Statistical Analysis of Functional Data

Anirvan Chakraborty

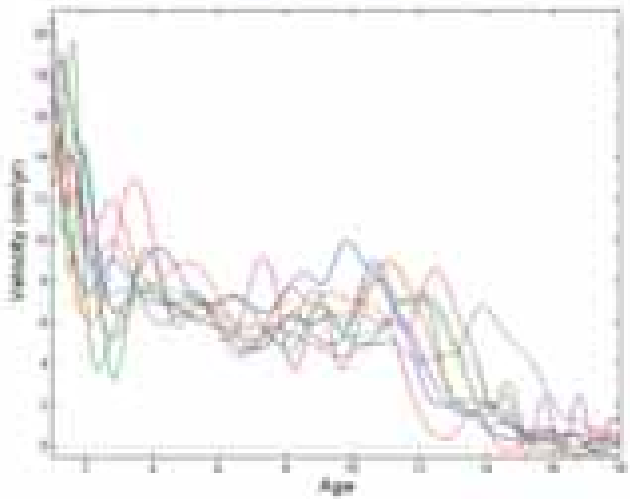
85th Annual Meeting

Indian Academy of Sciences, Bengaluru

- Q. What is FUNCTIONAL DATA ANALYSIS (FDA)?
- Broadly speaking, statistical analysis of data which can be looked upon as "curves", "surfaces" etc. is called functional data analysis. These are the atoms of FDA.
- Examples:
  - Growth curves of boys and girls
  - Temperature curves over a year for various weather stations
  - Hip and knee angles in the sagittal plane over time through a gait cycle
  - Trace of the tip of the pen while writing a letter/word, e.g., "fda"
  - ... (the list goes on and on)
- One could also have more complex data, e.g., functional time series (where the value at each time point is itself a function), spatially and temporally varying curves, images of the brain obtained from an MRI etc.

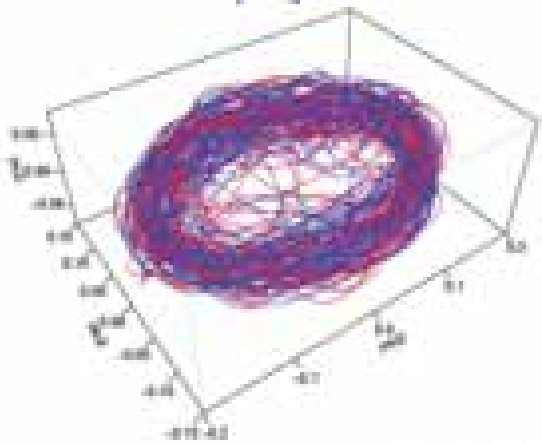
Functional data analysis (contd.)

Genetics



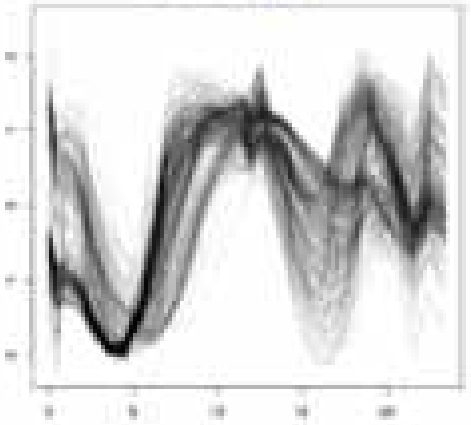
Handwriting Analysis

Biophysics



Brain Imaging

Energy



Environmetrics

- Although each curve/surface/image can only be recorded at finitely many points on the domain, still, these atoms of FDA are intrinsically infinite dimensional.
- Need to consider such data as elements of appropriate infinite dimensional spaces - involves new challenges.
- Indeed, many well-known statistical notions and techniques may break down completely! Need to develop a new branch of mathematical statistics.
- Aspects that play crucial roles are broadly:
  - Geometry: The atoms are nice geometrical objects with an inherent ordering.
  - Randomness: Being objects in an infinite dimensional space, there is a need to understand notions of probability distributions and statistical models in those spaces.
  - Complexity: Often, there is a natural structure that expresses the features of interest, which may be low dimensional.

- Aim: Probe the law of the random function $\{X(t) : t \in [a, b]\}$.
- Data: n i.i.d. realizations $\{X_i(t) : t \in [a, b]\}_{i=1}^n$
- Setup: Assume $X_i \in L^2[a, b]$ almost surely for all $i = 1, 2, \ldots, n$.

