函數型資料分析簡介
Introduction to Functional Data Analysis

李百靈 淡江大學 統計學系

邱俊捷 中研院統計所 (TA)

2025統計研習營 16 July 2025

Outline

1. Introduction
2. Representing Functional Data
3. Exploratory Data Analysis (EDA)
4. The fda Package
5. Functional Linear Models

References

- **For this 3-hour lecture:**
  - Functional Data Analysis, 2nd ed., Ramsay & Silverman, 2005.
    https://link.springer.com/book/10.1007/b98888
  - Functional Data Analysis with R and MATLAB, Ramsay, Hookder, and Graves, 2009.
    https://link.springer.com/book/10.1007/978-0-387-98185-7
  - R package: fda
    https://cran.r-project.org/web/packages/fda/index.html
- **Other reference:**
  - Functional Data Analysis with R, 1st ed., Ciprian M. Crainiceanu, Jeff Goldsmith, Andrew Leroux, and Erjia Cui, 2024.
    https://functionaldataanalysis.org/

What are functional data?

- There is actually an increasing number of situations coming from different fields of applied sciences (environmetrics, chemometrics, biometrics, medicine, econometrics, . . .) in which the collected data are *curves.*

- Functional data is multivariate data with an ordering on the dimensions.

  - Key assumption is *smoothness*:

$$y_{ij} = x_i(t_{ij}) + \varepsilon_{ij},$$

where t in a continuum (usually time), and $x_i(t)$ is smooth.

Data on the Growth of Girls

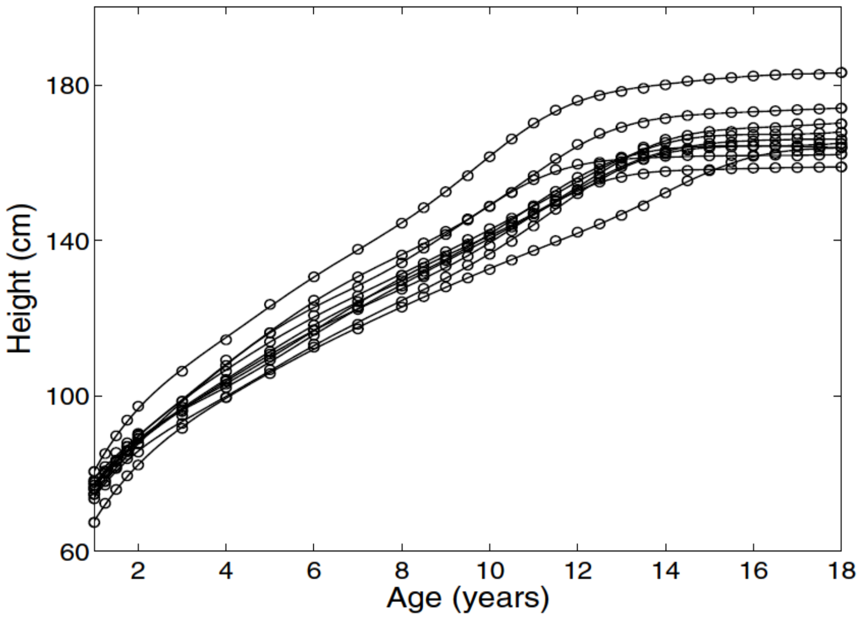


Figure 1 The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

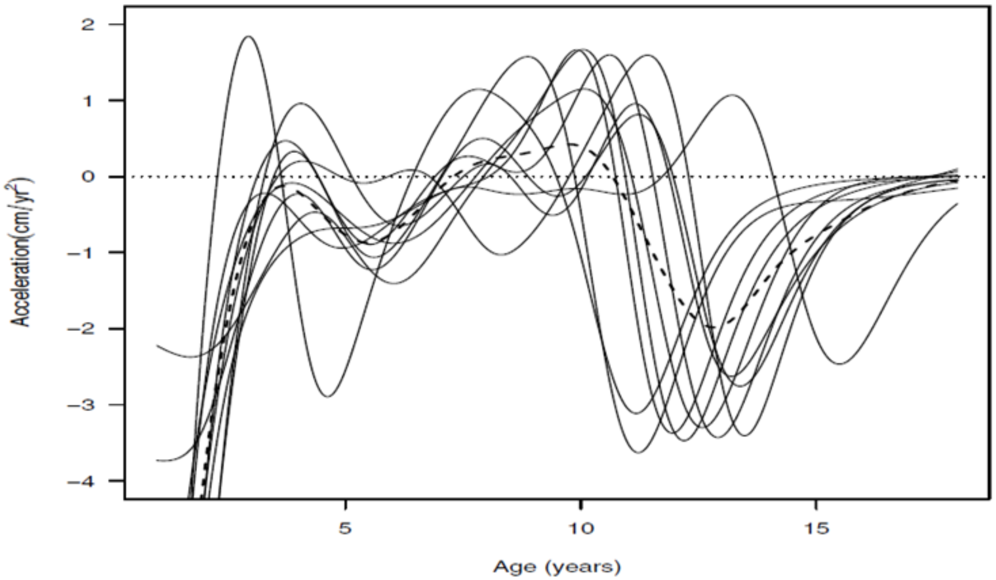
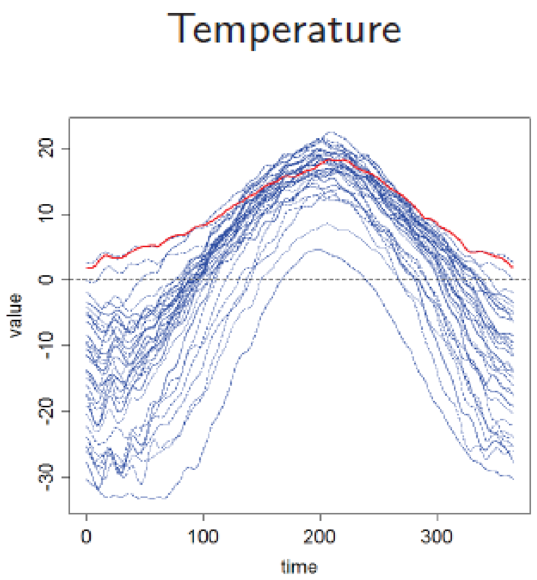


Figure 2 The estimated accelerations of height for 10 girls, measured in centimeters per year. The heavy dashed line is the cross-sectional mean and is a rather poor summary of the curves.

Weather In Canada



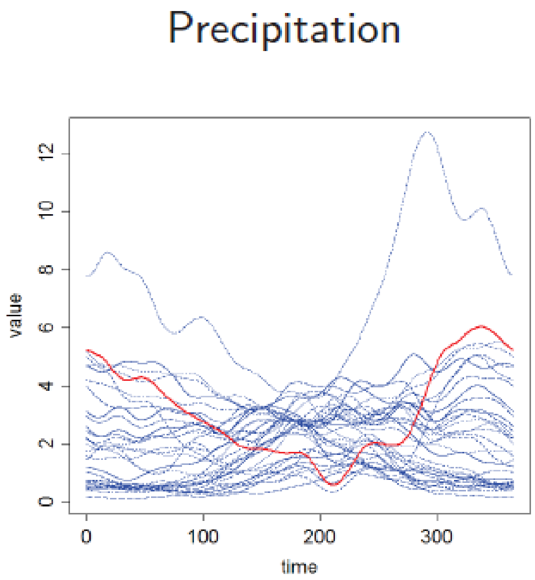


Figure 3 Average daily temperature and precipitation records in 35 weather stations across Canada.

Handwriting Data

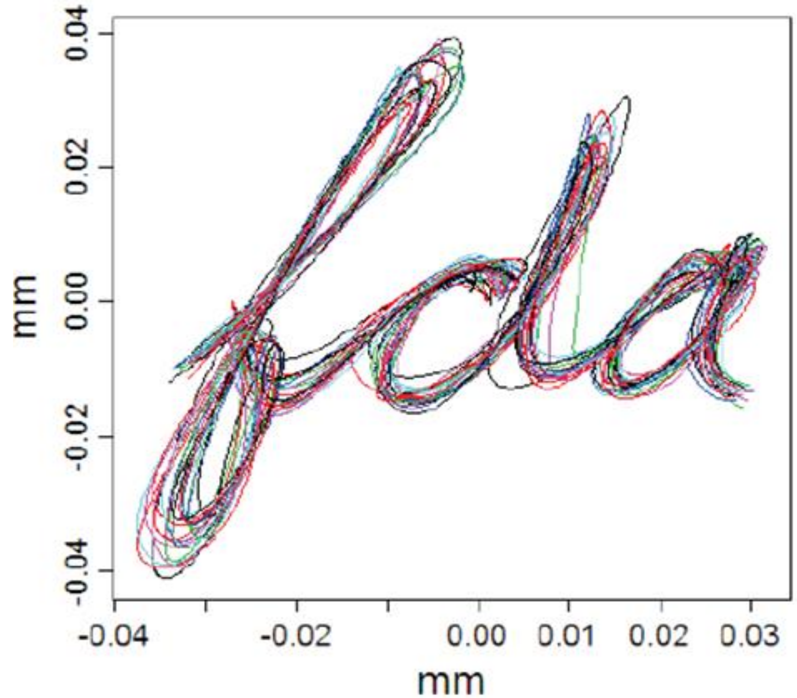


Figure 4 Measures of position of nib of a pen writing "fda". 20 replications, measurements taken at 200 hertz.

