Functional Data Analysis

Tools for exploring functional data and Functional PCA

Hyunsung Kim

Department of Statistics
Chung-Ang University

April 26, 2019

Outline

Outline

Introduction

- FDA의 notation과 concept 정의
- FDA에서 사용하는 statistics 정의
- Matrix decompositions, projections, and the constrained maximization of quadratic forms에 대한 자세한 내용은 Appendix 참고

Outline

Scalars, vectors, functions and matrices

- x : a vector
  $\Rightarrow x_i$: scalar (the elements of vector x)
- x : a function
  $\Rightarrow x(t)$: scalar (the values of function x)
  $\Rightarrow x(\mathbf{t})$: vector $\mathbf{t}$에 대한 function value (p-dim function)
- If x_i or $x(t)$ is a vector, we use $\boldsymbol{x}_i$ or $\boldsymbol{x}(t)$
- Standard notation을 요약하여 사용
  - `Temp` : a temperature record
  - `Knee` : a knee angle
  - `LMSSE` : a squared error fitting criterion for a linear model
  - `RSQ` : a squared correlation measure.

Derivatives and integrals

- D : $operator$ (함수 x를 함수 Dx로 변환하는 $operator$)
- $D^m x$: the derivative of order m of a function x ($\frac{d^m x}{dt^m}$와 동일)
- $D^0 x$: x
  $s.t.\ D^1 D^{-1} x = D^0 x = x$,
  when $D^{-1} x$가 x의 부정적분(indefinite integral)
- $\int x$: $\int_a^b x(t)dt$ (t의 적분 범위가 clear할 때)

Inner products

- Inner product for functions

$$\langle x, y \rangle = \int x(t)y(t)dt$$

- L_2 norm

$$\|x\|^2 = \langle x, x \rangle = \int x^2(t)dt$$

Functions of functions

- functional composition (합성함수)

$$x^* = x \circ h$$

- function value

$$x^*(t) = (x \circ h)(t) = x[h(t)]$$

- inverse function h^{-1}

$$(h \circ h^{-1})(t) = (h^{-1} \circ h)(t) = t$$

- functional transformations *operations* (or *operators*)
  ex) $D : x \to Dx$

Outline

Functional means and variances

- Mean function

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$$

- Variance function

$$Var_X(t) = \frac{1}{N-1}\sum_{i=1}^{N}[x_i(t) - \bar{x}(t)]^2$$

Functional means and variances

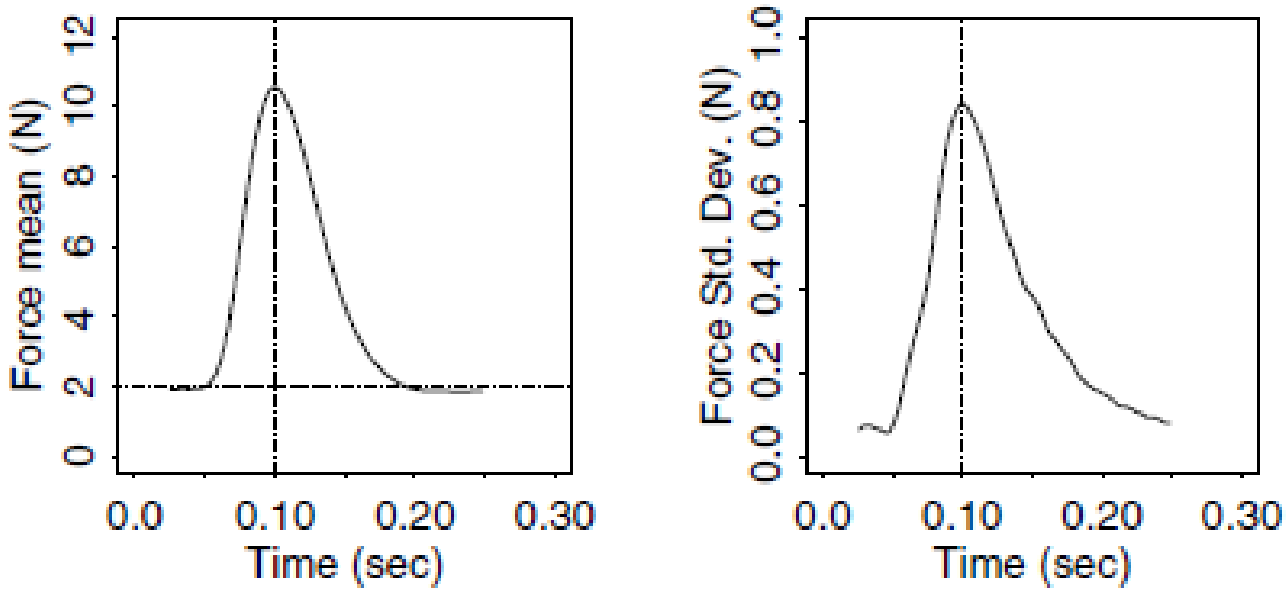


Figure 2.1. The mean and standard deviation functions for the 20 pinch force observations in Figure 1.11 after they were aligned or registered.

