Functional Data Analysis

Ping-Han Huang

Arizona State University

November 26, 2024

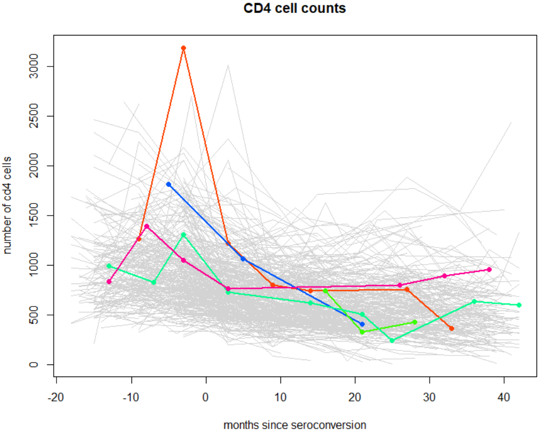


Figure 1: CD4 Data (Source: Staicu & Park, 2016)

Examples of FDA—DTI

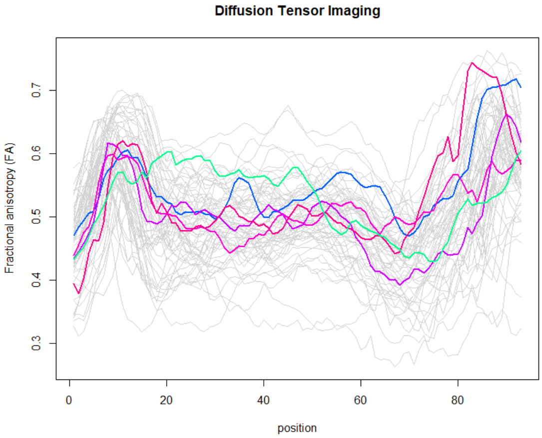


Figure 2: DTI Data (Source: Staicu & Park, 2016)

Stochastic Process

- Underlying random function: $\{X(t); t \in \mathcal{T}\}$
- m i.i.d. sample paths (realizations of random functions): $\{X_i(t); t \in \mathcal{T}\}$
- Subsamples of m sample paths: $x_i(t_{ij})$, $i = 1, ..., m$ and $j = 1, ..., n_i$

Second-Order Process

- $\{X(t); t \in \mathcal{T}\}$ is a second-order process if, for each t, $X(t)$ has finite second moment, i.e.,

$$E|X(t)|^2 < \infty$$

  - Continuous mean function:

$$\mu_X(t) = E\{X(t)\}$$

  - Continuous and nonnegative definite covariance function:

$$\Gamma_X(s,t) = Cov\{X(s), X(t)\}, \text{for all } s, t \in \mathcal{T}$$

Functional Principal Component Analysis (fPCA)

- Reduce dimensionality
- Capture main modes of variation
- Express $X(t)$ as

$$X(t) = \mu_X(t) + \sum_{k=1}^{K} \zeta_k \phi_k(t)$$

where ζ_k is the kth FPC score and ϕ_k is the kth eigenfunction

PCA for Multivariate Data

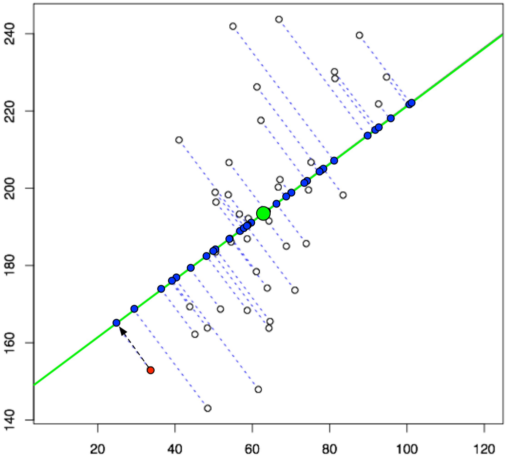


Figure 3: PCA (Source: Pachter, 2014)