Traffic Data

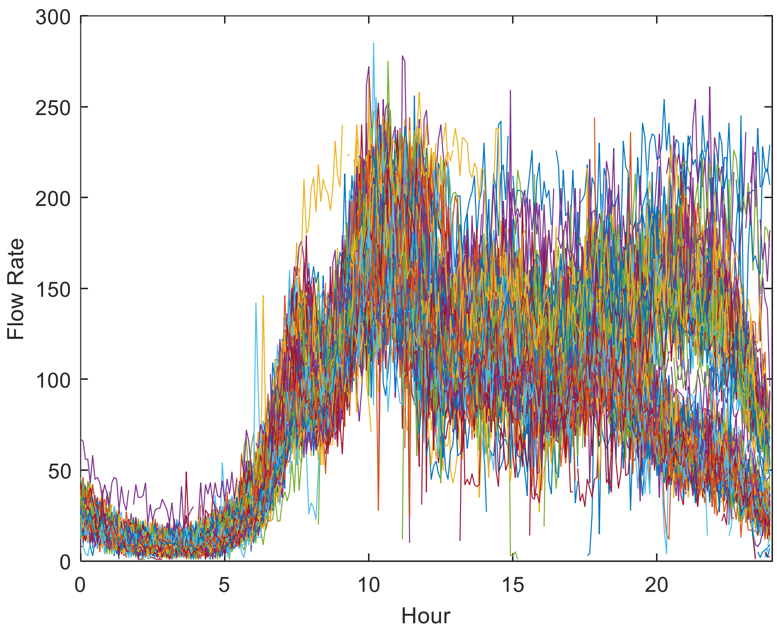


Figure 5 Daily traffic flow rate trajectories for 240 non-holidays.

Modelling

- **Functional variable**

  A random variable X is called functional variable if it takes values in an infinite dimensional space (or functional space).

$$X = \{X(t), t \in T\}$$

  An observation x of X is called a functional data.

- **Functional datasets**

  A functional dataset $x_1, \dots, x_n$ is the observation of n functional variables $X_1, \dots, X_n$ identically distributed as X.

- **Functional Modeling**

  Assume that the data are realizations of n independent random functions $\{X_i(t), t \in T\}, i = 1, \cdots, n$, over an entire interval T.

Modelling

➢ **Longitudinal Data**

$\{(t_{ij}, y_{ij}), 1 \le j \le m_i, 1 \le i \le n\}$

$$y_{ij} = x_i(t_{ij}) + \varepsilon_{ij}$$

t_{ij}: the ith recording time of the jth subject

y_{ij}: the measurement of the ith subject observed at t_{ij}.

➢ **Sampling design**

- regular or irregular
- densely or sparsely sampled

Weather In Vancouver

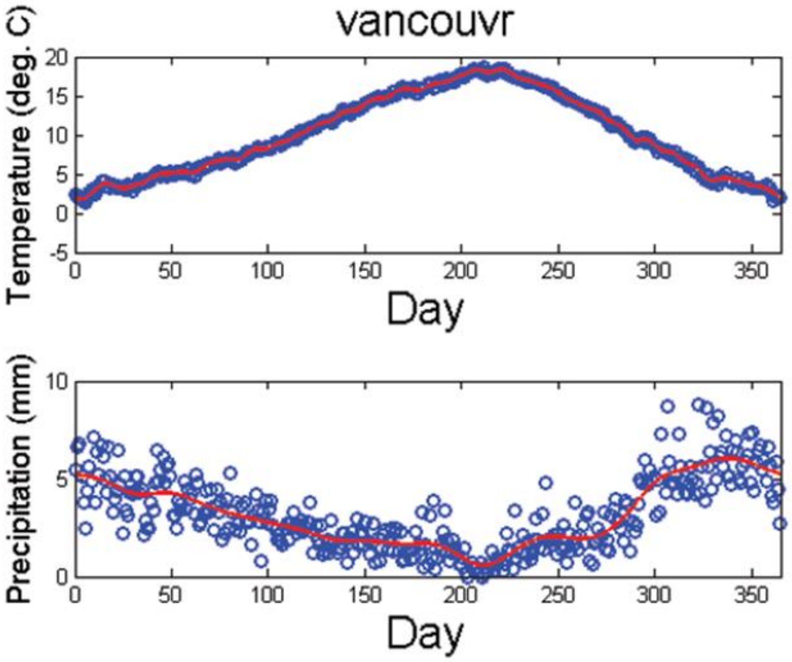


Figure 6 Measure of climate: daily precipitation and temperature in Vancouver, BC averaged over 40 years.

The Goals of FDA

- Represent the data in ways that aid further analysis.
- Display the data so as to highlight various characteristics.
- Study important sources of patterns and variation in the data.
- Explain variation in an outcome or dependent variable by using input or independent variable information
- Compare two or more sets of data with respect to certain types of variation, where two sets of data can contain different sets of replicates of the same functions, or different functions for a common set of replicates.

The First Steps in FDA

- **Data representation: smoothing and interpolation**

  If the discrete values are assumed to be errorless, then the process is interpolation, but if they have some observational error that needs removing, then the conversion from discrete data to functions may involve smoothing.

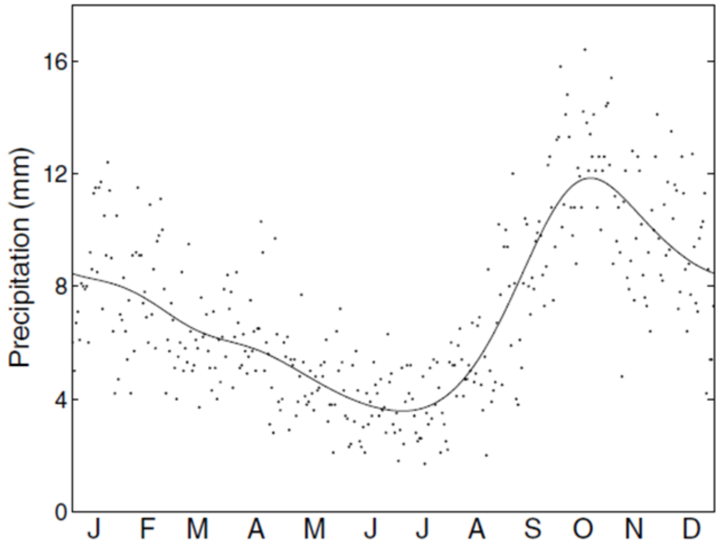


Figure 7 The points indicate average daily rainfall at Prince Rupert on the northern coast of British Columbia. The curve was fit to these data using a roughness penalty method..

The First Steps in FDA

- **Data registration or feature alignment**

  The start of the pinch is located arbitrarily in time, and a first step is to align the records by some shift of the time axis.

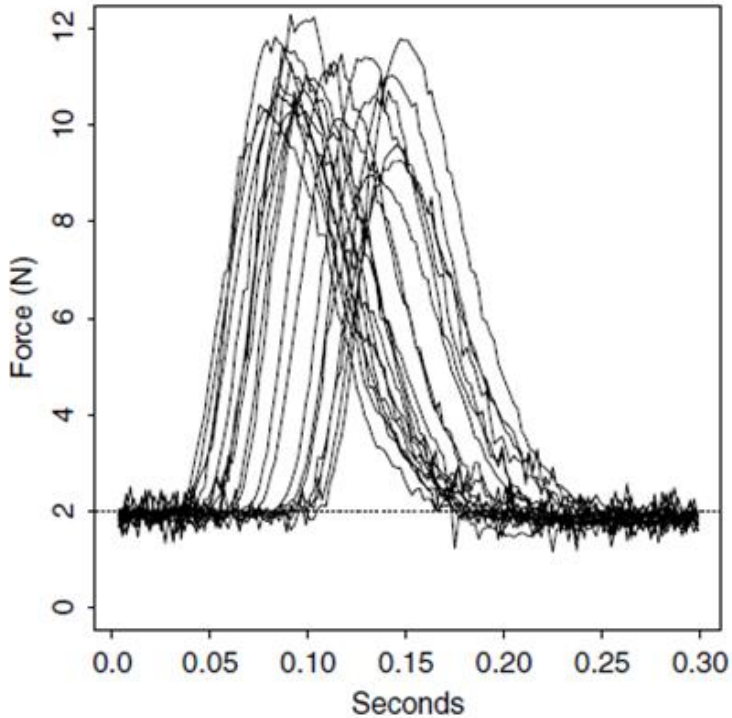


Figure 8 Twenty recordings of the force exerted by the thumb and forefinger where a constant background force of two newtons was maintained prior to a brief impulse targeted to reach 10 newtons. Force was sampled 500 times per second.

The First Steps in FDA

- **Data display**

Displaying the results of a functional data analysis can be a challenge.

Different displays of data can bring out different features of interest, and that the standard plot of $x(t)$ against t is not necessarily the most informative.

It is impossible to be prescriptive about the best type of plot for a given set of data or procedure.

- **Plotting pairs of derivatives**

Helpful clues to the processes giving rise to functional data can often be found in the relationships between derivatives.

The First Steps in FDA

➢ **Example** - Mean temperature at Montreal

In Figure 9, casual inspection does indeed suggest a strongly sinusoidal relationship between mean temperature and month, but the right panel shows that things are not so simple.

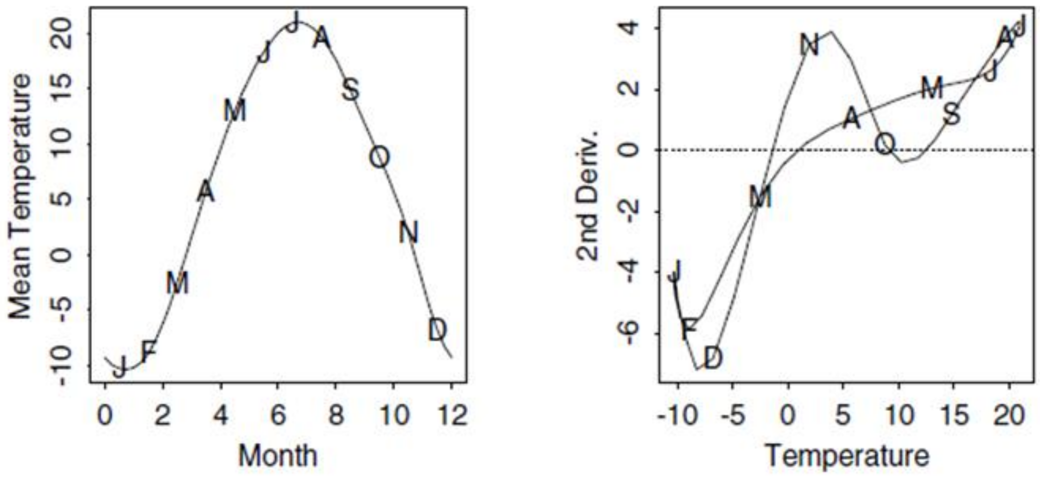


Figure 9 The left panel gives the annual variation in mean temperature at Montreal. The times of the mid-months are indicated by the first letters of the months. The right panel displays the relationship between the second derivative of temperature and temperature less its annual mean. Strictly sinusoidal or harmonic variation in temperature would imply a linear relationship.

