Bayesian Functional Overlapping Clusters

Nicholas Marco   Donatello Telesca

4/26/2021

Functional Data Analysis

- Functional Data Analysis (FDA) focuses methods to analyze the sample path of an underlying continuous stochastic process $Y : \mathcal{T} \to \mathbb{R}$.
- To describe the data, we can use the mean function and the covariance function:

$$\mu(t) = \mathbb{E}(Y(t)), \quad C(s,t) = \mathrm{Cov}(Y(s), Y(t)); \quad s,t \in \mathcal{T}$$

- Functional data often takes the form of:

$$y_i(t_j); \quad j = 1, \ldots, p_i, \; i = 1, \ldots, n, \quad t_j \in \mathcal{T}$$

- Using the common assumption that Y is smooth, we can approximate $y_i(t)$ using basis functions:

$$y_i(t) \approx \sum_{l=1}^{P} \theta_{(i,l)} b_l(t); \quad t \in \mathcal{T}$$

Functional Clustering

- Functional clustering is an unsupervised technique that classifies functional observations into different groups
  - Classically, each observation belongs to exactly one group

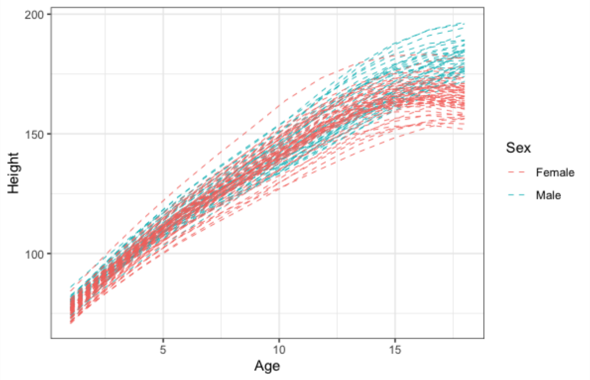

Figure 1: Observations from the Berkeley Growth Study

Overlapping Clustering

- Overlapping Clustering allows for each observation to belong to multiple clusters
  - Examples: Clustering movies by genre

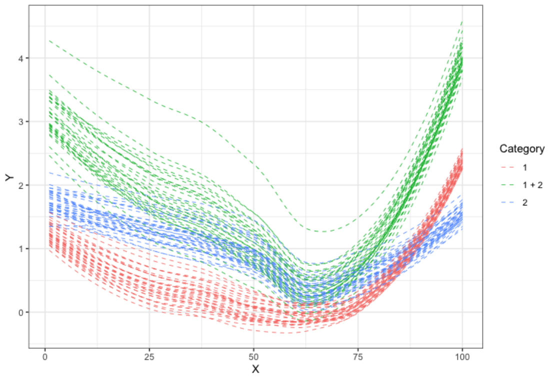

Figure 2: Simulated Data

Overlapping Clusters Setup

- We will assume that the number of clusters (K) is known.
- Let $z_{ik} \in \{0, 1\}$ be a latent variable denoting whether the i^{th} observation is a member of the k^{th} cluster.
- Let $f^{(k)}(.)$ represent the underlying stochastic process associated with the k^{th} cluster.
  - $\mathbb{E}(f^{(k)}(.)) = \mu^{(k)}(.)$
  - $\text{Cov}(f^{(k_1)}(t_{k_1}), f^{(k_2)}(t_{k_2})) = C_{(k_1,k_2)}(t_{k_1}, t_{k_2}); \;\; t_{k_1}, t_{k_2} \in \mathcal{T}$
- We will assume an additive model.
- Assumptions on $f^{(k)}(.)$:
  - $||f^{(k)}(.)||_2 = \left(\int_{\mathcal{T}} |f^{(k)}(t)|^2 dt\right)^{1/2} < \infty$ $(f^{(k)}(.) \in L^2(\mathcal{T}))$
  - $f^{(k)}(.)$ is is a smooth function and is in the P-dimensional subspace $span\{b_1, \dots, b_p\}$ of $L^2(\mathcal{T})$:
    - $f^{(k)}(t) = \sum_{l=1}^{P} \theta_{(k,l)} b_l(t) = \mathbf{B}'(t)\boldsymbol{\theta}_k; \;\; t \in \mathcal{T}$

Estimating the Covariance Functions

- If we assume independence between the underlying stochastic processes, we only have K covariance matrices to estimate.
  - Assumes that the variances are additive.
- Not assuming independence, there are $2^K - 1$ different covariance functions that we have to model.
- Using the additive structure of the model, we have:

$$\begin{aligned}\mathrm{Var}\left(\sum_{k=1}^{K} z_{ik} f^{(k)}(t)\right) &= \sum_{k=1}^{K} z_{ik} \mathrm{Var}(f^{(k)}(t)) \\ &\quad + 2 \sum_{k_1 < k_2} z_{ik_1} z_{ik_2} \mathrm{Cov}(f^{(k_1)}(t), f^{(k_2)}(t))\end{aligned}$$

- Using the additive structure, we can reduce the number of covariance functions that we have to estimate from $2^K - 1$ to $K + \frac{K(K-1)}{2}$.

