Functional Principal Component Analysis

Jiguo Cao, PhD

Canada Research Chair in Data Science

Associate Professor, Department of Statistics and Actuarial Science

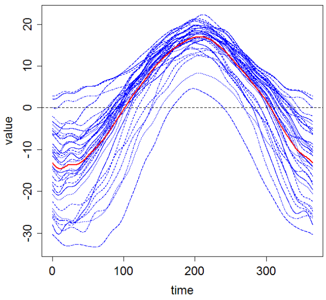
Associate Faculty Member, School of Computing Science

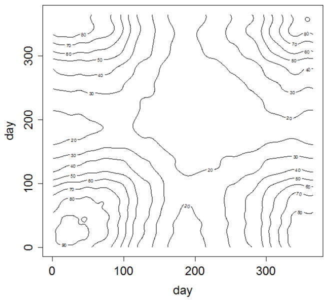
Simon Fraser University, Vancouver, Canada

Understanding the Distribution of Collections of Functions

Summary statistics:

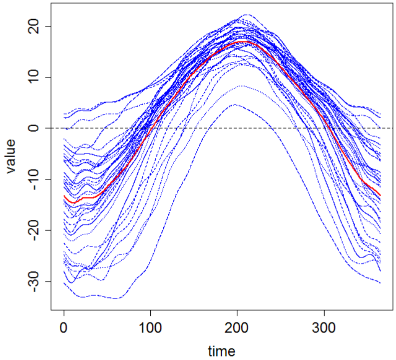
- mean $\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$
- covariance $\sigma(s,t) = \frac{1}{n}\sum_{i=1}^{n}(x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t))$

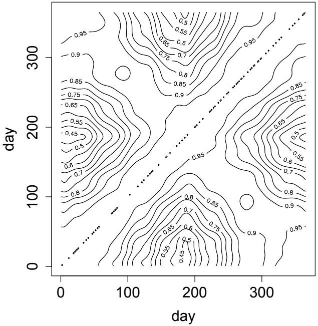




Correlation of Functional Data

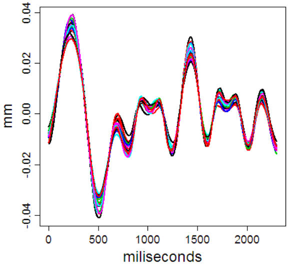
$$\rho(s,t) = \frac{\sum_{i=1}^{n}(x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t))}{\sqrt{\sum_{i=1}^{n}(x_i(s) - \bar{x}(s))^2}\sqrt{\sum_{i=1}^{n}(x_i(t) - \bar{x}(t))^2}}$$

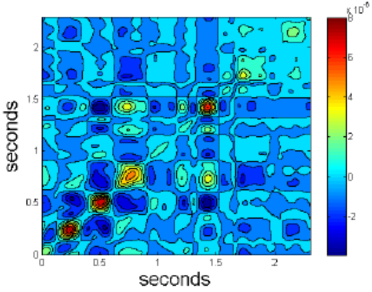




Exploring Functional Covariance

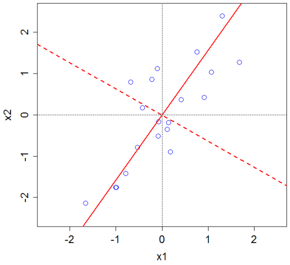
Covariance surfaces provide insight but do not describe the major directions of variation.



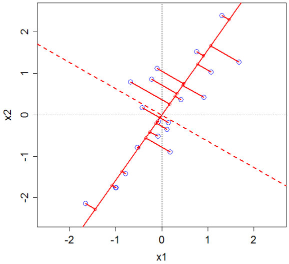


Multivariate Principal Components Analysis

- Directions of greatest variation



- Dimension reduction – subspace closest to the data



Frequently picks out interpretable contrasts.