Mean function

Assume $\mathbb{E}(||X||) < \infty$. Define $m(t) = \mathbb{E}[X(t)]$. Then, $m \in L^2[a, b]$.

$\hookrightarrow$ Characterizes "location" of the random function.

Covariance kernel

Assume $\mathbb{E}(||X||^2) < \infty$. Define $k(t, s) = \mathrm{Cov}(X(t), X(s))$. Then, $k \in L^2([a, b] \times [a, b])$.

$\hookrightarrow$ Characterizes the fluctuations of the random function around its mean
$\hookrightarrow$ Still uncountably infinite! Reduce to countable variation?

Assume that $\mathbb{E}(||X||^2) < \infty$. Then,

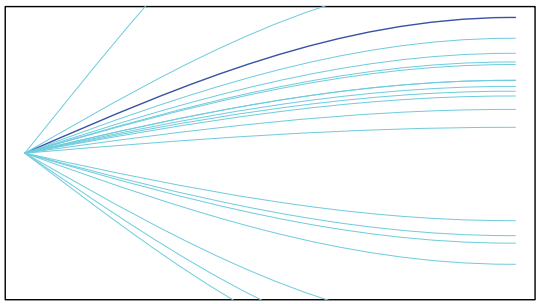
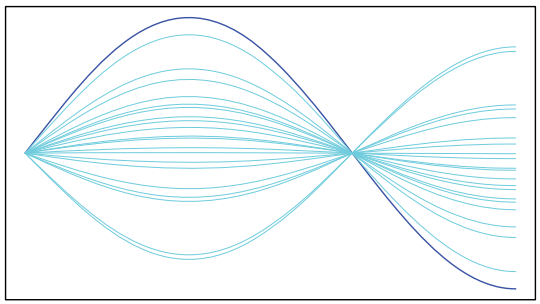
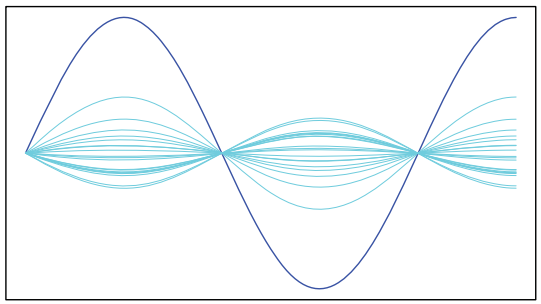
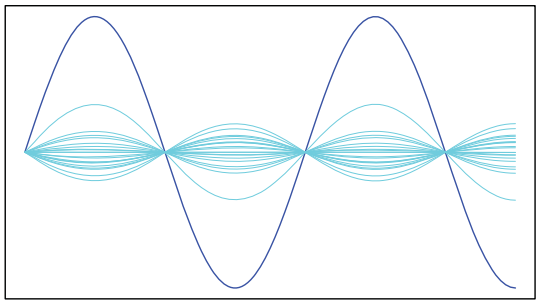
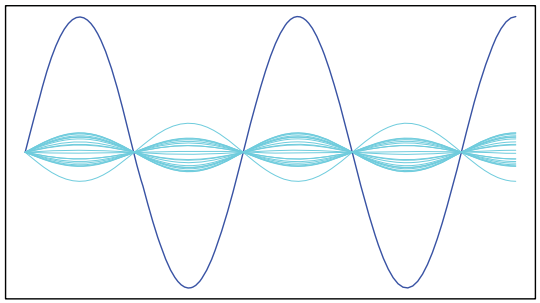
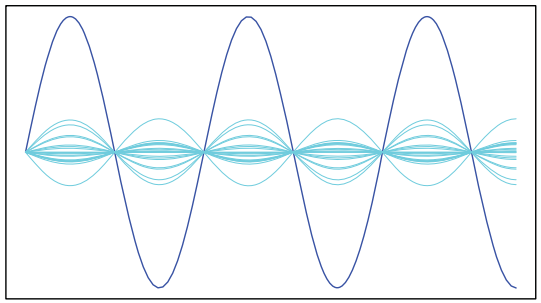
$$X(t) - m(t) = \sum_{n=1}^{\infty} \xi_n \phi_n(t)$$

where $\xi_n = \langle X - m, \phi_n \rangle$, called the *principal component scores*, are zero mean uncorrelated rvs with variance λ_n.

