\breve{N}(t_j, z, z'; \hat{\zeta}) = \sum_{i=1}^{n} \pi^{-1}\{t_j, \boldsymbol{Z}_i; \hat{\zeta}\} I\{Z_i(t_j) = z, Z_i(t_{j+1}) = z', t_j < B_i\}$ and $\breve{N}(t_j, z; \hat{\zeta}) = \sum_{i=1}^{n} \pi^{-1}\{t_j, \boldsymbol{Z}_i; \hat{\zeta}\} I\{Z_i(t_j) = z, t_j < B_i\}$.
- Thus we arrive at an inverse probability estimator for $p(t_j, z, z')$.

Current project:

- Searching for methods which provides a consistent estimator of ζ.
- Development of simulation studies to study transition probability estimator performance under similar origin setting.
- Construction of the transition probability estimator under the setting of dissimilar origins.
- Study of theoretical properties and simulation studies based on the new setup.

Thank You.


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


Delaigle, A. and Hall, P. (2013). Classification using censored functional data. *Journal of the American Statistical Association*, **108**, 1269–1283.


Delaigle, A. and Hall, P. (2016). Approximating fragmented functional data by segments of Markov chains. *Biometrika*.


Molenberghs, G., Fitzmaurice, G., Kenward, G.M., Tsiatis, A. and Verbeke, G. (2015). Handbook of missing data methodology. *CRC Press*.