A Little Analysis

- If $\mathbf{x}$ has covariance Σ, the variance of $u^T\mathbf{x}$ is $u^T\Sigma u$.
- Objective: Maximize $u^T\Sigma u$

A Little Analysis

- If $\mathbf{x}$ has covariance Σ, the variance of $u^T\mathbf{x}$ is $u^T\Sigma u$.
- Objective: Maximize $u^T\Sigma u$ with $u^T u = 1$

A Little Analysis

- If $\mathbf{x}$ has covariance Σ, the variance of $u^T\mathbf{x}$ is $u^T\Sigma u$.
- Objective: Maximize $u^T\Sigma u$ with $u^T u = 1$
- This is equivalent to solve the eigen-equation

$$\Sigma u = \lambda u$$

Mechanics of PCA

- Estimate covariance matrix: $\Sigma = \frac{1}{n} \sum \mathbf{x}_i \mathbf{x}_i^T$
- Take the eigen-decomposition of $\Sigma = U^T D U$
- Columns of U are orthogonal; represent a new basis
- D is diagonal; entries give variances of data along corresponding directions U.
- $d_k / \sum d_k =$ "proportion of variance explained".
- Order D, U in terms of decreasing d_i.
- u_k is the k-th column of U. It is the k-th principal component.
- From original data, $\mathbf{x}_i$, $\mathbf{x}_i^T u_k$ is the k-th principal component score; co-ordinate in new basis.

Functional Principal Component Analysis

In functional data analysis,

- Functional principal component analysis (FPCA) plays a crucial role;

Functional Principal Component Analysis

In functional data analysis,

- Functional principal component analysis (FPCA) plays a crucial role;
- Top K FPCs $w_1(t), \ldots, w_K(t)$
- Top functional principal components (FPC) summarize major sources of variation among multiple curves $X_i(t)$, $i = 1, \ldots, n$;
- $X_i(t)$ is projected to $s_{i1}, \ldots, s_{iK}$: $X_i(t) = \sum_{k=1}^{K} s_{ik} w_k(t)$

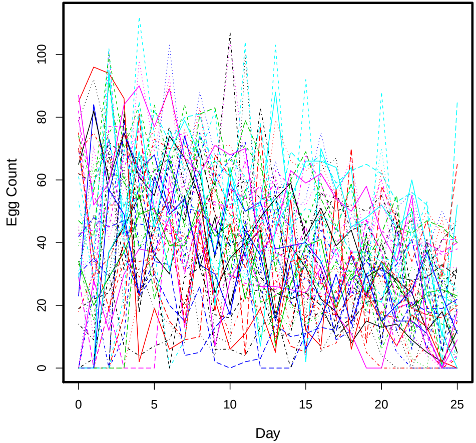
Functional Principal Component Analysis

In functional data analysis,

- Functional principal component analysis (FPCA) plays a crucial role;
- Top K FPCs $w_1(t), \ldots, w_K(t)$
- Top functional principal components (FPC) summarize major sources of variation among multiple curves $X_i(t)$, $i = 1, \ldots, n$;
- $X_i(t)$ is projected to $s_{i1}, \ldots, s_{iK}$: $X_i(t) = \sum_{k=1}^{K} s_{ik} w_k(t)$
- Top functional principal components (FPC) are represented by a set of flexible basis functions such as B-spline basis.
- Shapes of top FPCs are simple.

One example

- Number of eggs laid by 50 Mediterranean fruit flies over 25 days
- Objective: Exploring major modes of variability in 50 curves



Introduction of FPCA

- Data - n curves: $X_1^*(t), X_2^*(t), \ldots, X_n^*(t)$

Introduction of FPCA

- Data - n curves: $X_1^*(t), X_2^*(t), \ldots, X_n^*(t)$
- Pre-processing: $X_i(t) = X_i^*(t) - \mu(t)$
- Mean Curve: $\mu(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^*(t)$

Introduction of FPCA

- Data - n curves: $X_1^*(t), X_2^*(t), \ldots, X_n^*(t)$
- Pre-processing: $X_i(t) = X_i^*(t) - \mu(t)$
- Mean Curve: $\mu(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^*(t)$
- First FPC: $w_1(t)$;
- First FPC score: $s_{i1} = \int w_1(t)X_i(t)dt$

Introduction of FPCA

- Data - n curves: $X_1^*(t), X_2^*(t), \ldots, X_n^*(t)$
- Pre-processing: $X_i(t) = X_i^*(t) - \mu(t)$
- Mean Curve: $\mu(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^*(t)$
- First FPC: $w_1(t)$;
- First FPC score: $s_{i1} = \int w_1(t)X_i(t)dt$
- Maximize $\sum_{i=1}^{n} s_{i1}^2$ subject to $\int w_1^2(t)dt = 1$.