Exploring Variability in Functional Data

- **Interests**

  1. Representations of the distribution of functions

     Mean, variation, covariation

  2. Relationships of functional data to

     Covariates, responses, and other functions

  3. Relationships between derivatives of functions.

- **Some Methods**

  Functional descriptive statistics

  Functional principal components analysis

  Functional canonical correlation

  Functional linear models

Exploring Variability in Functional Data

- **What are the challenges?**

1. Estimation of functional data from noisy, discrete observations.
2. Numerical representation of infinite-dimensional objects
3. Representation of variation in infinite dimensions.
4. Description of statistical relationships between infinite dimensional objects.
5. $n < p = \infty$, and use of smoothness.
6. Measures of variation in estimates.

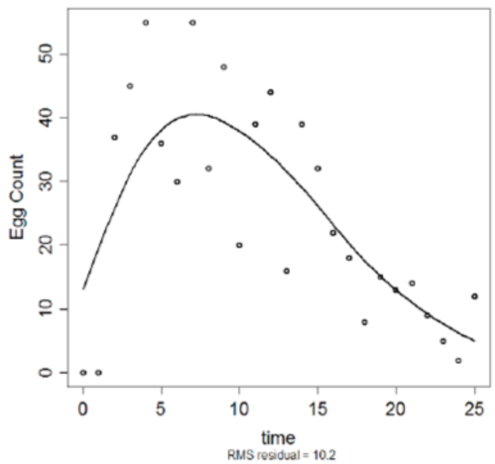
…

Representing Functional Data

- **From discrete to functional data**

Represent data recorded at discrete times as a continuous function in order to

- Allow evaluation of record at any time point.
- Evaluate rates of change.
- Reduce noise.
- Allow registration onto a common time-scale.



$$y_i = x(t_i) + \varepsilon_i, i = 1, \dots, n.$$

Representing Functional Data

- **Method 1: Representing functions by basis functions**

  - *Basis-expansion methods* (e.g. regression spline smoothing)

    Basic function procedures represent a function x by a linear expansion

    $$x(t) = \sum_{k=1}^{K} c_k \phi_k(t), \qquad (1)$$

    in terms of K known basis functions $\phi_k(t)$.

    - Several basis systems available: e.g. Fourier and B-splines
    - Matrix notation : $\vec{x} = \vec{c}'\vec{\phi}$
    - Smoothing the functional data by least squares

$$SMSSE(y|c) = \sum_{j=1}^{n}\left[y_j - \sum_{k=1}^{K} c_k \phi_k(t)\right]^2 = (y - \Phi c)'(y - \Phi c) = \|y - \Phi c\|^2$$

$$\hat{y} = \Phi\hat{c} = \Phi(\Phi'\Phi)^{-1}\Phi' y$$

Representing Functional Data

- **Method 2: Reducing noise in measurements**

  - *Smoothing penalties* (e.g. spline smoothing)

    Smoothing the functional data with a roughness penalty.

$$\vec{c} = argmin \sum_{i=1}^{n} (y_i - x(t))^2 + \lambda \int [Lx(t)]^2 \, dt \qquad (2)$$

in terms of K known basis functions $\phi_k(t)$.

- $Lx(t)$: measures roughness of x
- λ: a smoothing parameter that trade-off fit to the y_i and roughness; must be chosen. (GCV)

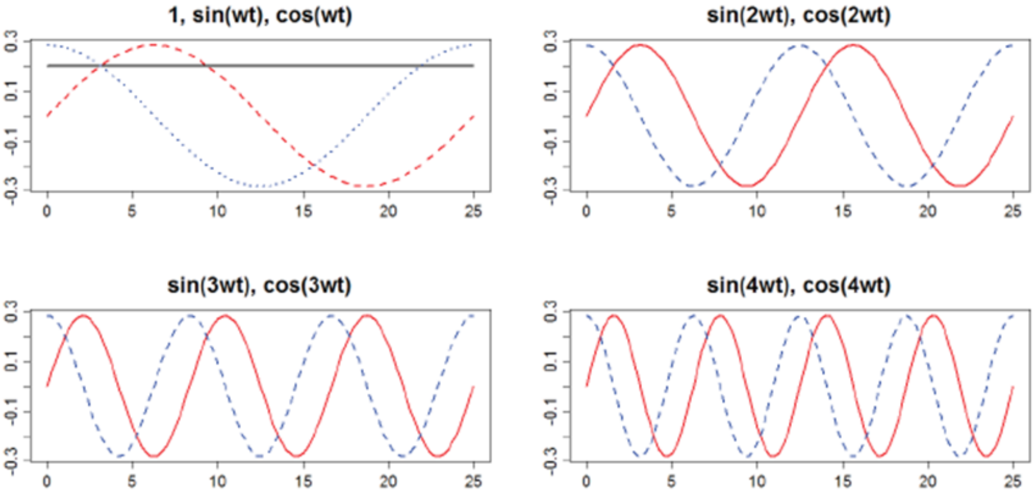
$$PENSSE_\lambda(x|y) = [y - x(t)]' W [y - x(t)] + \lambda PEN_2(x). \qquad PEN_2(x) = \int [D^2 x(s)^2] ds.$$

$$PENSSE_m(y|c) = [y - \Phi c]' W [y - \Phi c] + \lambda c' R c. \qquad \hat{c} = (\Phi' W \Phi + \lambda R)^{-1} \Phi' W y$$

Representing Functional Data

- The Fourier basis system for periodic data

$$1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), \ldots, \sin(\alpha\omega t), \cos(\alpha\omega t), \ldots$$



Constant α=$2\pi/P$ defines the period P of oscillation of the first sine/cosine pair.

Representing Functional Data

- **The spline basis system for open-ended data**
  - Spline functions are the most common choice of approximation system for non-periodic functional data or parameters.
  - Splines combine the fast computation of polynomials with substantially greater flexibility, often achieved with only a modest number of basis functions.
  - B-splines are a particularly useful means of incorporating the constraints. (Reference: de Boor, 2001, "A Practical Guide to Splines")

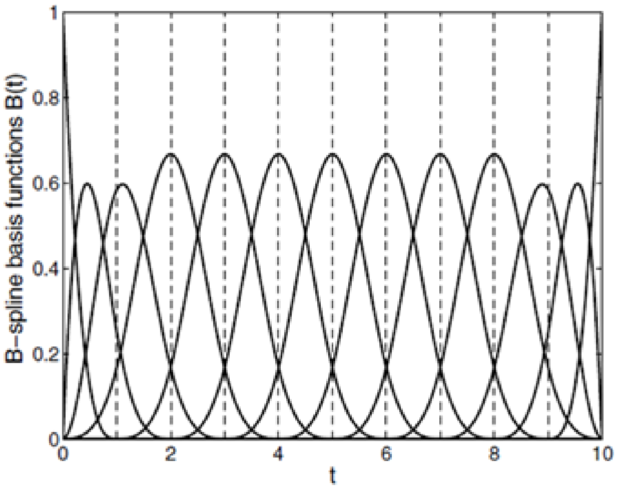
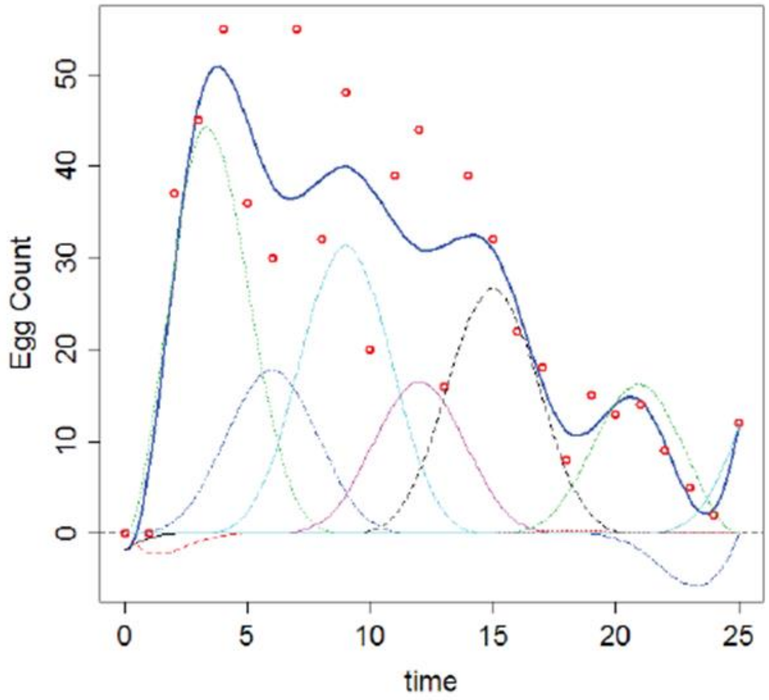


Figure 10
The thirteen basis functions defining an order four spline with nine interior knots, shown as vertical dashed lines.

Representing Functional Data

- An illustration of basis expansions for B-splines



Representing Functional Data

- **Properties of B-splines**

  - Number of basis functions

    order + number interior knots e.g. 4 + 9 = 13

  - Order m splines: derivatives up to (m − 2) are continuous.
  - Support on m adjacent intervals => highly sparse design matrix.

  - Advice
    - Flexibility comes from knots; derivatives from order.
    - Frequently, fewer knots will do just as well (approximation properties can be formalized).

Representing Functional Data

Local polynomial smoothing

- Basic idea:

$$SMSSE_t(y|c)=\sum_{j=1}^{n}w_j(t)\left[y_j-\sum_{k=1}^{K}c_k\phi_k(t)\right]^2 \qquad \text{or}$$

$$SMSSE_t(y|c)=(y-\Phi c)'W(t)(y-\Phi c)$$

where the weight functions $w_j(t)$ are constructed from the kernel function.