Captures **complete curve dynamics** - canonical FDA framework:

- Separation of variables (stochastic versus functional)
- Quantification of smoothness
  $\hookrightarrow$ ϕ_n contributes as $\lambda_n / \sum_j \lambda_j$
  $\hookrightarrow$ rate of decay of λ_n
- Variance components / functional fluctuations around mean
- Optimal finite dimensional representation
  (modelling/methodology + inference/regularization)

$$X(t) = \sum_{n=1}^{\infty} \underbrace{\xi_n \sqrt{2}\, \sin\left(\left(k - \frac{1}{2}\right)\pi t\right)}_{Z_k(t)}, \qquad \xi_n \overset{indp}{\sim} \mathsf{N}\left(0, \frac{1}{\left(k - \frac{1}{2}\right)^2 \pi^2}\right)$$

1947/49 Independent introduction by Karhunen & Loéve

$\hookrightarrow$ Linear filtering of stochastic process and series representation

1950 Ulf Grenander shows importance in statistics (birth of FDA?)

$\hookrightarrow$ Uses as coordinate representation for likelihood ratios

1958 C. R. Rao hints potential usefulness for growth curves

$\hookrightarrow$ Components of variance interpretation

$\vdots$

1973 Kleffe considers empirical version $(n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}) \otimes (X_i - \overline{X}))$

$\hookrightarrow$ Large sample convergence

1982 Dauxois, Pousse & Romain develop asymptotics of empirical version

1986 Besse & Ramsay (psychometrics) use as PCA

1991 Rice & Silverman: "Estimating Mean and Covariance when data are curves"

Subject then takes off...

- Let y be a random variable and X be a random element in a separable Hilbert space $\mathcal{H}$.
- Given observations $\{(y_i, X_i) : 1 \leqslant i \leqslant n\}$, we intend to fit the functional linear model (with a scaler response and a functional covariate)

$$y = \alpha + \langle X, \beta \rangle + \epsilon$$

  to the data, where $\beta \in \mathcal{H}$ is the slope function, $\alpha \in \mathbb{R}$ is the intercept, and ϵ is a random error term independent of X and is assumed to be zero mean.
- Typically, β is the main parameter of interest. Once we estimate β by $\widehat{\beta}$, then once can estimate α by $\widehat{\alpha} := \overline{y} - \langle \overline{X}, \widehat{\beta} \rangle$.

- Method of Least Squares leads to the linear system

$$\hat{\mathcal{K}}\beta = \hat{C},$$

where $\hat{C} = n^{-1}\sum_{i=1}^{n}(y_i - \overline{y})(X_i - \overline{X})$, and
$\hat{\mathcal{K}} = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}) \otimes (X_i - \overline{X})$.

- The infinite dimensional nature of the problem implies that the linear systems are ill-posed, and thus the estimation of β is an ill-posed inverse problem.
- Specifically, solving $\hat{\mathcal{K}}\beta = \hat{C}$ is equivalent to setting

$$\hat{\beta} = \sum_{j=1}^{n-1} \hat{\lambda}_j^{-1} \langle \hat{C}, \hat{\phi}_j \rangle \hat{\phi}_j.$$

- Roughly, can only include those j's such that the growth of λ_j^{-1} can be "controlled" by n^{-1}
  $\hookrightarrow$ delicate balance between sample size n and the rate of decay of eigenvalues.
- However, cannot take only finitely many terms since this would induce asymptotically non-negligible bias.
- Bias-variance trade-off crucially depends on rate of decay of eigenvalues as well as the number of terms contributing to the expression of $\widehat{\beta}$ above
  $\hookrightarrow$ need appropriate regularization
- Broadly, regularization in functional linear models can be categorized into
  (a) Sieve methods $\longrightarrow$ Spectral tuncation
  (b) Penalization methods $\longrightarrow$ Tikhonov regularization

- Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two samples from probability measures P_1 and P_2 (equivalently, two populations Π_1 and Π_2) on a Hilbert space $\mathcal{H}$.
- Problem: Given a new observation Z, how do we determine whether Z belongs to Π_1 or Π_2?
- In the multivariate setting, the "optimal" classifier is called the Bayes classifier, and is given by

$$\text{Classify } Z \text{ to } \Pi_1 \text{ if } f_P(Z)/f_Q(Z) > c \text{ for a suitable } c$$

- However, for functional data, f_P and f_Q are not obtainable in most cases.
- Q. What are the ways around this problem?