Introduction of FPCA

- Data - n curves: $X_1^*(t), X_2^*(t), \ldots, X_n^*(t)$
- Pre-processing: $X_i(t) = X_i^*(t) - \mu(t)$
- Mean Curve: $\mu(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^*(t)$
- First FPC: $w_1(t)$;
- First FPC score: $s_{i1} = \int w_1(t) X_i(t) dt$
- Maximize $\sum_{i=1}^{n} s_{i1}^2$ subject to $\int w_1^2(t) dt = 1$.
- $w_1(t)$ represents the strongest and most important mode of variation in n curves.

Introduction of FPCA

- Data - n curves: $X_1^*(t), X_2^*(t), \ldots, X_n^*(t)$
- Pre-processing: $X_i(t) = X_i^*(t) - \mu(t)$
- Mean Curve: $\mu(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^*(t)$
- Second FPC: $w_2(t)$;
- Second FPC score: $s_{i2} = \int w_2(t)X_i(t)dt$

Introduction of FPCA

- Data - n curves: $X_1^*(t), X_2^*(t), \ldots, X_n^*(t)$
- Pre-processing: $X_i(t) = X_i^*(t) - \mu(t)$
- Mean Curve: $\mu(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^*(t)$
- Second FPC: $w_2(t)$;
- Second FPC score: $s_{i2} = \int w_2(t)X_i(t)dt$
- Maximize $\sum_{i=1}^{n} s_{i2}^2$ subject to $\int w_2^2(t)dt = 1$ and $\int [w_1(t)w_2(t)]dt = 0$.

Introduction of FPCA

- Data - n curves: $X_1^*(t), X_2^*(t), \ldots, X_n^*(t)$
- Pre-processing: $X_i(t) = X_i^*(t) - \mu(t)$
- Mean Curve: $\mu(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^*(t)$
- Second FPC: $w_2(t)$;
- Second FPC score: $s_{i2} = \int w_2(t)X_i(t)dt$
- Maximize $\sum_{i=1}^{n} s_{i2}^2$ subject to $\int w_2^2(t)dt = 1$ and $\int [w_1(t)w_2(t)]dt = 0$.
- $w_2(t)$ represents the second strongest and most important mode of variation in n curves.

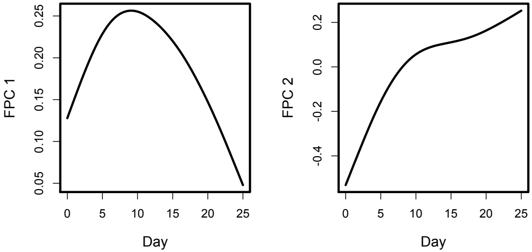
Introduction of FPCA

- Data - n curves: $X_1^*(t), X_2^*(t), \ldots, X_n^*(t)$
- Pre-processing: $X_i(t) = X_i^*(t) - \mu(t)$
- Mean Curve: $\mu(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^*(t)$
- Second FPC: $w_2(t)$;
- Second FPC score: $s_{i2} = \int w_2(t)X_i(t)dt$
- Maximize $\sum_{i=1}^{n} s_{i2}^2$ subject to $\int w_2^2(t)dt = 1$ and $\int [w_1(t)w_2(t)]dt = 0$.
- $w_2(t)$ represents the second strongest and most important mode of variation in n curves.
- Similarly, we obtain the rest top K FPCs $w_3(t), \ldots, w_K(t)$. FPCs are orthogonal to each other.