Covariance and correlation functions

- Covariance function

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

- Correlation function

$$Corr_X(t_1, t_2) = \frac{Cov_X(t_1, t_2)}{\sqrt{Var_X(t_1) Var_X(t_2)}}$$

Covariance and correlation functions

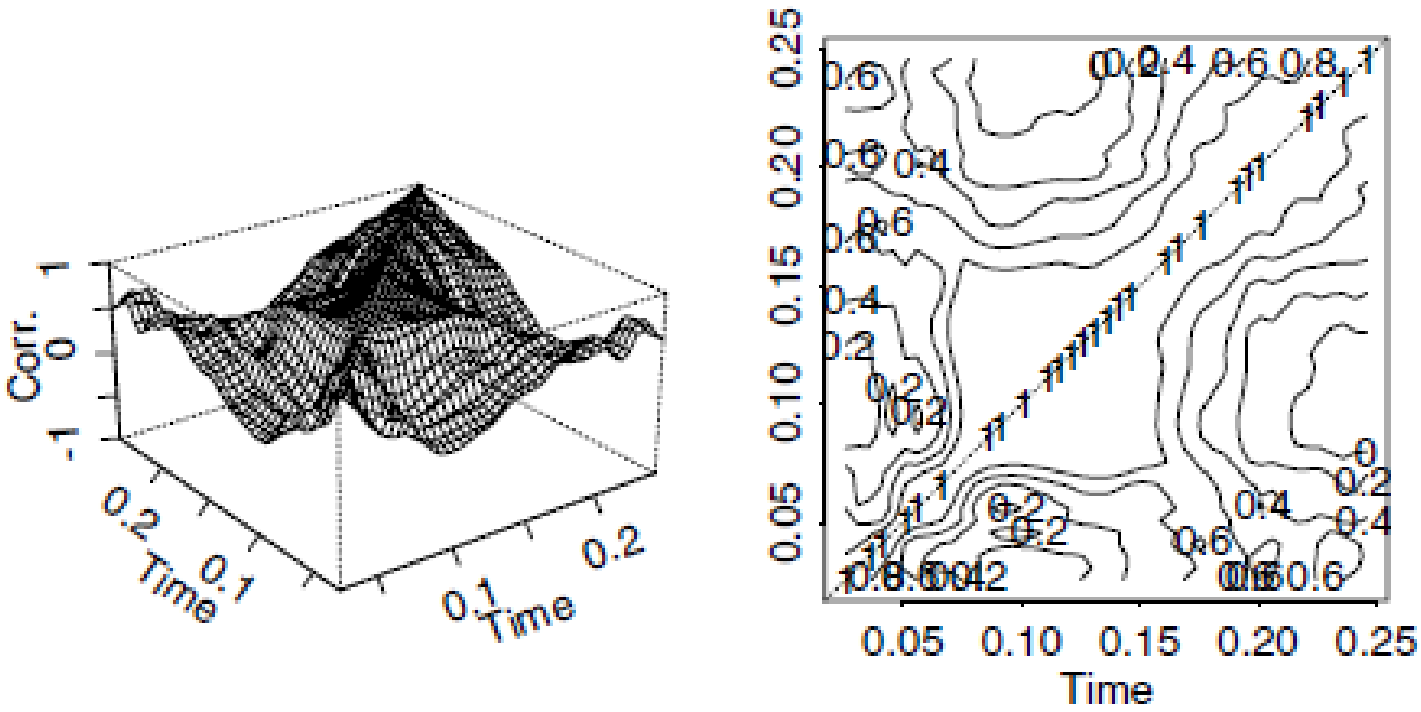


Figure 2.2. The left panel is a perspective plot of the bivariate correlation function values $r(t_1, t_2)$ for the pinch force data. The right panel shows the same surface by contour plotting. Time is measured in seconds.