- Use a low order polynomial basis.

$$SMSSE_t(y|c)=\sum_{j=1}^{n}Kern_h(t_j,t)\left[y_j-\sum_{\ell=0}^{L}c_\ell(t-t_j)^\ell\right]^2$$

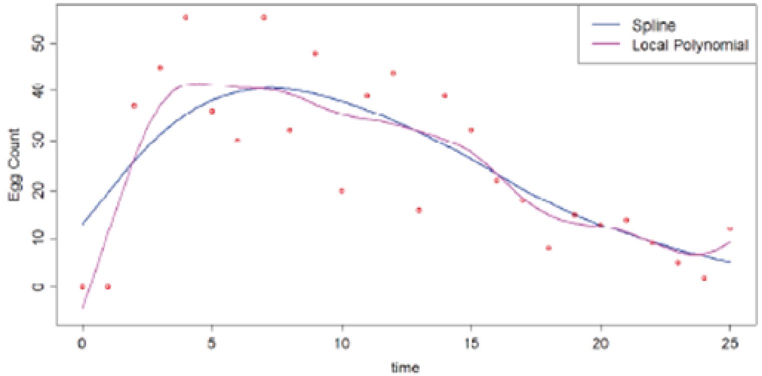
smoothing parameter: bandwidth h

- WSLE:

$$\hat{c}(t)=\left(\Phi'W(t)\Phi\right)^{-1}\Phi'W(t)y$$

Representing Functional Data

- **Example**
  - Local linear regression



$$(\hat{\beta}_0(t), \hat{\beta}_1(t)) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1(t - t_i))^2 K\left(\frac{t - t_i}{\lambda}\right)$$

Estimate $\hat{x}(t) = \hat{\beta}_0(t)$, $\widehat{Dx}(t) = \hat{\beta}_1(t)$.

EDA

- **Functional means and variances**
  - Mean function

$$\mu(t) = E(X(t))$$

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$$ (The average of the functions point-wise across replications.)

  - Variance function

$$\sigma^2(t) = Var(X(t))$$

$$\mathrm{var}_X(t) = \frac{1}{N-1}\sum_{i=1}^{N}\left[x_i(t) - \bar{x}(t)\right]^2$$

EDA

- Example - Pinch force data

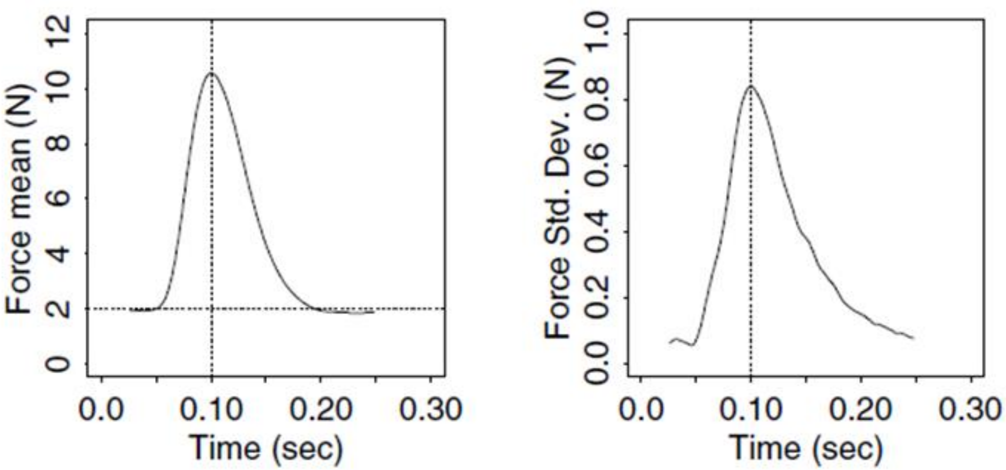


Figure 11 The mean and standard deviation functions for the 20 pinch force observations in Figure 8 after they were aligned or registered.

EDA

- Covariance and correlation functions
  - Covariance function

$$Cov(s,t)=Cov(X(s),X(t))$$

$$\operatorname{cov}_X(t_1,t_2)=\frac{1}{N-1}\sum_{i=1}^{N}\{x_i(t_1)-\bar{x}(t_1)\}\{x_i(t_2)-\bar{x}(t_2)\}$$

  - Correlation function

$$Corr(s,t)=Corr(X(s),X(t))$$

$$corr_X(t_1,t_2)=\frac{\operatorname{cov}_X(t_1,t_2)}{\sqrt{\operatorname{var}_X(t_1)\operatorname{var}_X(t_2)}}$$

EDA

- Example – Pinch force data

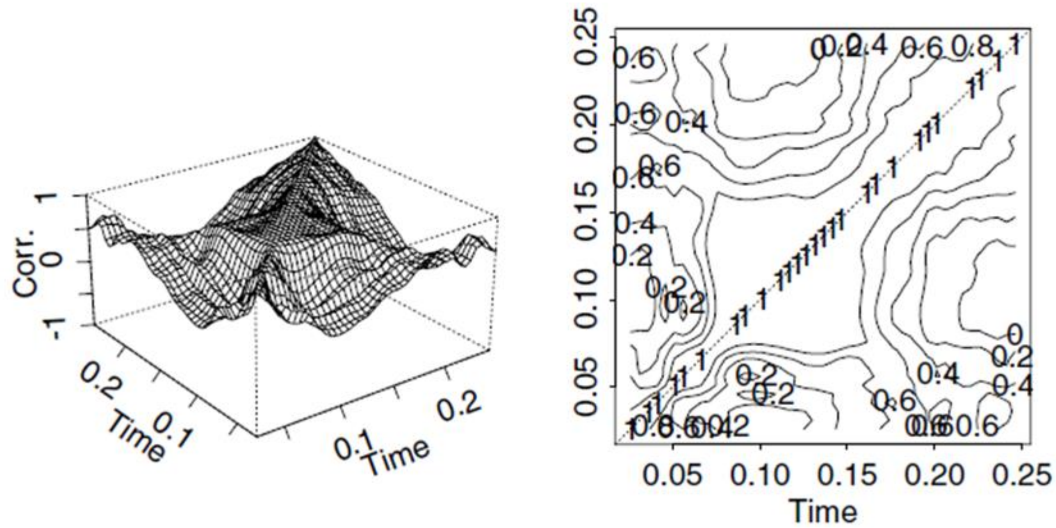


Figure 12 The left panel is a perspective plot of the bivariate correlation function values $r(t_1, t_2)$ for the pinch force data. The right panel shows the same surface by contour plotting. Time is measured in seconds.

EDA

- Cross-covariance and cross-correlation functions
  - Cross-covariance function

$$Cov\big(X(s),Y(t)\big)=E[(X(s)-\mu_X(s))(Y(t)-\mu_Y(t))]$$

$$\mathrm{cov}_{X,Y}(t_1,t_2)=\frac{1}{N-1}\sum_{i=1}^{N}\{x_i(t_1)-\bar{x}(t_1)\}\{y_i(t_2)-\bar{y}(t_2)\}$$

  - Cross-correlation function

$$Corr\big(X(s),Y(t)\big)=\frac{Cov(X(s),Y(t))}{\sqrt{Var\big(X(s)\big)Var(Y(t))}}$$

$$corr_{X,Y}(t_1,t_2)=\frac{\mathrm{cov}_{X,Y}(t_1,t_2)}{\sqrt{\mathrm{var}_X(t_1)\,\mathrm{var}_Y(t_2)}}$$

EDA

- Example - Canadian weather data (refer to Fig.3)

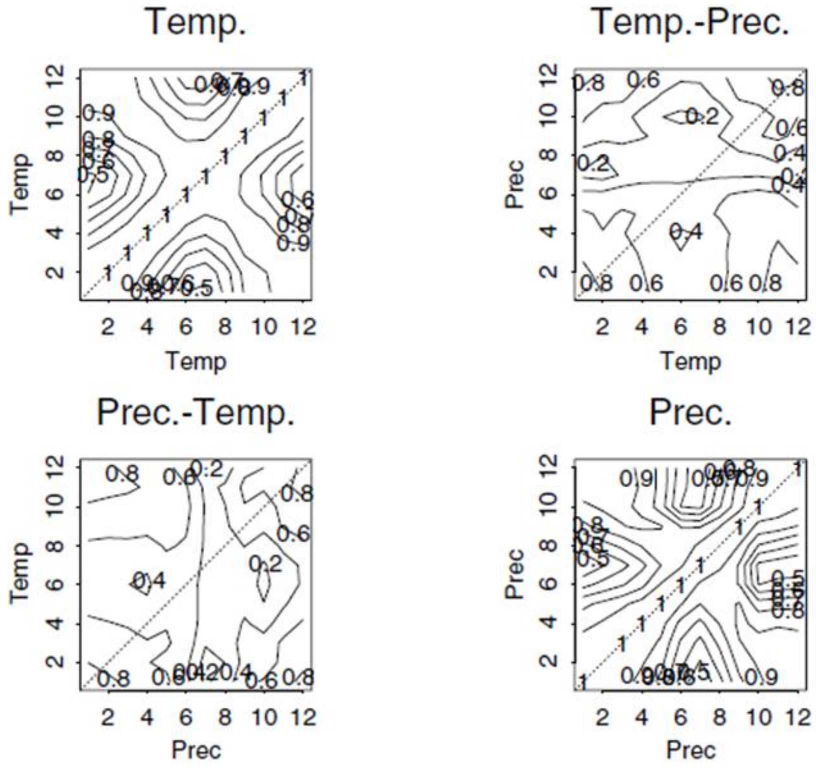


Figure 13 Contour plots of the correlation and cross-correlation functions for 35 Canadian weather stations for temperature and log precipitation. The cross-correlation functions are those in the upper right and lower left panels.

EDA

Functional PCA

- Principal components analysis (PCA) of functional data is a key technique to explore the features characterizing typical functions.
- A low-dimensional summary/interpretation.
- Multivariate PCA uses eigen-decomposition of the covariance matrix Σ :

$$\Sigma = U'\Lambda U = \sum_{j=1}^{p} \lambda_j \vec{u}_j \vec{u}_j',$$

and $\vec{u}_i' \vec{u}_j = 1\ (i = j)$.

- For functional PCA, use the covariance function Γ :

$$\Gamma(s,t) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(s) \varphi_j(t),$$

where $\int \varphi_i(t)\varphi_j(t)\, dt = 1(i = j)$.