Introduction of FPCA

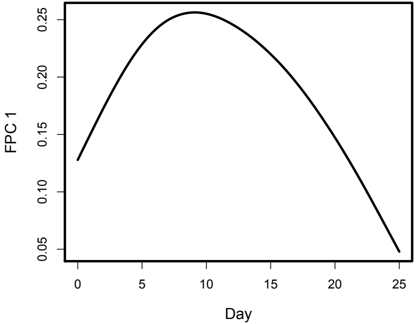
- Data - n curves: $X_1^*(t), X_2^*(t), \ldots, X_n^*(t)$
- Pre-processing: $X_i(t) = X_i^*(t) - \mu(t)$
- Mean Curve: $\mu(t) = \frac{1}{n}\sum_{i=1}^n X_i^*(t)$
- Second FPC: $w_2(t)$;
- Second FPC score: $s_{i2} = \int w_2(t)X_i(t)dt$
- Maximize $\sum_{i=1}^n s_{i2}^2$ subject to $\int w_2^2(t)dt = 1$ and $\int [w_1(t)w_2(t)]dt = 0$.
- $w_2(t)$ represents the second strongest and most important mode of variation in n curves.
- Similarly, we obtain the rest top K FPCs $w_3(t), \ldots, w_K(t)$. FPCs are orthogonal to each other.
- $X_i(t)$ is projected to $s_{i1}, \ldots, s_{iK}$: $X_i(t) = \sum_{k=1}^K s_{ik} w_k(t)$

Top Two FPCs of the medfly data



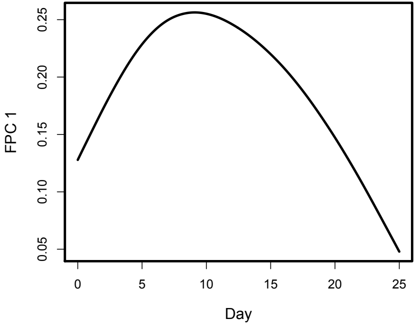
- Represented by B-spline basis functions;
- Explain 91.6% of total variations of data;
- Simple trends.

Interpretation of First FPC



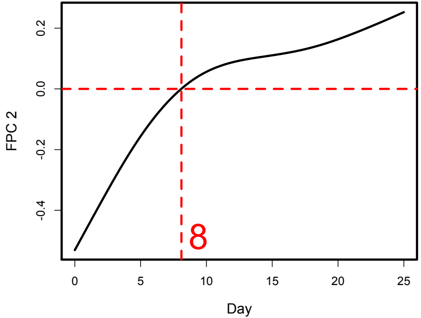
- First FPC score: $s_{i1} = \int w_1(t) X_i(t) dt$
- First FPC: Explains around 62.2% of total variability.
- First FPC: Positive over the whole time interval.

Interpretation of First FPC



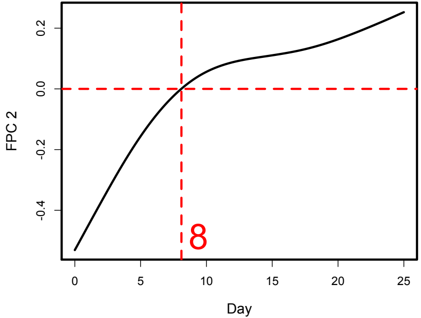
- First FPC score: $s_{i1} = \int w_1(t)X_i(t)dt$
- First FPC: Explains around 62.2% of total variability.
- First FPC: Positive over the whole time interval.
- Interpretation of First FPC score: Weighted average of the number of eggs laid by 50 Mediterranean fruit flies over 25 days.

Interpretation of Second FPC



- Second FPC score: $s_{i2} = \int w_2(t) X_i(t) dt$
- Second FPC: Explains around 29.4% of total variability.
- Second FPC: Negative before Day 8 and Positive after Day 8.

Interpretation of Second FPC



- Second FPC score: $s_{i2} = \int w_2(t)X_i(t)dt$
- Second FPC: Explains around 29.4% of total variability.
- Second FPC: Negative before Day 8 and Positive after Day 8.
- Interpretation of Second FPC score: Change of the number of eggs laid by 50 Mediterranean fruit flies after 8 days.