Cross-covariance and cross-correlation functions

- Cross-covariance function

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

- Cross-correlation function

$$Corr_{X,Y}(t_1, t_2) = \frac{Cov_{X,Y}(t_1, t_2)}{\sqrt{Var_X(t_1) Var_Y(t_2)}}$$

Cross-covariance and cross-correlation functions

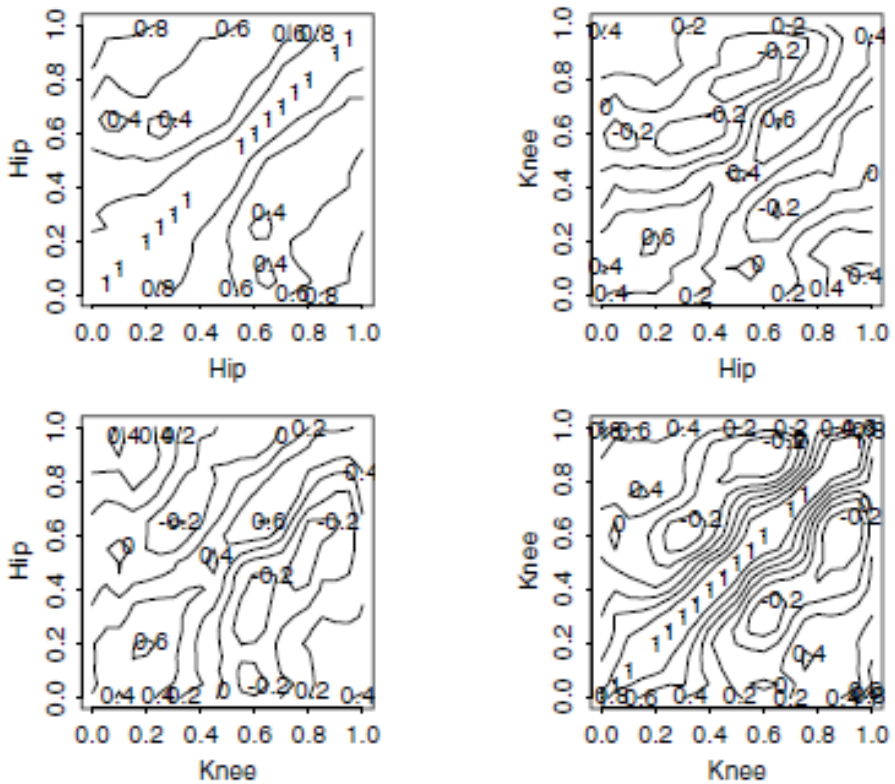
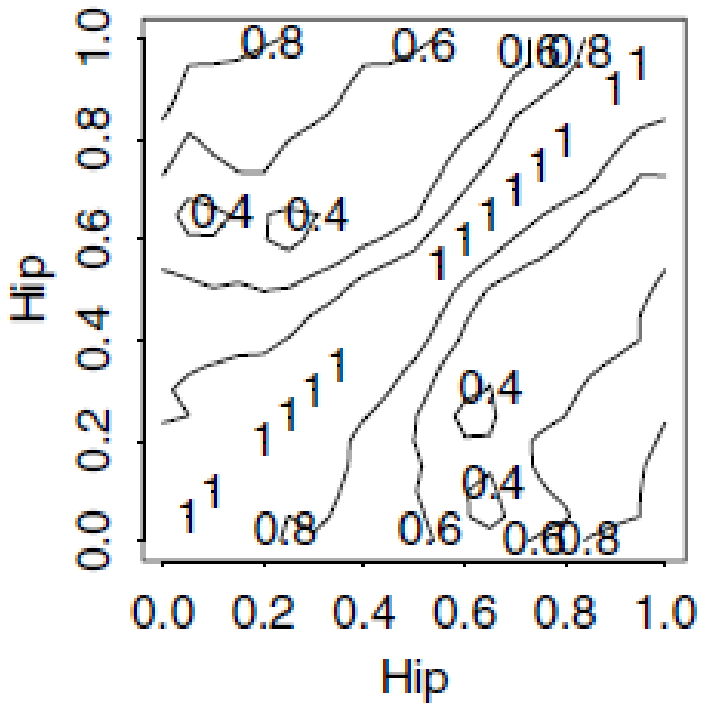
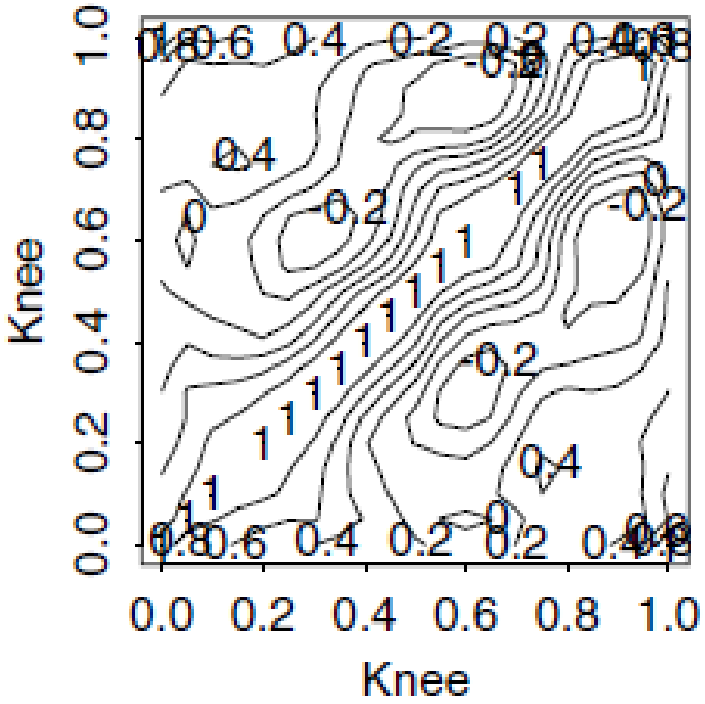


Figure 2.3. Contour plots of the correlation and cross-correlation functions for the gait data. In each panel t_1 is plotted on one axis and t_2 on the other; the legends indicate which observations are being correlated against each other.