EDA

- Karhunen-Loève decomposition / FPCA model
  - FPC scores:

$$\xi_{ij} = \int \big(x_i(t) - \bar{x}(t)\big)\varphi_j(t)dt\,, j = 1,2, \dots .$$

  - Reconstruction of $x_i(t)$:

$$x_i(t) = \bar{x}(t) + \sum_{j=1}^{\infty} \xi_{ij}\varphi_j(t)$$

EDA

- Example – Canadian Weather Data

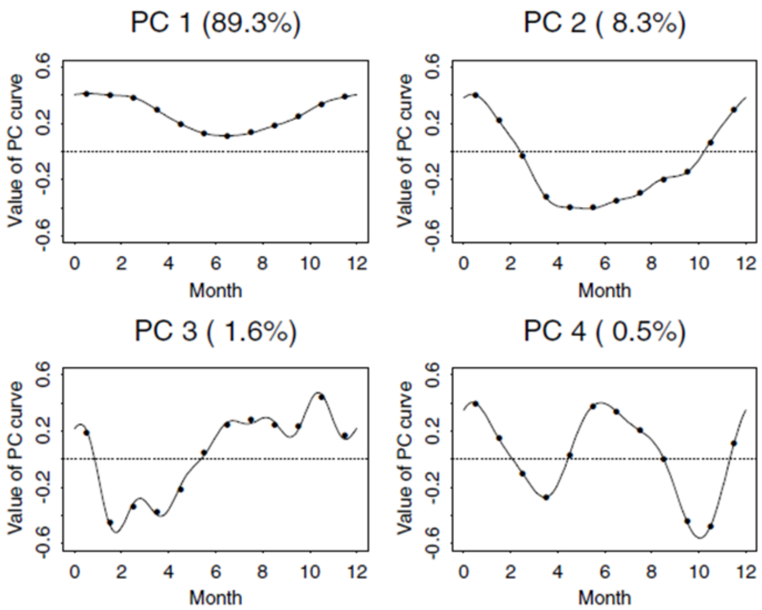


Figure 12 The first four principal component curves of the Canadian temperature data (refer to Fig.4) estimated by two techniques. The points are the estimates from the discretization approach, and the curves are the estimates from the expansion of the data in terms of a 12-term Fourier series. The percentages indicate the amount of total variation accounted for by each principal component.

EDA

- Example – Canadian Weather Data

  (1) Plotting components as perturbations of the mean:

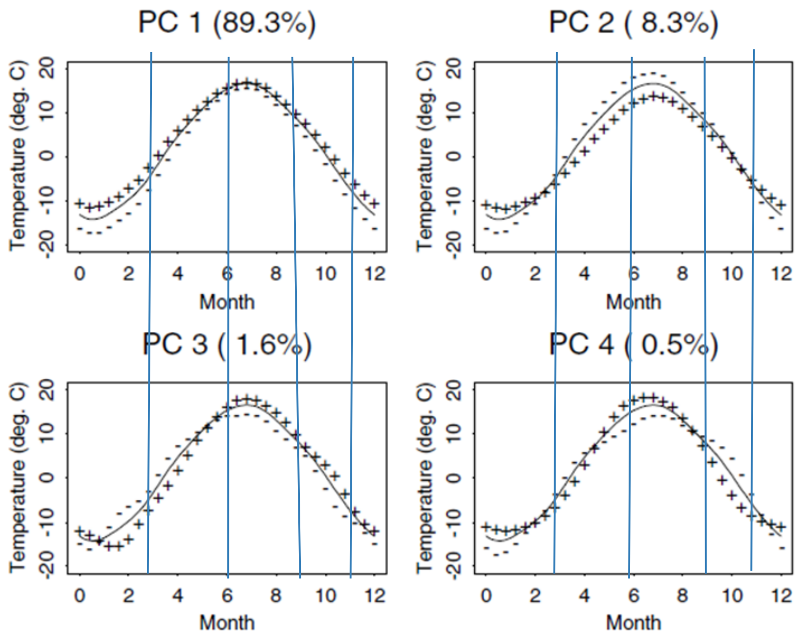
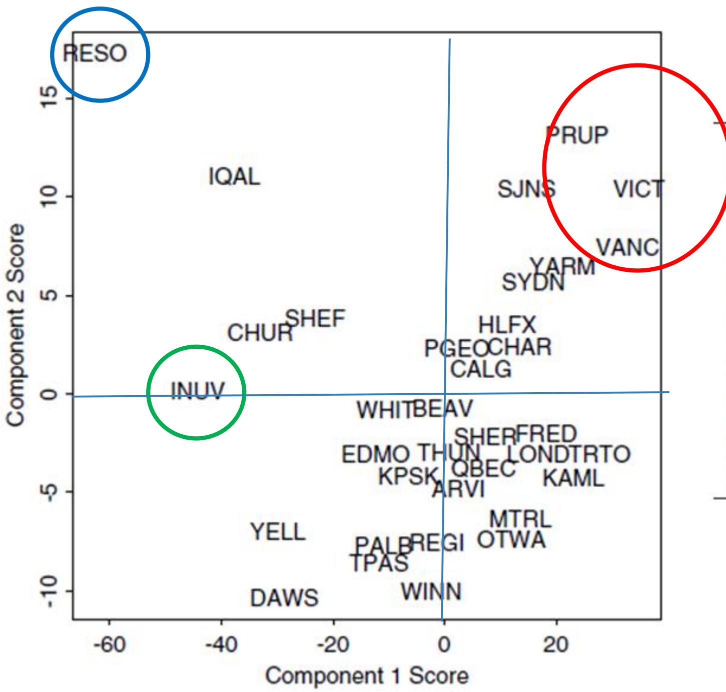


Figure 13 The mean temperature curves and the effects of adding (+) and subtracting (−) a suitable multiple of each PC curve.

EDA

- Example － Canadian Weather Data

  (2) Plotting principal component scores:



Arvida, Que.	Kapuskasing, Ont.	St. John's, Nfld
Beaverlodge, B.C.	London, Ont.	Sydney, N.S.
Calgary, Alta.	Montreal, Que.	The Pas, Man.
Charlottetown, P.E.I.	Ottawa, Ont.	Thunder Bay, Ont.
Churchill, Man.	Prince Albert, Sask.	Toronto, Ont.
Dawson, Yukon	Prince George, B.C.	Vancouver, B.C.
Edmonton, Alta.	Prince Rupert, B.C.	Victoria, B.C.
Fredericton, N.B.	Quebec City, Que.	Whitehorse, Yukon
Halifax, N.S.	Regina, Sask.	Winnipeg, Man.
Inuvik, N.W.T.	Resolute, N.W.T.	Yarmouth, N.S.
Iqualuit, N.W.T.	Schefferville, Que.	Yellowknife, N.W.T.
Kamloops, B.C.	Sherbrooke, Que.	

Figure 14 The scores of the weather stations on the first two principal components of temperature variation. The location of each weather station is shown by the four-letter abbreviation of its name.

The fda Package

- **Fda package**
  - J. O. Ramsay, Hadley Wickham, Spencer Graves, Giles Hooker.
  - https://cran.r-project.org/web/packages/fda/index.html

- **Fda Objects**
  - basis objects define basis systems that can be used.
  - fd objects store functional objects.
  - bifd objects store functions of two-dimensions.
  - Lfd objects define smoothing penalties.
  - fdPar objects collect fdobj, Lfodj, and a smoothing parameter.

The fda Package

- **Basis Objects**
  - Define basis systems of various types.
  - ✓ rangval Range of values for which basis is defined.
  - ✓ nbasis Number of basis functions.
  - Ex. Create a fourier basis on [0 365] with 21 basis functions.

    `fbasis = create.fourier.basis(c(0,365),21)`

- **Bspline Basis Objects**
  - Requre
  - ✓ norder Order of the splines
  - ✓ breaks Knots for the splines.
  - `nbasis = length(knows) + norder - 2`

The fda Package

- Ex. Create a B-spline basis of order 6 on the year [0 365] with knots at the months.
- ✓ nbasis = 13 + 6 - 2 = 17
- ✓ noder = 6
- ✓ months = cumsum(c(0,31,28,31,30,31,30,31,31,30,31,30,31))
- ✓ bbasis = create.bspline.basis(c(0,365), nbasis, norder, months)

➢ **Manipulation of basis objects**

- Plots bbasis. plot(bbasis)
- Evaluate fbasis at time 0:365. eval.basis(0:365, fbasis)
- Produces the inner product matrix $J_{ij} = \int \phi_i(t)\psi_j(t)dt$.

  inprod(bbasis, fbasis)

The fda Package

- fd Objects
  - Creates a functional data object:
    - coef array of coefficients
    - basis basis object
    - fdnames defines dimension names
  - Ex. Creates a functional data object with coeffcients coefs and basis bbasis.

    fdobj = fd(coefs, bbsis)
  - The basis bbasis coefs has three dimensions corresponding to:
    - index of the basis function
    - replicate
    - dimension

The fda Package

- **Manipulation of fd objects**
  - Pointwise calculation of fd objects:
    - ✓ fdobj1 + fdoj2
    - ✓ fdobj1^k
    - ✓ fdobj1*fdobj2
  - Subset of fd objects: fdobj[3,2]
  - Return an array of values of fdobj on 0:365:
    eval.fd(0:365,fdobj)
  - Give the nderiv-th derivative of fdobj:
    deriv.fd(fdobj,nderiv)
  - Plot fdobj:
    plot(fdobj)

The fda Package

Lfd Objects

- Define smoothing penalties:

$$Lx = D^m x - \sum_{j=0}^{m-1} \alpha_j(t) D^j x$$

and require that α_j to be given as a list of fd objects.

- Two common ways:
  - ✓ int2Lfd(k) creates an Lfd object $Lx = D^m x$
  - ✓ vec2Lfd(a) for vector *a* of length m creates an Lft object

$$Lx = D^m x - \sum_{j=0}^{m-1} a_j D^j x$$

- Ex. Creates a Harmonic acceleration penalty $Lx = D^3 x - \frac{2\pi}{365} D x$

  vec2Lfd(c(0,-2*pi/365,0))

The fda Package

- **fdPar Objects**
  - For imposing smoothing. It collects
    - ✓ fdobj
    - ✓ Lfdobj
    - ✓ lambda a smoothing parameter

- **bifd Objects**
  - Represent a function of two dimension s and t:

$$x(s,t) = \sum_{i=0}^{K_1} \sum_{j=1}^{K_2} \phi_i(s)\,\psi_j(t) c_{ij}$$

and require that α_j to be given as a list of fd objects.