- PCA: Covariance matrix Σ

- PCA: Covariance matrix Σ
- Eigen-decomposition: $\Sigma = U^T D U = \sum d_i u_i u_i^T$

- PCA: Covariance matrix Σ
- Eigen-decomposition: $\Sigma = U^T D U = \sum d_i u_i u_i^T$
- FPCA: Functional covariance function $\sigma(s, t)$.

- PCA: Covariance matrix Σ
- Eigen-decomposition: $\Sigma = U^T D U = \sum d_i u_i u_i^T$
- FPCA: Functional covariance function $\sigma(s, t)$.
- Karhunen-Loève decomposition $\sigma(s, t) = \sum_{j=1}^{\infty} d_j w_j(s) w_j(t)$

- PCA: Covariance matrix Σ
- Eigen-decomposition: $\Sigma = U^T D U = \sum d_i u_i u_i^T$
- FPCA: Functional covariance function $\sigma(s, t)$.
- Karhunen-Loève decomposition $\sigma(s, t) = \sum_{j=1}^{\infty} d_j w_j(s) w_j(t)$
- FPC score: $s_{ij} = \int w_j(t) X_i(t) dt$
- $d_j = \mathrm{Var}(s_{ij})$ for given j

- PCA: Covariance matrix Σ
- Eigen-decomposition: $\Sigma = U^T D U = \sum d_i u_i u_i^T$
- FPCA: Functional covariance function $\sigma(s, t)$.
- Karhunen-Loève decomposition $\sigma(s, t) = \sum_{j=1}^{\infty} d_j w_j(s) w_j(t)$
- FPC score: $s_{ij} = \int w_j(t) X_i(t) dt$
- $d_j = \text{Var}(s_{ij})$ for given j
- d_j represents amount of variation in the direction $w_j(t)$.
- $\frac{d_j}{\sum_{j=1}^{\infty} d_j}$ is the proportion of variance explained.

Computing FPCA

Components solve the eigen-equation

$$\int \sigma(s,t)w_i(t)dt = \lambda w_i(t)$$

Option 1

1. take a fine grid $\mathbf{t} = [t_1, \ldots, t_K]$
2. find the eigen-decomposition of $\Sigma(\mathbf{t}, \mathbf{t})$
3. interpolate the eigenvectors

Option 2 (in `fda` library)

1. if the $x_i(t)$ have a common basis expansion, so must the eigen-functions
2. can re-express eigen-equation in terms of co-efficients

Backstage Linear Algebra

- $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T$

Backstage Linear Algebra

- $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T$
- $\phi(t) = (\phi_1(t), \ldots, \phi_J(t))^T$

Backstage Linear Algebra

- $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T$
- $\phi(t) = (\phi_1(t), \ldots, \phi_J(t))^T$
- Centered curves $\mathbf{x}(t) = \mathbf{C}\phi(t)$

Backstage Linear Algebra

- $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T$
- $\phi(t) = (\phi_1(t), \ldots, \phi_J(t))^T$
- Centered curves $\mathbf{x}(t) = \mathbf{C}\phi(t)$
- $\sigma(s,t) = n^{-1}\mathbf{x}(t)\mathbf{x}^T(t) = n^{-1}\phi^T(s)\mathbf{C}^T\mathbf{C}\phi(t)$

Backstage Linear Algebra

- $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T$
- $\phi(t) = (\phi_1(t), \ldots, \phi_J(t))^T$
- Centered curves $\mathbf{x}(t) = \mathbf{C}\phi(t)$
- $\sigma(s,t) = n^{-1}\mathbf{x}(t)\mathbf{x}^T(t) = n^{-1}\phi^T(s)\mathbf{C}^T\mathbf{C}\phi(t)$
- $w(t) = \phi^T(t)\mathbf{b}$