Cross-covariance and cross-correlation functions

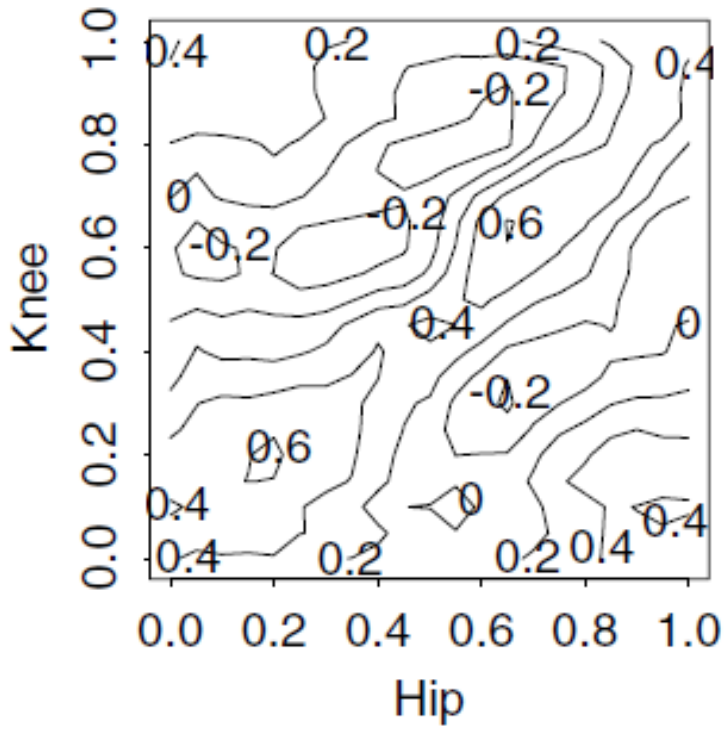
Cantour plots of correlation functions

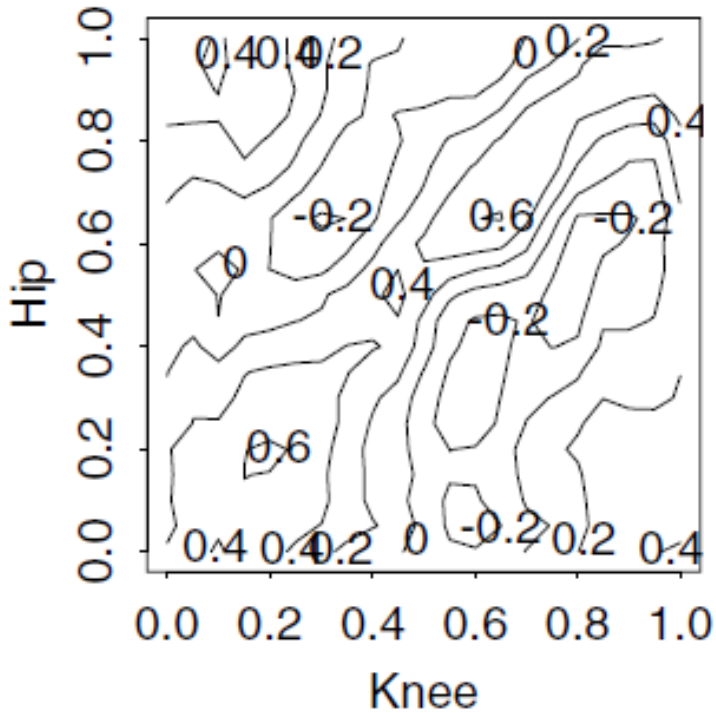




Cross-covariance and cross-correlation functions