The fda Package

- bifd Objects
  - Requires
    - ✓ coefs for the matrix of c_{jj}
    - ✓ sbasis basis object for defining $\phi_i(s)$
    - ✓ tbasis basis object for defining $\psi_j(t)$
  - Can be evaluated but not plotted.
  - bifdPar store bifd and Lfd objects and λ for each of s and t.

The fda Package

- **Smoothing Functions**
  - Main function: smooth.basis
  - Ex. Smooths the Canadian temperature data with a second derivative penalty, $\lambda = 0.01$.

```
data(daily)
argval = (1:365) – 0.5
fdParobj = fdPaf(fbasis, int2Lfd(2), 1e-2)
tempSmooth = smooth.basis(argvals, daily$tempav, fdParobj)
```

  - fd object returns:
    - ✓ df equivalent degrees of freedom
    - ✓ SSE total sum of squared errors
    - ✓ gcv vector giving GCV for each smooth
  - Typically, λ is chosen to minimize average gcv.

The fda Package

- **Functional Statistics**
  - Basic Statistics:
    - ✓ mean.df mean fd object
    - ✓ var.df Variance or covariance (bifd object)
    - ✓ sd.df Standard deviation (root diagonal of var.fd)
  - FPCA: (Smoothing not strictly necessary)
    - ✓ temppca = pca.fd(tempfd$fd, nharm =4, fdParobj)
  - Outputs of pca.fd:
    - ✓ harmonics fd objects giving eigen-fucntions
    - ✓ values eigen values
    - ✓ scores PCA scores
    - ✓ varprop Proportion of variance explained
  - Diagnostics plots:
    - ✓ plot(temppca)

Examples for R

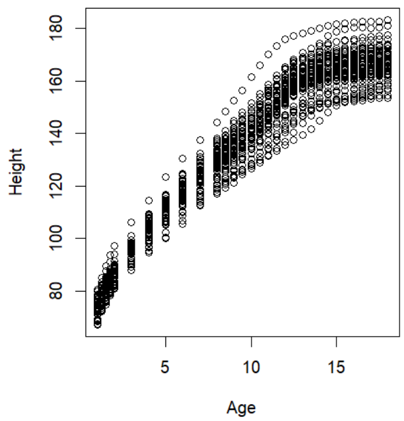
Please refer to the file 「R Examples for FDA_SS2025.r 」.

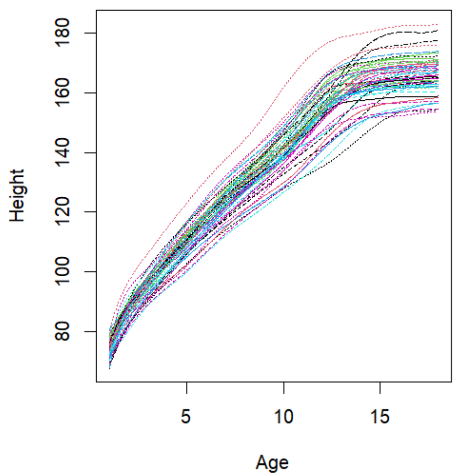
Exercise

- ➢ Berkeley Growth Study Data
- ◘ The growth data set in the fda package for R contains the heights of 54 girls measured at a set of 31 ages in the Berkeley Growth Study.

a. Fit these data by using a cubic B-spline basis (with norder = 4) with 12 basis functions. Plot the 54 fitted growth curves.

Hint: Use smooth.basis().

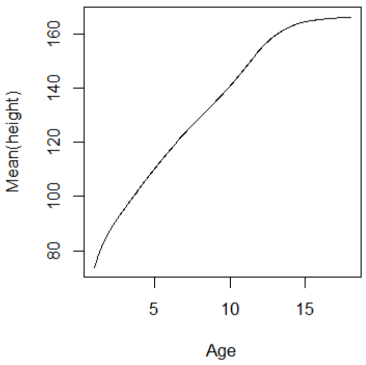


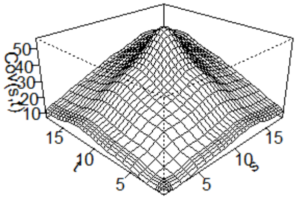


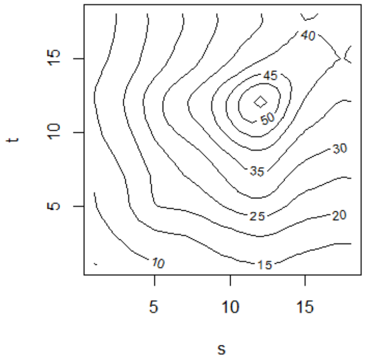
Exercise

b. Obtain the mean function and covariance functions.

Hint: Use mean.fd() and var.fd().





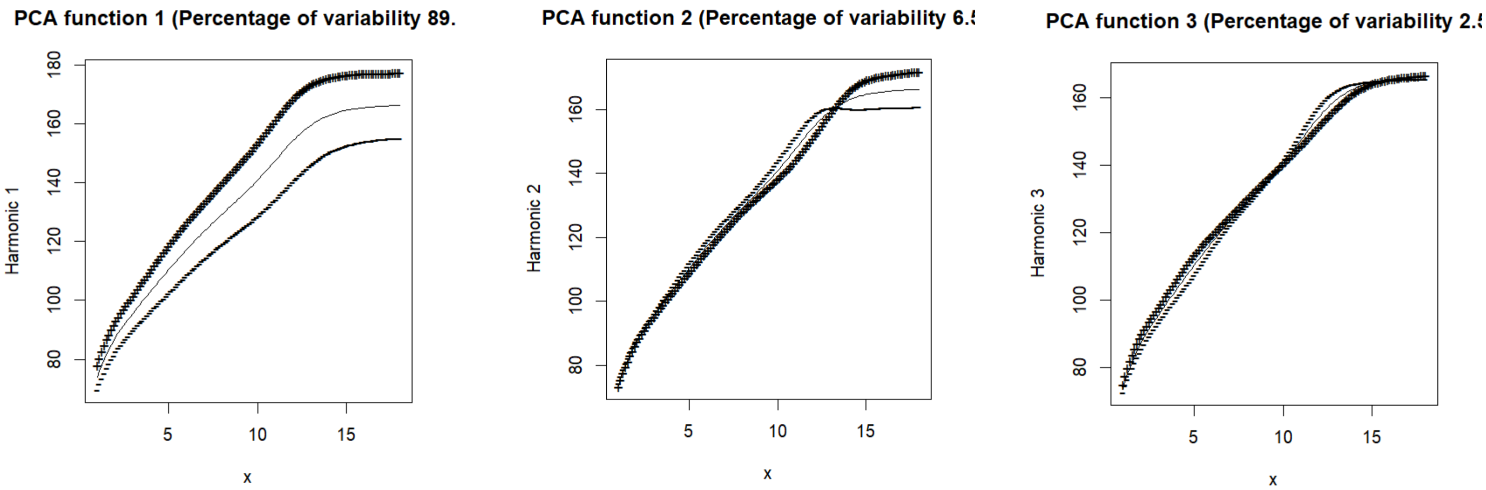


Exercise

c. Conduct a functional principal components analysis with 3 components using these smooths. Plot the first three principal component functions (or eigenfunctions) and provide the percentages of variability of each component.

Are the components interpretable? How many do you need to retain to recover 90% of the variation.

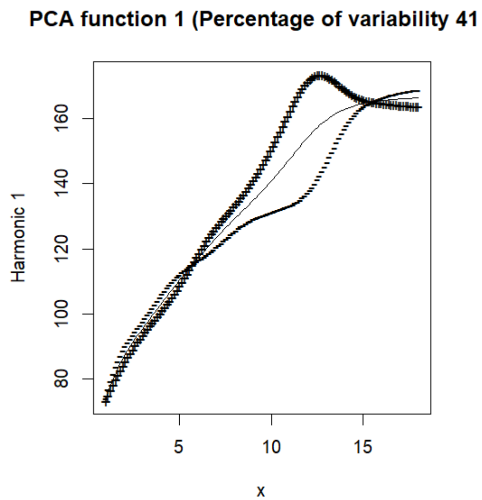
Hint: Use pca.fd().

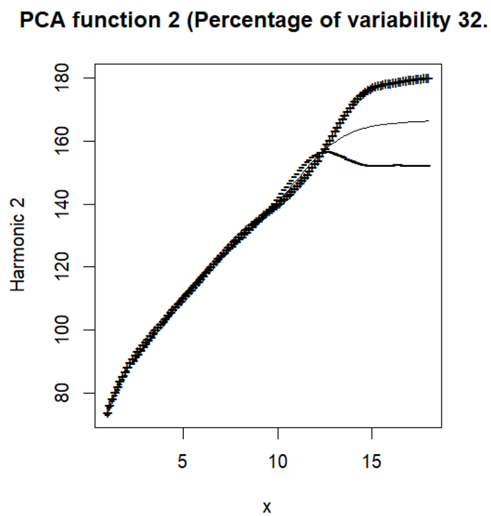


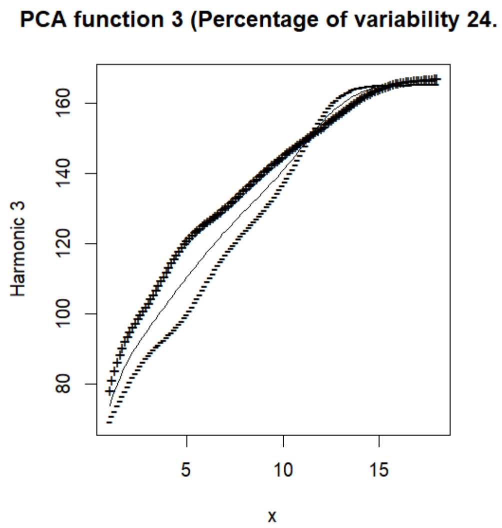
Exercise

d. Conduct a rotation of functional principal components by using the VARIMAX rotation algorithm. Plot the first three rotated principal component functions. Can the rotated components revel more meaningful components of variation? How?