Backstage Linear Algebra

- $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T$
- $\phi(t) = (\phi_1(t), \ldots, \phi_J(t))^T$
- Centered curves $\mathbf{x}(t) = \mathbf{C}\phi(t)$
- $\sigma(s,t) = n^{-1}\mathbf{x}(t)\mathbf{x}^T(t) = n^{-1}\phi^T(s)\mathbf{C}^T\mathbf{C}\phi(t)$
- $w(t) = \phi^T(t)\mathbf{b}$
- Want to maximize

$$\int \sigma(s,t)w(t)dt = \rho w(s)$$

  subject to

$$\int [w(t)]^2 dt = \mathbf{b}^T \int \phi(t)^T \phi(t) dt \mathbf{b} = 1$$

- $\mathbf{W} = \int \phi(t)\phi^T(t)dt$

Backstage Linear Algebra

- Centered curves $\mathbf{x}(t) = \mathbf{C}\phi(t)$

Backstage Linear Algebra

- Centered curves $\mathbf{x}(t) = \mathbf{C}\phi(t)$
- $\sigma(s,t) = n^{-1}\mathbf{x}(t)\mathbf{x}^T(t) = n^{-1}\phi^T(s)\mathbf{C}^T\mathbf{C}\phi(t)$
- $w(t) = \phi^T(t)\mathbf{b}$
- $\mathbf{W} = \int \phi(t)\phi^T(t)dt$
- Want to maximize

$$\int \sigma(s,t)w(t)dt = \rho w(s)$$

$$\int n^{-1}\phi^T(s)\mathbf{C}^T\mathbf{C}\phi(t)\phi^T(t)\mathbf{b}dt = \rho\phi^T(s)\mathbf{b}$$

$$n^{-1}\phi^T(s)\mathbf{C}^T\mathbf{C}\int \phi(t)\phi^T(t)dt\mathbf{b} = \rho\phi^T(s)\mathbf{b}$$

$$n^{-1}\mathbf{C}^T\mathbf{C}\mathbf{W}\mathbf{b} = \rho\mathbf{b}; \mathbf{b}^T\mathbf{W}\mathbf{b} = 1$$

$$n^{-1}\mathbf{C}^T\mathbf{C}\mathbf{W}\mathbf{b} = \rho\mathbf{b}; \text{subjective to } \mathbf{b}^T\mathbf{W}\mathbf{b} = 1$$

$$n^{-1}\mathbf{C}^T\mathbf{C}\mathbf{W}\mathbf{b} = \rho\mathbf{b}; \text{subjective to } \mathbf{b}^T\mathbf{W}\mathbf{b} = 1$$

Substitute $\mathbf{u} = \mathbf{W}^{1/2}\mathbf{b}$

$$n^{-1}\mathbf{C}^T\mathbf{CWb} = \rho\mathbf{b}; \text{subjective to } \mathbf{b}^T\mathbf{Wb} = 1$$

Substitute $\mathbf{u} = \mathbf{W}^{1/2}\mathbf{b}$

$n^{-1}\mathbf{W}^{1/2}\mathbf{C}^T\mathbf{CW}^{1/2}\mathbf{W}^{1/2}\mathbf{b} = \rho\mathbf{W}^{1/2}\mathbf{b}$

$$n^{-1}\mathbf{C}^T\mathbf{C}\mathbf{W}\mathbf{b} = \rho\mathbf{b}; \text{subjective to } \mathbf{b}^T\mathbf{W}\mathbf{b} = 1$$

Substitute $\mathbf{u} = \mathbf{W}^{1/2}\mathbf{b}$

$n^{-1}\mathbf{W}^{1/2}\mathbf{C}^T\mathbf{C}\mathbf{W}^{1/2}\mathbf{W}^{1/2}\mathbf{b} = \rho\mathbf{W}^{1/2}\mathbf{b}$

$n^{-1}\mathbf{W}^{1/2}\mathbf{C}^T\mathbf{C}\mathbf{W}^{1/2}\mathbf{u} = \rho\mathbf{u}$; subjective to $\mathbf{u}^T\mathbf{u} = 1$

$$n^{-1}\mathbf{C}^T\mathbf{C}\mathbf{W}\mathbf{b} = \rho\mathbf{b}; \text{subjective to } \mathbf{b}^T\mathbf{W}\mathbf{b} = 1$$