Cantour plots of cross-correlation functions





Cross-covariance and cross-correlation functions

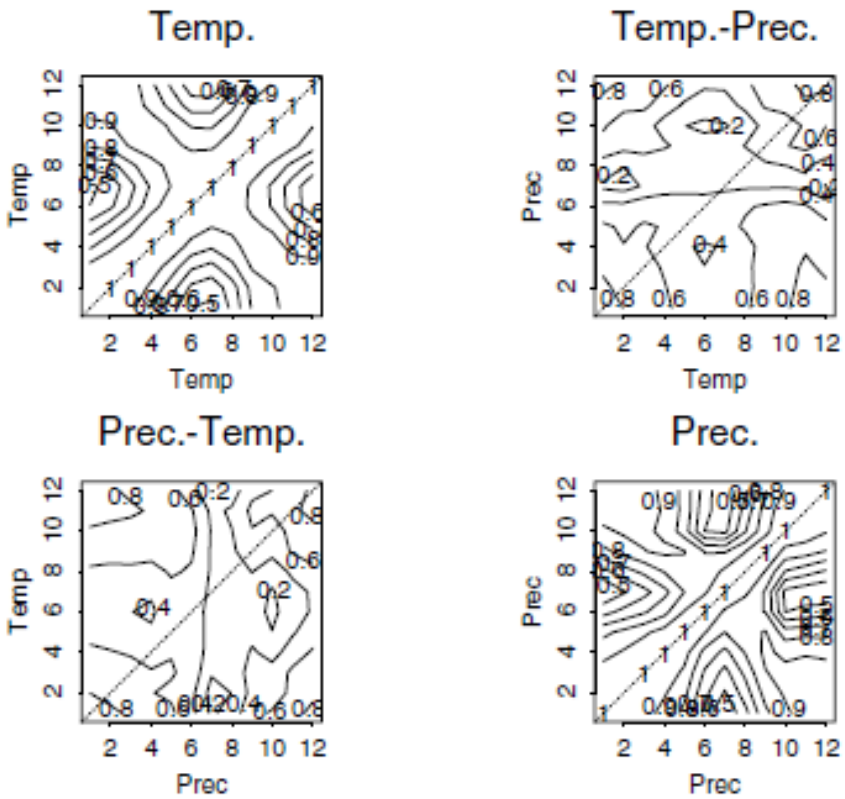
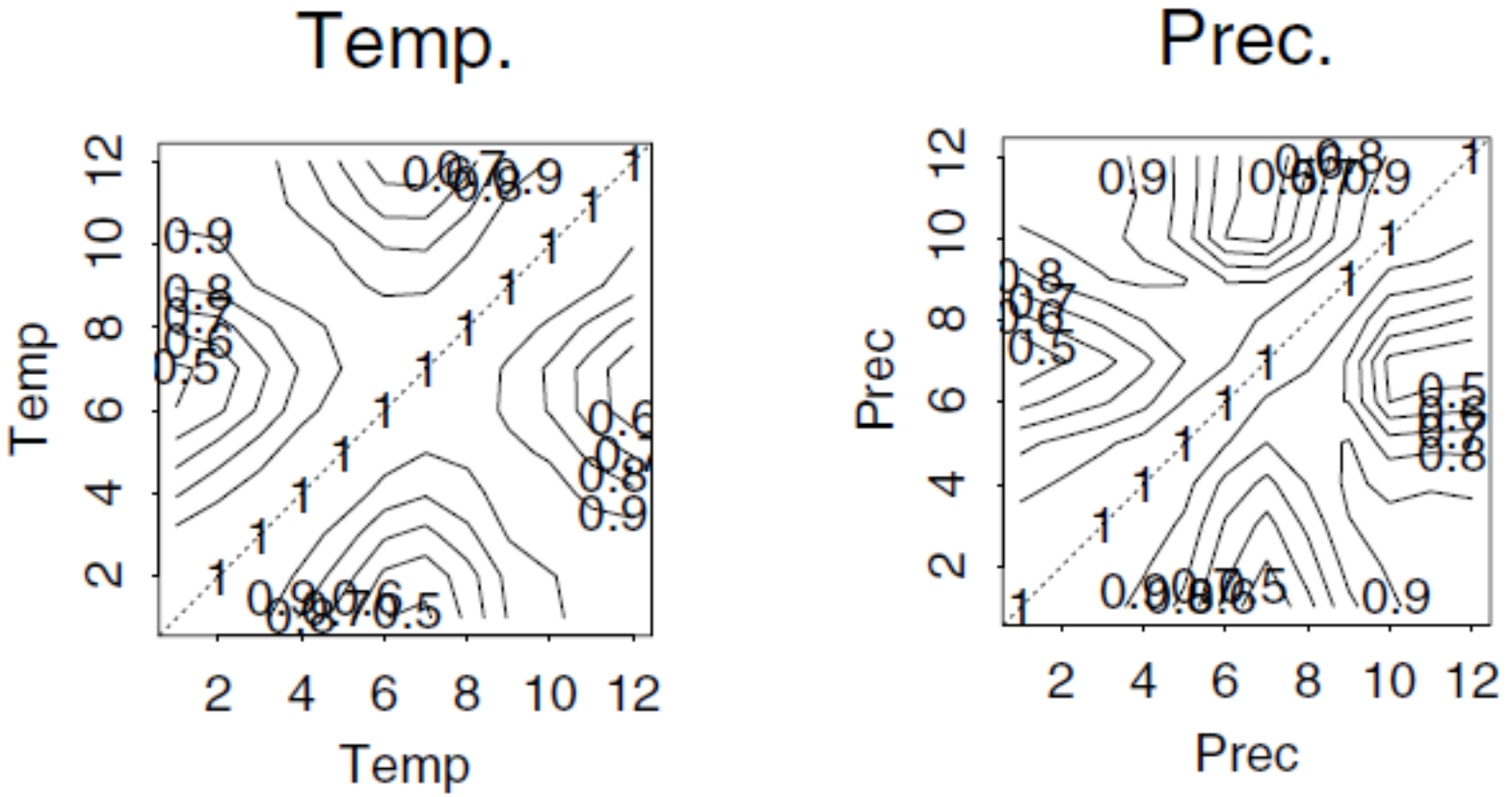