Hint: Use varmx.pca.fd().

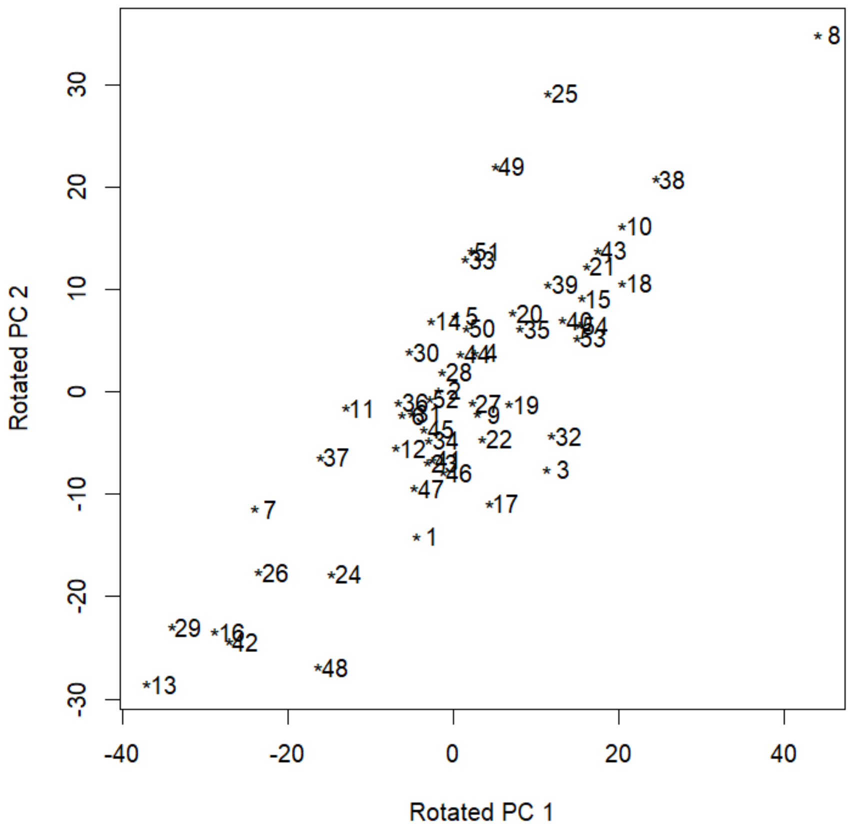






Exercise

e. Plot the first two FPC scores by a scatter plot. Explain the result.



Functional Linear Models

➢ Functional Linear Regression

- We wish to examine predictive relationships -> generalization of linear models.

$$y_i = \alpha + x_i \beta + \varepsilon_i$$

- Three different scenarios for x_i, y_i:

✓ Functional covariate, scalar response

✓ Scalar covariate, functional response

✓ Functional covariate, functional response

Functional Linear Models

- **Scalar Response Models**
  - We observe $x_i(t), y_i$.
  - Model: $y_i = \alpha + \int \beta(t) x_i(t) dt + \varepsilon_i$
  - Estimate β by minimizing squared error:

$$\beta(t) = \operatorname{argmin} \sum_i \left(y_i - \alpha - \int \beta(t) x_i(t) dt \right)^2$$

  - Smoothing:

$$PENSSE_\lambda(\beta) = \sum_{i=1}^{n} \left(y_i - \alpha - \int \beta(t) x_i(t) dt \right)^2 + \lambda \int [L\beta(t)]^2 dt$$

$$\beta(t) = \sum c_i \phi_i(t)$$

  - Extension: $y_i = \alpha + \sum_{j=1}^{p} \int \beta_j(t) x_{ij}(t) dt + \varepsilon_i$

Functional Linear Models

- **Scalar Response Models**
  - Functional Principal Components Regression

FPCA: $x_i(t) = \bar{x}(t) + \sum_{j=1}^{\infty} \xi_{ij}\varphi_j(t)$

Let $\beta(t) = \sum_{j=1}^{\infty} \beta_j \varphi_j(t)$

$$y_i = \beta_0 + \int \beta(t) x_i(t) dt + \varepsilon_i$$

$$y_i = \beta_0 + \sum \int \beta_j \varphi_j(t) x_i(t) dt + \varepsilon_i$$

$$= \beta_0 + \sum \beta_j \int \varphi_j(t) x_i(t) dt + \varepsilon_i$$

Functional Linear Models

- **Functional Response Models**
  - Case 1: Scalar Covariates $(x_i, y_i(t))$
  - Model: $y_i(t) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i(t)$
  - Conduct a linear regression at each time t (also works for ANOVA effects).
  - Smoothing:

$$PENSISE = \sum_{i=1}^{n} \int \left(y_i(t) - \hat{y}_i(t)\right)^2 dt + \sum_{j=0}^{p} \lambda_j \int \left[L_j \beta_j(t)\right]^2 dt$$

Usually keep λ_j, L_j all the same.

Functional Linear Models

Functional Response Models

- Case 2: Concurrent Linear Model for $(x_i(t), y_i(t))$

- Model: $y_i(t) = \beta_0(t) + \beta_1(t)x_i(t) + \varepsilon_i(t)$

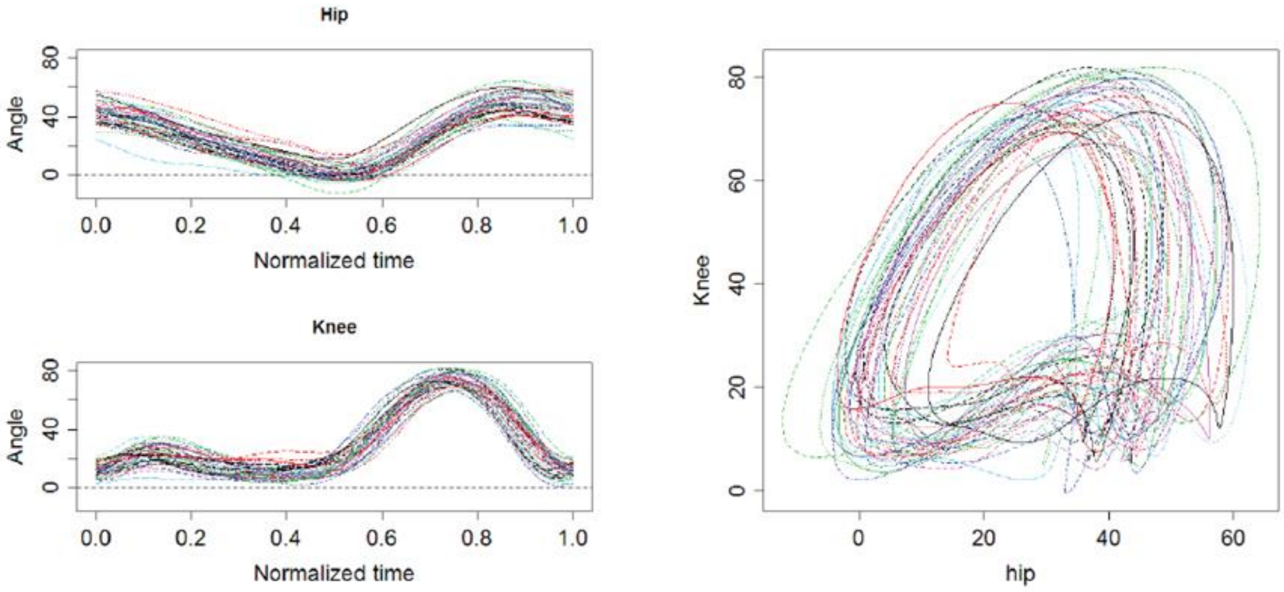
✓ $(x_i(t), y_i(t))$ must be measured in the same time domain.

✓ Must be appropriate to compare observations time-point by time-point.

✓ Especially useful if $y_i(t)$ is a derivative of $x_i(t)$.

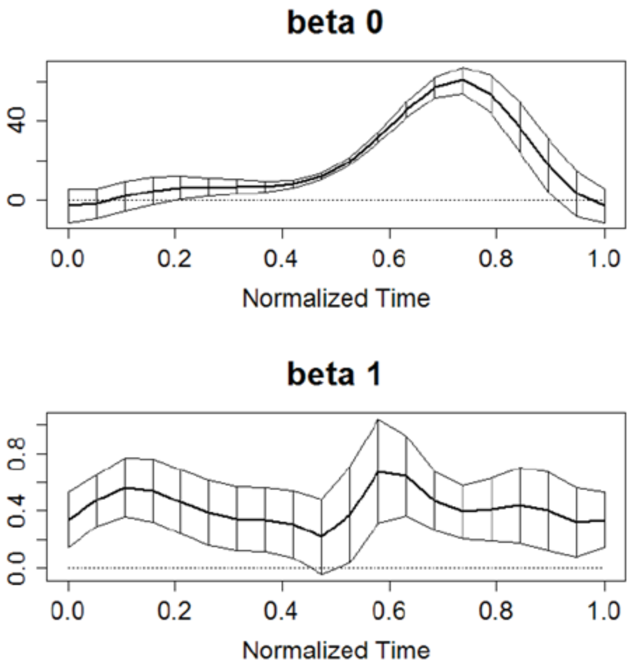
Functional Linear Models

- **Example: Gait Data**
  - Records of the angle of hip and knee of 39 subjects taking a step.
  - Interest in kinetics of walking.