Multivariate Karhunen-Loève Theorem

- The Multivariate Karhunen-Loève decomposition can be used to jointly represent the K stochastic processes as a linear combination of eigenfunctions.
- Let $f(\mathbf{t}) = (f^{(1)}(t_1), \ldots, f^{(K)}(t_K)) \in \mathbb{R}^K$.
    - Thus we have $f \in \mathcal{H} := L^2(\mathcal{T}) \times \cdots \times L^2(\mathcal{T})$.
- Defining the inner product as:

$$\langle f, g \rangle = \sum_{k=1}^{K} \int_{\mathcal{T}} f^{(k)}(t) g^{(k)}(t) \mathrm{d}t \quad f, g \in \mathcal{H},$$

  we have that $\mathcal{H}$ is a Hilbert space.
- We can define the Covariance Operator, $\mathcal{K}$, in the following way:

$$\left(\mathcal{K}g\right)^{(i)}(\mathbf{t}) = \sum_{k=1}^{K} \int_{\mathcal{T}} C_{(k,i)}(t_k, t_i) g^{(k)}(t_k) \mathrm{d}t_k.$$

Multivariate Karhunen-Loève Theorem (cont.)

- Since $\mathcal{H}$ is a Hilbert space, under some assumptions on $\mathcal{K}$, there exists a complete orthonormal basis of eigenfunctions $\Psi_m \in \mathcal{H}$ and eigenvalues λ_m such that:

$$\mathcal{K}\Psi_m = \lambda_m \Psi_m.$$

- Mercer's Theorem tells us that:

$$\mathrm{Cov}(f^{k_1}(t_{k_1}), f^{(k_2)}(t_{k_2})) = \sum_{m=1}^{\infty} \lambda_m \Psi_m^{(k_1)}(t_{k_1}) \Psi_m^{(k_2)}(t_{k_2})$$

- Using the Multivariate Karhunen-Loève Theorem, we have that:

$$f(\mathbf{t}) = \mu(\mathbf{t}) + \sum_{m=1}^{\infty} \rho_m \Psi_m(\mathbf{t}) \approx \mu(\mathbf{t}) + \sum_{m=1}^{M} \rho_m \Psi_m(\mathbf{t}),$$

where $\mu(\mathbf{t}) = \left(\mu^{(1)}(t_1), \dots, \mu^{(K)}(t_K)\right)$, $\mathbb{E}(\rho_m) = 0$, and $Cov(\rho_m, \rho_n) = \lambda_m \delta_{mn}$.

Multivariate Karhunen-Loève Theorem (cont.)

- We have decomposed the stochastic process into a linear combination of eigenfunctions and mean functions, however we do not know these functions.
- Using the assumption that $f^{(k)}(t) = \mathbf{B}'(t)\boldsymbol{\theta}_k$, we can represent these functions as a finite number of unknown parameters:
  - $\mu^{(k)}(t) = \mathbf{B}'(t)\boldsymbol{\nu}_k; \;\; \boldsymbol{\nu}_k \in \mathbb{R}^P$
  - $\Psi_m^{(k)}(\mathbf{t}) = \mathbf{B}'(t)\tilde{\boldsymbol{\phi}}_{km}; \;\; \tilde{\boldsymbol{\phi}}_{km} \in \mathbb{R}^P$ for $\lambda_m > 0$
- Therefore we have:

$$f^{(k)}(t) \approx \mathbf{B}'(t)\boldsymbol{\nu}_k + \sum_{m=1}^{M} \rho_m \mathbf{B}'(t)\tilde{\boldsymbol{\phi}}_{km}$$

BFOC Model

- Letting $\phi_{km} = \sqrt{\lambda_m}\tilde{\phi}_{km}$, we have our likelihood:

$$y_i(t)|\mathbf{\Theta} \sim \mathcal{N}\left(\sum_{k=1}^{K} z_{ik}\underbrace{\left(\mathbf{B}'(t)\boldsymbol{\nu}_k + \sum_{m=1}^{M}\chi_{im}\mathbf{B}'(t)\phi_{km}\right)}_{f^{(k)}(t)}, \sigma^2\right)$$

- Priors:
  - $p(\boldsymbol{\nu}_k|\tau_k) \propto exp\left\{-\frac{\tau_k}{2}\sum_{h=2}^{P}(\nu_{(h,k)} - \nu_{(h-1,k)})^2\right\}$
    - Prevents overfitting of the mean function
  - We can use the multiplicative gamma process shrinkage prior (Bhattacharya and Dunson 2011)
    - Shrinks the magnitude of ϕ_{km} as m increases
    - $\Psi_m(t)$, $m = 1, \ldots, M$, are no longer orthogonal

Further Work

- Run simulation studies
    - Ensure that we are able to recover the parameters or functions of the paramters
- Speed up convergence of MCMC
    - To speed up convergence of the Markov chain, we can pick "good" initial states
- Expand model to higher dimensional functional data