Figure 2.4. Contour plots of the correlation and cross-correlation functions for 35 Canadian weather stations for temperature and log precipitation. The cross-correlation functions are those in the upper right and lower left panels.

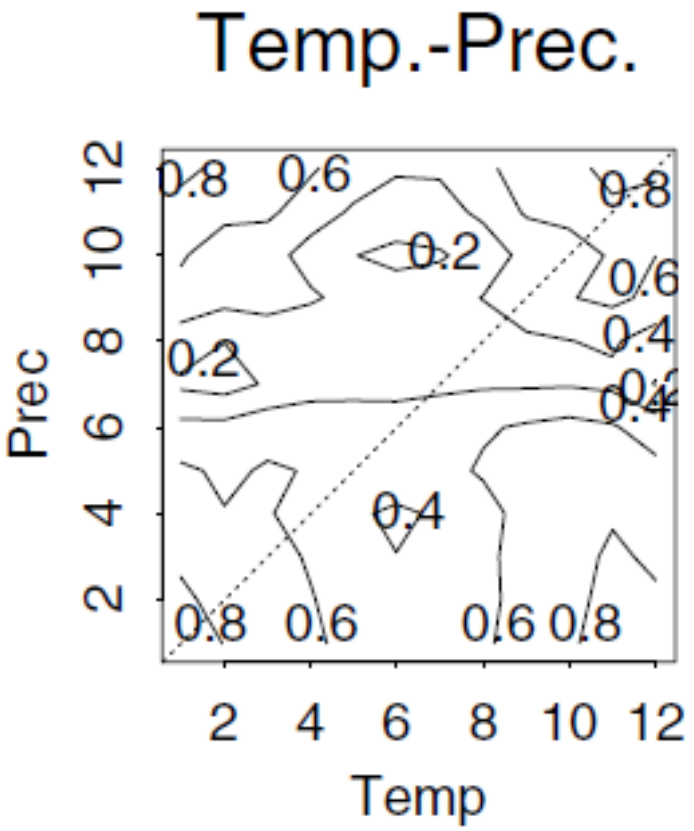
Cross-covariance and cross-correlation functions