PCA for Multivariate Data

1. The data is $\vec{X} = (X_1, ..., X_m)^T$
2. Eigen decomposition of $Cov(\vec{X})$ to get eigenvectors $\mathbf{\Phi}$ and eigenvalues $\vec{\lambda}$

$$Cov(\vec{X}) = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^T = \sum_{m=1}^{M} \lambda_m \phi_m \phi_m^T$$

3. Obtain

$$\boldsymbol{Y} = \boldsymbol{P}\vec{X}_c = \sum_{m=1}^{M} [\phi_m^T \vec{X}_c] \phi_m$$

- $\vec{X}_c = \vec{X} - \boldsymbol{\mu_X}$
- $\boldsymbol{P} = \Phi(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ is the projection matrix
- $\boldsymbol{Y}$ is the re-representation of the data

- Then we have $\vec{X} = \boldsymbol{\mu}_X + \boldsymbol{\Phi}\zeta$, where $\zeta = \boldsymbol{\Phi}^T \vec{X_c}$.
  - $\zeta = (\zeta_1, ..., \zeta_m)^T$
- $\zeta_m = \phi_m^T(\vec{X} - \boldsymbol{\mu}_X)$
  - $E(\zeta_m) = 0$
  - $Var(\zeta_m) = \lambda_m$
  - $Cov(\zeta_m, \zeta_m') = 0$
- The principal component scores are rank-ordered by their variances

From PCA to fPCA

- Mercer's Theorem

$$\Gamma_X(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t), \text{ for all } s,t \in \mathcal{T}$$

  - λ_k: mth eigenvalue of $X(t)$
  - $\phi_k(t)$: mth eigenfunction of $X(t)$
- Karhunen-Lóeve Representation

$$X(t) = \mu_x(t) + \sum_{k=1}^{\infty} \zeta_k \phi_k(t)$$

  - $\zeta_k = \int_{\mathcal{T}} [X(t) - \mu_x(t)]\phi_k(t)\,dt$: kth FPC score for $X(t)$
  - $E(\zeta_k) = 0, var(\zeta_k) = \lambda_k, cov(\zeta_k, \zeta_{k'}) = 0$

Number of FPC (K)

- Fraction of variation explained (FVE)
  - $FVE = \frac{\sum_{k=1}^{K} \lambda_k}{\sum_{k=1}^{\infty} \lambda_k}$
- Information criteria
  - AIC
  - BIC
- Cross validation (CV)
  - Minimize the cross-validation score based on the one-curve-leave-out squared prediction error:

$$CV(K) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \{ Y_{ij} - \hat{Y}_i^{(-i)}(T_{ij}) \}^2$$

After fPCA and selecting the number of FPCs, we can recover the trajectory $\hat{X}_i(t)$ for the ith subject as

$$\hat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\zeta}_{ik}\hat{\phi}_k(t)$$

The estimation is based on noisy observations $\{(Y_{i1}, t_{i1}), ..., (Y_{in_i}, t_{in_i})\}$, where

$$Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$$

Comparisons of LMEM and FDA

Observed data: $\{(Y_{i1}, t_{i1}), ..., (Y_{in_i}, t_{in_i})\}$

- LMEM: $Y_i = \boldsymbol{X}_i \vec{\beta} + \boldsymbol{Z}_i \vec{b}_i + \vec{e}_i$
  - parametric assumptions for the model
  - parametric methods for estimation
  - objective: inference
- FDA: $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$
  - no assumption for the model covariance
  - nonparametric approach for estimation
  - objective: recovering subject-specific trajectories

R Demonstration

- Functional Linear Regression Models (FLR)
  - Scalar-on-Function Regression

$$Y_i = \alpha + \int \beta(t) X_i(t)\, dt + \epsilon_i$$

  - Function-on-Scalar Regression

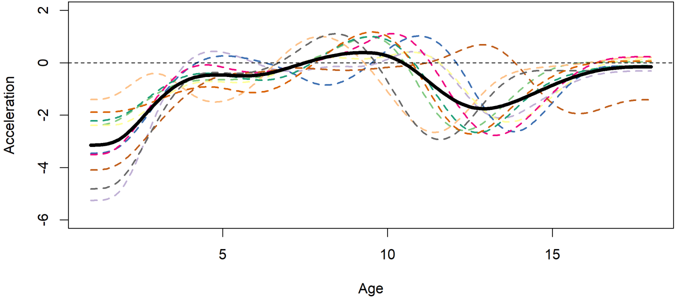
$$Y_i(t) = \beta_0(t) + \sum_{j=1}^{p} \beta_i(t) X_{ij} + \epsilon_i(t)$$