- Suppose that the two functional populations differ only in their means.
- A simple classifier would be to the "centroid classifier" constructed by projecting the data onto a single pre-chosen direction.
- Assuming $\mu_1 = \mu$ and denoting $\mu_2 = 0$, we

$$\text{Classify } Z \text{ to } \Pi_1 \text{ if } (\langle Z - \mu, \psi \rangle)^2 - (\langle Z, \psi \rangle)^2 < 0,$$

  where ψ is the pre-chosen projection direction.
- So, the minimum value (over choices of ψ) of the total misclassification error is

$$1 - \Phi \left[\frac{1}{2} \sum_{j=1}^{\infty} \lambda_j^{-1} m_j^2 \right].$$

- If $\sum_{j=1}^{\infty} \lambda_j^{-2} m_j^2 < \infty$, the minimum is achieved with

$$\psi \propto \sum_{j=1}^{\infty} \lambda_j^{-1} m_j \phi_j = \mathcal{K}^{-1} \mu.$$

- If $\sum_{j=1}^{\infty} \lambda_j^{-2} m_j^2 = \infty$, then the minimum cannot be attained since there is no valid $\psi \in \mathcal{H}$ satisfying the above condition. If, further, $\sum_{j=1}^{\infty} \lambda_j^{-1} m_j^2 < \infty$, then the minimum misclassification error is strictly positive.
- The minimum error is *zero* if $\sum_{j=1}^{\infty} \lambda_j^{-1} m_j^2 = \infty$
  $\hookrightarrow$ perfect classification is achieved!
- Classifier based on the likelihood ratio test is equivalent to applying the centroid based classifier.
- The *perfect classification* also holds under non-Gaussianity.

In theory: one observed complete curves $X_1, X_2, \ldots, X_n$

In practice: one observed each curve discretely with potential measurement error

$$W_{ij} = X_i(T_{ij}) + \epsilon_{ij}, \quad , j = 1, 2, \ldots, N_i, \ i = 1, 2, \ldots, n,$$

where

- $\{T_{ij} : j = 1, 2, \ldots, N_i\}$ can be a deterministic or random grid over $[a, b]$ for each i.
  $\hookrightarrow$ each set of grid points may be completely different from the other sets of grid points
- ϵ_{ij}'s are measurement errors and are assumed to be independent of the X_i's. They are themselves i.i.d. with zero mean and variance σ^2.

- Generally, if $\min_{1 \leqslant i \leqslant n} N_i$ grows to infinity with n, then the data is called "densely observed".
- On the other hand, if $\max_{1 \leqslant i \leqslant n} N_i$ stays bounded as $n \to \infty$, the data is called "sparsely observed".
- Issues:

  (a) We do not observe the true functional data, and the observations are contaminated with error. So, it may not possible to directly estimate population parameters.

  (b) Since the grid may differ across observations, even simple estimators like taking the mean of the observations may not be useful and appropriate (may get only one observation per time point!).

  (c) Estimation procedure and their performance will heavily depend on the behaviour of the grid – different performance for dense and sparse sampling.

Two major approaches:

(1) First find smooth estimates of each curve X_i from the discrete set of observations $\{W_{ij} : j = 1, 2, \ldots, N_i\}$, using some smoothing technique.
Then, use the smoothed estimates (functions) to carry out other analyses – leads to independent estimates of curves

and the converse approach,

(2) Using the pooled data, first find an estimate of the covariance kernel. Then use the estimated covariance kernel to provide estimates of the true curves – leads to dependent estimates of curves
$\hookrightarrow$ PACE approach

- Dense sampling: $\min_{1 \leqslant i \leqslant n} N_i \gg n^{1/4}$ as $n \to \infty$

  Under appropriate smoothing and other conditions, the estimators of μ, K, λ_j's and ϕ_j's obtained under approaches (1) and (2) are $\sqrt{n}$-consistent – parametric rate of convergence.

  $\hookrightarrow$ as good as observing the fully functional data.

- Sparse sampling: $\max_{1 \leqslant i \leqslant n} N_i \leqslant C < \infty$ as $n \to \infty$

  (a) Approach (1) will be inconsistent since there will be non-negligible bias in the estimates of the individual curves.

  (b) Approach (2) will be consistent under certain assumptions provided that the sampling points are randomly distributed over the entire of $[0, 1]$, or else, it will also be inconsistent like approach (1).

  (c) Estimators of μ, K and ϕ_j's have slower (non-parametric) rates of convergence. However, the estimators of the λ_j's continue to be $\sqrt{n}$-consistent.

Thank you!