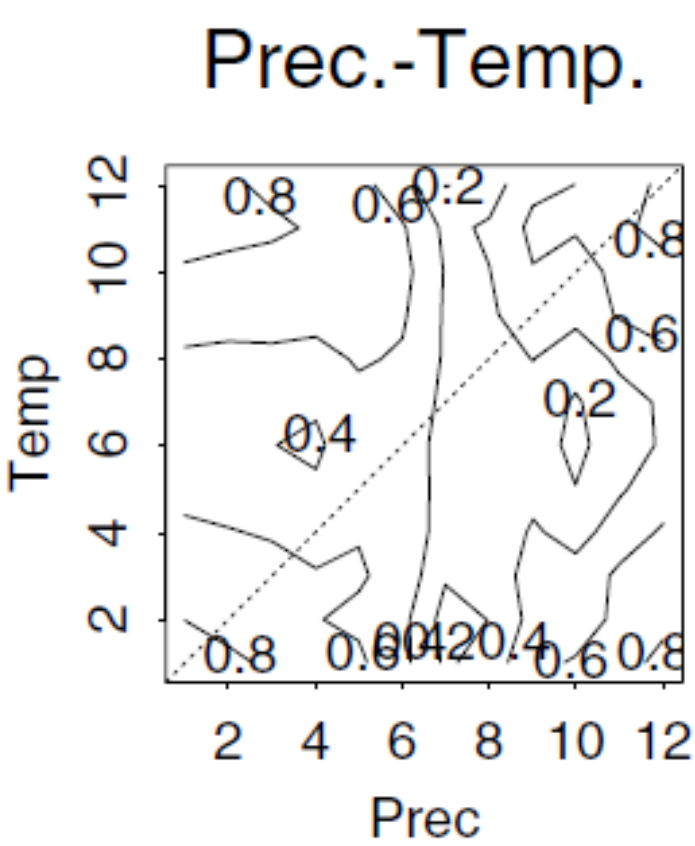
Cantour plots of correlation functions



Cross-covariance and cross-correlation functions

Cantour plots of cross-correlation functions





Outline

Introduction

- 전처리와 시각화 후, data의 특성을 파악하기 위해 PCA를 사용
- Classical multivariate anlaysis에서는 variance-covariance와 correlation을 설명하기 힘든 경우가 많음
- PCA는 유용한 정보를 담고 있는 covariance structure를 파악하는데 도움을 준다.
- PCA를 통해 이후의 분석에서 발생할 수 있는 문제를 사전에 고려할 수 있다. (ex - multicollinearity)
- FPCA(functional PCA)는 smoothing 되어있는 경우에 특성이 더 잘 나타난다. (smoothing 과정에서 $regularization$ issue 발생)

Outline

PCA for multivariate data

Concept of multivariate PCA

- Linear combination of X

$$f_i = \sum_{j=1}^{p} \beta_j x_{ij}, \ i = 1, ..., N$$

where β_j : weighting coefficient, x_{ij} : ith obs of jth variable

- Vectorized form

$$f_i = \boldsymbol{\beta}' \boldsymbol{x}_i, \ i = 1, ..., N$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)', \ \boldsymbol{x}_i = (x_{i1}, ..., x_{ip})'$

PCA for multivariate data

How to find PC

- 1. Find the weight vector $\xi_1 = (\xi_{11}, ..., \xi_{p1})'$ for

$$f_{i1} = \sum_j \xi_{j1} x_{ij} = \boldsymbol{\xi}_1' \boldsymbol{x}_i$$

s.t. maximize $\frac{1}{N}\sum_i f_{i1}^2$ subject to $\sum_j \xi_{j1}^2 = \|\xi_1\|^2 = 1$

- 2. 1번 과정을 반복하며 동시에

$$\sum_j \xi_{jk}\xi_{jm} = \boldsymbol{\xi}_k'\boldsymbol{\xi}_m = 0, \ k < m$$

을 만족하는 $\xi_2, ..., \xi_m$을 찾는다.

PCA for multivariate data

Summary

- 1. Mean square(변수들 간 variation)를 maximize하는 방향의 unit vector $\boldsymbol{\xi}_1$을 찾는다.
- 2. 2nd PC부터는 mean square를 maximize함과 동시에 이전 PC loading($\boldsymbol{\xi}_i$)과 orthogonal한 $\boldsymbol{\xi}_2, ...\boldsymbol{\xi}_k$ $(k < p)$을 찾는다.
- Data의 mean을 뺀 후에 PCA를 하는 것이 일반적이다. (Centering $\Rightarrow \max MS(f_{ij}) = \max Var(f_{ij})$)
- Weight vector $\boldsymbol{\xi}_i$는 unique하지 않다. (Sign change)
- PC score f_{im}은 특정 사례 또는 반복실험의 특징 측면에서 변동(variation)의 의미를 설명하는데 도움을 준다.

Defining PCA for functional data

Concept of functional PCA

- Inner product of integration version is defined by

$$\int \beta x = \int \beta(s)x(s)ds$$

- PC score

$$f_i = \int \beta x_i = \int \beta(s)x_i(s)ds$$

where β : weight function

Defining PCA for functional data

How to find functional PCA

- 1. Find the weight function $\xi_1(s)$ for

$$f_i = \int \xi_1(s) x_i(s) ds$$

s.t. maximize $\frac{1}{N}\sum_i f_{i1}^2 = \frac{1}{N}\sum_i (\int \xi_1 x_i)^2$
subject to $\|\xi_1\|^2 = \int \xi_1(s)^2 ds = \int \xi_1^2 = 1$

- 2. 1번 과정을 반복하며 동시에

$$\int \xi_k \xi_m = 0, \ k < m$$

을 만족하는 $\xi_2, ..., \xi_m$을 찾는다.