  - Function-on-Function Regression

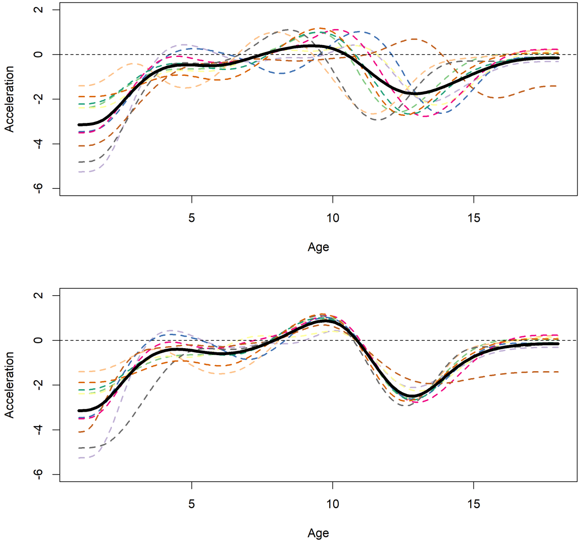
$$Y_i(t) = \beta_0(t) + \int \beta(s,t) X_i(t)\, dt + \epsilon_i(t)$$

Extensions: Curve Registration

- Phase displacement or amplitude variability in the data.
- Align the curves through time warping.

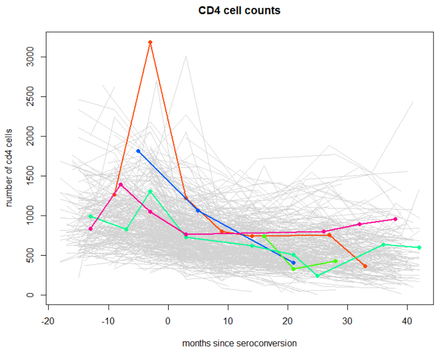


Extensions: Curve Registration



Extensions: Sparse Data

- Principal Component Analysis through Conditional Expectation (PACE) Method for Sparse Data
  - We have “sparse” data when the number of measurements per subject (n_i) is very low.



References

Textbook:

- Hsing, T., & Eubank, R. (2015). Theoretical foundations of functional data analysis, with an introduction to linear operators (Vol. 997). John Wiley & Sons. `https://onlinelibrary.wiley.com/doi/book/10.1002/9781118762547`
- Kokoszka, P., & Reimherr, M. (2017). Introduction to Functional Data Analysis (1st ed.). Chapman and Hall/CRC. `https://doi.org/10.1201/9781315117416`
- Ramsay, J., Hooker, G., Graves, S. (2009). Functional Data Analysis with R and MATLAB. Use R. Springer, New York, NY. `https://link.springer.com/book/10.1007/978-0-387-98185-7`

Paper:

- Yao, F., Müller, H.-G., & Wang, J.-L. (2005). Functional data analysis for SPARSE LONGITUDINAL DATA. Journal of the American Statistical Association, 100(470), 577–590. `https://doi.org/10.1198/016214504000001745`
- J. S. Marron. James O. Ramsay. Laura M. Sangalli. Anuj Srivastava. "Functional Data Analysis of Amplitude and Phase Variation." Statist. Sci. 30 (4) 468 - 484, November 2015. `https://doi.org/10.1214/15-STS524`

Online Lecture:

- Ana-Maria Staicu & So Young Park (2016). Short course on Applied Functional Data Analysis. Retrieved from `https://www4.stat.ncsu.edu/~staicu/FDAtutorial/index.html`
- Short course on functional data analysis. YouTube. `https://youtube.com/playlist?list=PLD2RXrMBJWfOEmYmYE5xlB1ARqbZrc0h9&feature=shared`

Thank you!