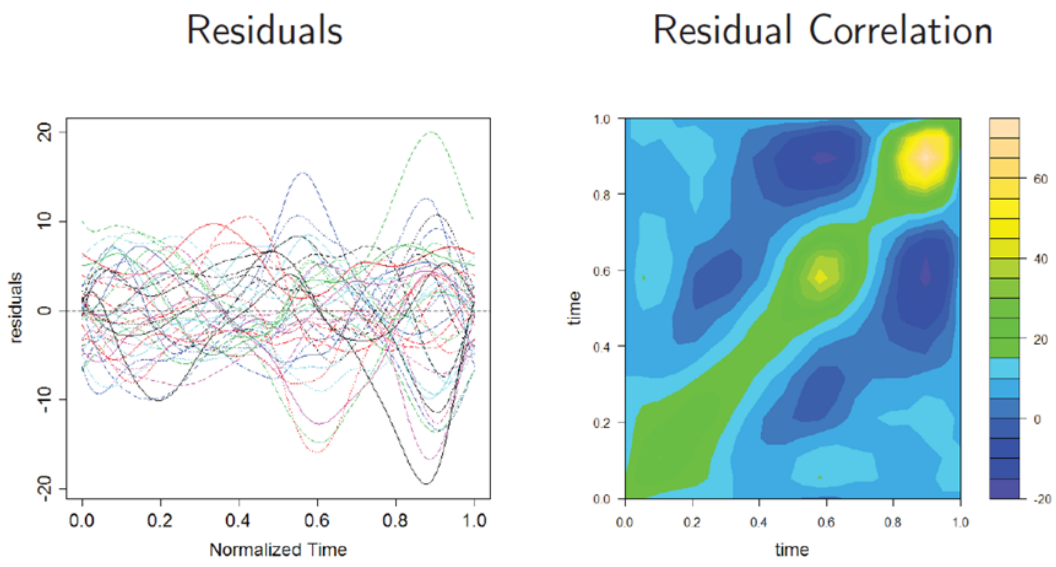
Functional Linear Models

- Gait Model
  - $knee(t) = \beta_0(t) + \beta_1(t)hip(t) + \varepsilon(t)$.
  - $\beta_0(t)$ indicates a well-defined autonomous knee cycle.
  - $\beta_1(t)$ modulation of cycle with respect to hip
  - More hip bend also indicates more knee bend; by a fairly constant amount throughout cycle.



Functional Linear Models

- **Gait Residuals: Covariance and Diagnostics**
  - Examine residual functions for outliers, skewness etc. (can be challenging).
  - Residual correlation may be of independent interest.



Functional Linear Models

- **Functional Response Models**
  - Case 3: Functional Response, Functional Covariate $(x_i(s), y_i(t))$
  - Model: $y_i(t) = \beta_0(t) + \int \beta_1(s,t) x_i(s) ds + \varepsilon_i(t)$
    - ✓ Same identification issues as scalar response models.
    - ✓ Usually penalize $\beta_1(s)$ in each direction separately.

$$\lambda_s \int [L_s \beta_1(s,t)]^2 dsdt + \lambda_t \int [L_t \beta_1(s,t)]^2 dsdt$$

Functional Linear Models

- **Summary**
  - Scalar Response Model
    - ✓ Functional covariate implies a functional parameter.
    - ✓ Use the smoothness of $\beta_1(t)$ to obtain identifiability.
    - ✓ Variance estimates come from sandwich estimators
  - Concurrent Linear Model
    - ✓ $y_i(t)$ only depends on $x_i(t)$ at the current time.
    - ✓ Scalar covariates = constant functions.
    - ✓ Will be used in dynamics.
  - Functional Covariate/Functional Response
    - ✓ Most general functional linear model.

R for Functional Linear Models

- **fRegress**
  - Main function for functional linear models.
  - Requires:
    - ✓ y response, either vector or fd object.
    - ✓ xlist list containing covariates; vectors or fd objects.
    - ✓ betalist list of fdPar objects to define bases and smoothing penalties for each coefficient.
  - Returns depend on y:
    - ✓ betaestlist list of fdPar objects with estimated β coefficients.
    - ✓ yhatfdobj predicted values, either numeric or fd.

R for Functional Linear Models

- **fRegress.stderr**
  - Produce pointwise standard errors for the $\hat{\beta}_j$.
  - ✓ model output of fRegress.
  - ✓ y2cmap smoothing matrix for the responses (obtained from smooth.basis)
  - ✓ SigmaE error covariance for the response.
  - Output
  - ✓ betastderrlist contains fd objects giving the pointwise standard errors.

R for Functional Linear Models

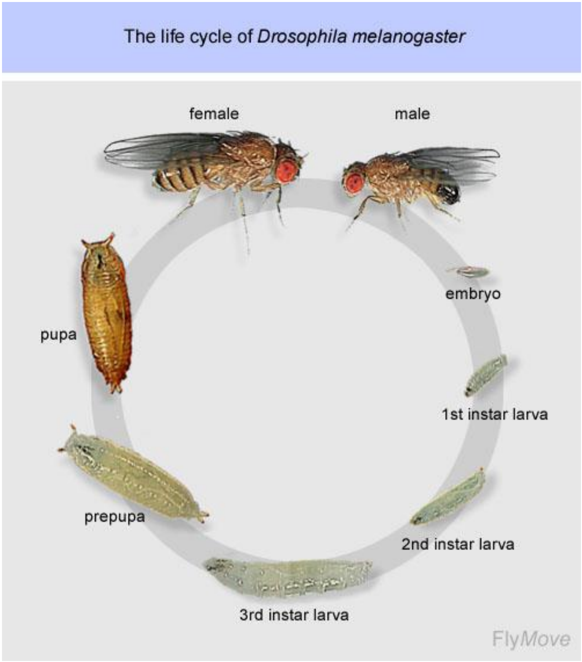
- fRgress.CV
  - Provides leave-one-out cross valdation.
  - Same arguements for fRegress, allows use of specific observations.
  - For concurrent linear models,

$$CV(\lambda) = \sum_{i=1}^{n} \int (y_i(t) - \hat{y}_{\lambda}^{-i}(t))^2$$

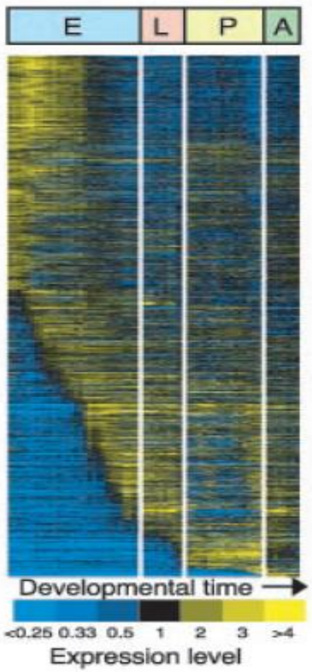
- ✓ $\hat{y}_{\lambda}^{-i}(t)$ is the prediction with smoothing parameter λ and without ith observation.
- ✓ plotbeta(betaestlist,betastderrliest) produces graphs with confidence regions.

More examples

- **Clustering** Gene Expression Data



Source :www.anatomy.unimelb.edu.au/researchlabs/whitington



Source : Arbeitman et al. (2002)

More examples

➢ **Clustering** Gene Expression Data

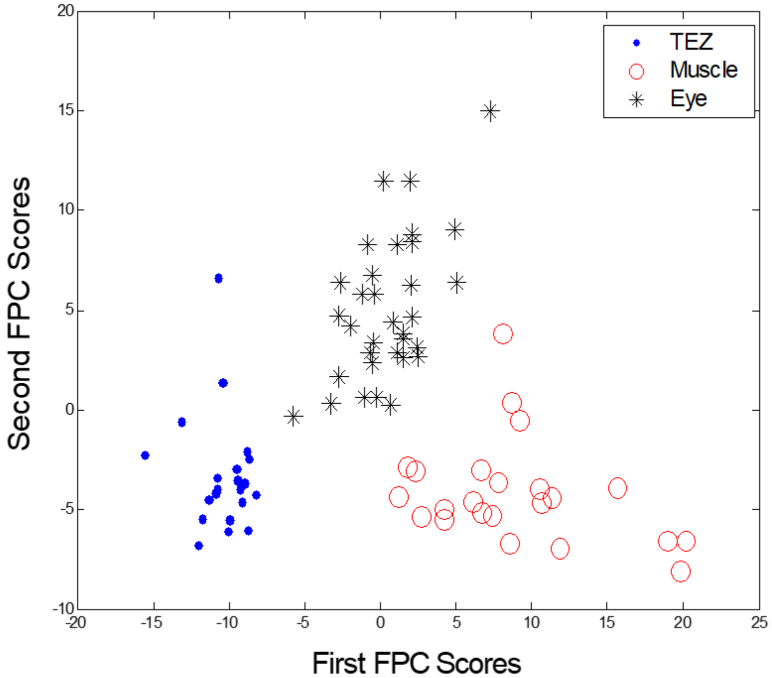


Figure 15 Scatter plot of the first two FPC scores obtained from the *k* –centers FC procedure for the gene expression profile data.

More examples

➢ **Clustering** Gene Expression Data

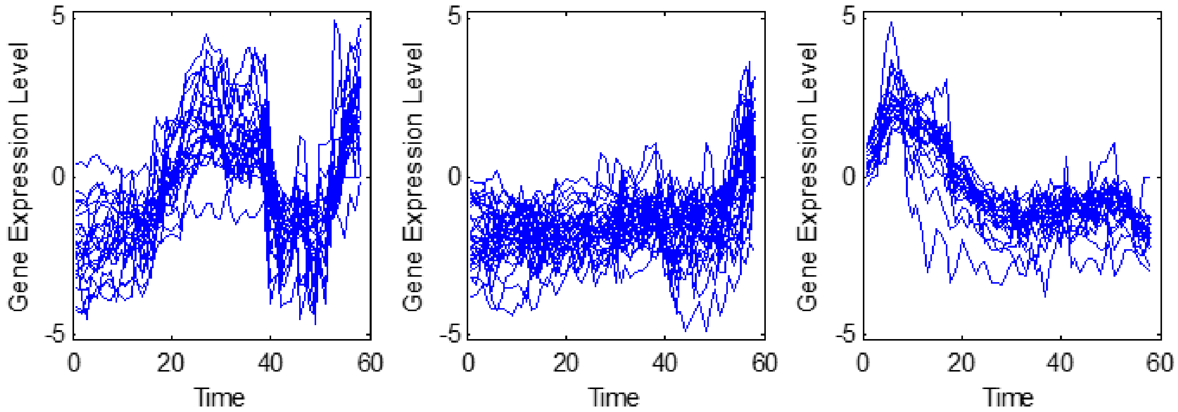


Figure 16 Gene expression profiles clustered using the k-centers FC procedure, corresponding the eye- and muscle-specific, and TEZ genes, respectively (from left to right for clusters 1, 2, and 3).

Thanks!

Contact me at:
plli@gms.tku.edu.tw