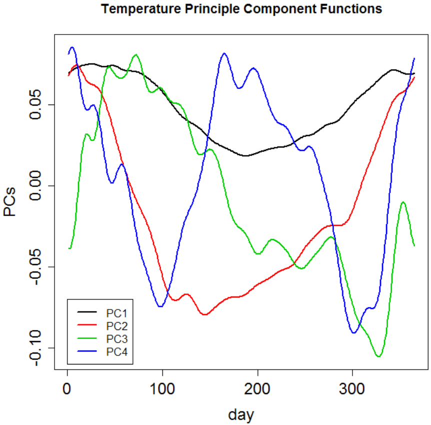
Substitute $\mathbf{u} = \mathbf{W}^{1/2}\mathbf{b}$

$n^{-1}\mathbf{W}^{1/2}\mathbf{C}^T\mathbf{C}\mathbf{W}^{1/2}\mathbf{W}^{1/2}\mathbf{b} = \rho\mathbf{W}^{1/2}\mathbf{b}$

$n^{-1}\mathbf{W}^{1/2}\mathbf{C}^T\mathbf{C}\mathbf{W}^{1/2}\mathbf{u} = \rho\mathbf{u}$; subjective to $\mathbf{u}^T\mathbf{u} = 1$

$\mathbf{u}$ is the eigenvector of the matrix $n^{-1}\mathbf{W}^{1/2}\mathbf{C}^T\mathbf{C}\mathbf{W}^{1/2}$

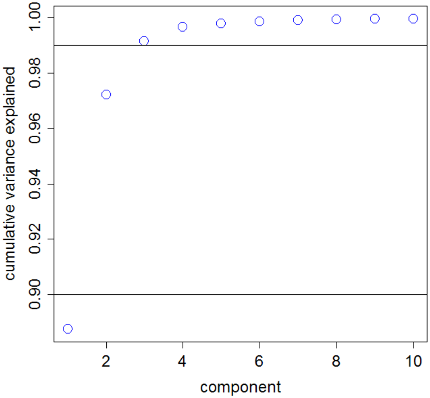
Canadian Temperature Data



- PC1 – over-all temperature
- PC2 – relative temperature of winter and summer
- PC3 – contrast between fall and spring
- PC4 – relative lengths of summer/winter

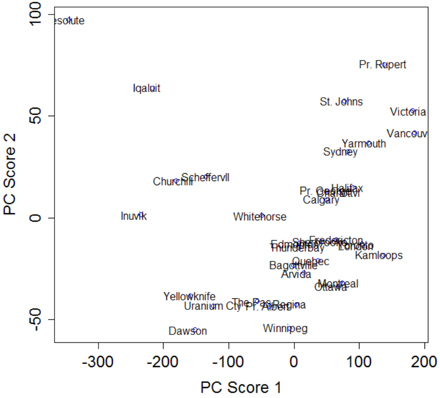
Canadian Temperature Data

We can alternatively calculate how many components are needed to capture 90% of the total variation in the data.



Canadian Temperature Data

Sanity check: we can plot the first two PC scores for each observation.

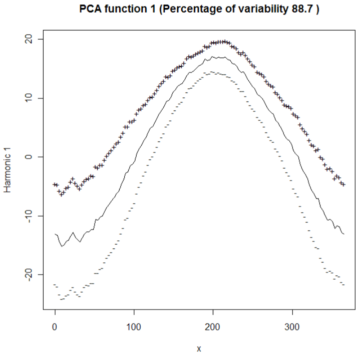


- First PC: over-all temperature.
- Second PC: contrast between Summer and Winter.

Display of Principal Components

Best way to obtain an idea of variation for each component is to plot

$$\bar{x}(t) \pm 2\sqrt{d_i} w_i(t)$$



Summary

- PCA = means of summarizing high dimensional covariation
- fPCA = extension to infinite-dimensional covariation
- Representation in terms of basis functions for fast(er) computation