Defining PCA for functional data

Summary

- 1. Mean square를 maximize하는 방향이고 $\|\xi_1\|^2 = 1$인 function $\xi_1(s)$를 찾는다.
- 2. 2nd PC부터는 mean square를 maximize함과 동시에 이전 PC loading($\xi_1(s)$)와 orthogonal한 $\xi_2(s), ...\xi_k(s)$ $(k < p)$을 찾는다.
- Data의 mean을 뺀 후에 PCA를 하는 것이 일반적이다. (Centering $\Rightarrow \max MS = \max Var$)
- Weight function $\xi_i(s)$는 unique하지 않다. (Sign change)

Defining PCA for functional data

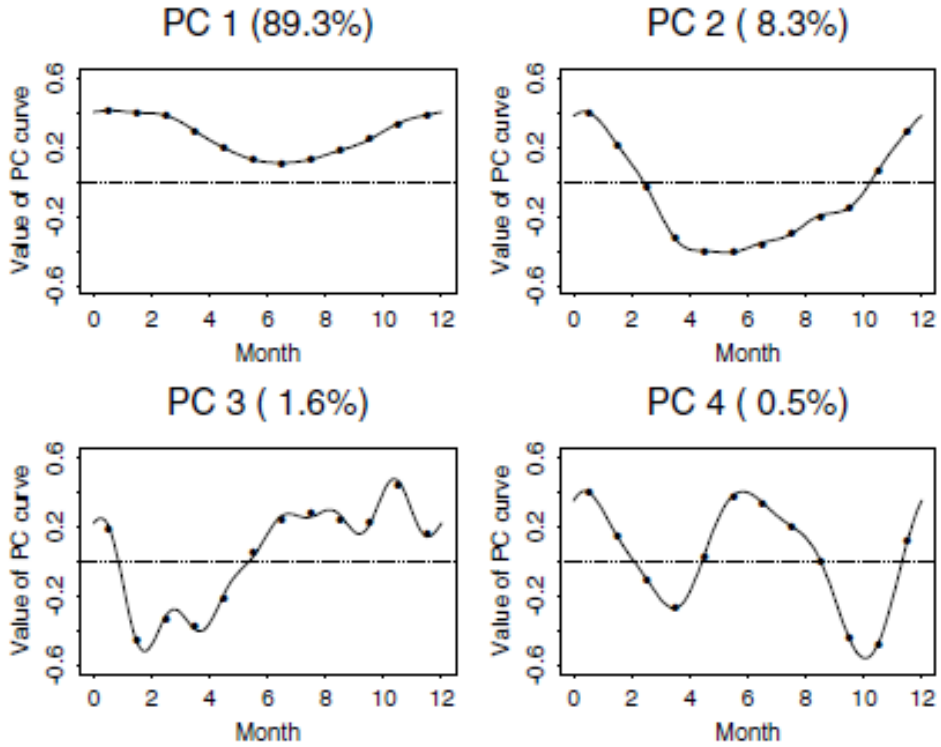


Figure 8.1. The first four principal component curves of the Canadian temperature data estimated by two techniques. The points are the estimates from the discretization approach, and the curves are the estimates from the expansion of the data in terms of a 12-term Fourier series. The percentages indicate the amount of total variation accounted for by each principal component.

Defining an optimal empirical orthonormal basis

- We want to find K orthonormal functions ξ_m.
- 즉, expansion했을 때 각 curve에 가장 잘 근사하는 K개의 orthonormal basis functions를 찾고 싶다!
- Expansion by the orthonormal basis functions

$$\hat{x}_i(t) = \sum_{k=1}^{K} f_{ik}\xi_k(t),$$

where f_{ik} is the principal component value $\int x_i\xi_k$

Defining an optimal empirical orthonormal basis

- Measure of approximation (PCASSE)

$$\texttt{PCASSE} = \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2$$

where $\|x_i - \hat{x}_i\|^2 = \int [x(s) - \hat{x}(s)]^2 ds$ (integrated squared error)

- Optimal orthonormal basis function = weight function ξ_m

$$\xi_m = \arg\min_{\xi} \texttt{PCASSE}$$

where ξ_m : $empirical\ orthonormal\ functions$

PCA and eigenanalysis

Multivariate PCA

- Assumtion : x_{ij} is centerized. $(x_{ij} - \frac{1}{N}\sum_i x_{ij})$
- Mean square criterian for finding the 1st PC

$$\max_{\boldsymbol{\xi}'\boldsymbol{\xi}=1} \frac{1}{N}\boldsymbol{\xi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\xi}$$

- Substitute variance-covariance matrix

$$\max_{\boldsymbol{\xi}'\boldsymbol{\xi}=1} \boldsymbol{\xi}'\boldsymbol{V}\boldsymbol{\xi}$$

where $\boldsymbol{V} = N^{-1}\boldsymbol{X}'\boldsymbol{X}$ is a $p \times p$ sample var-cov matrix
$\Rightarrow$ We can solve maximization problem using eigen decomposition!

PCA and eigenanalysis

Multivariate PCA

- Eigen equation

$$\boldsymbol{V}\boldsymbol{\xi} = \rho\boldsymbol{\xi}$$

  where ρ is largest eigen value
- 위 식을 풀면 $(\rho_j, \boldsymbol{\xi}_j)$ pairs가 생기고, 각 $\boldsymbol{\xi}_j$는 orthogonal하다.
- $\boldsymbol{V}$ has $\min\{p, N-1\}$ nonzero eigen values ρ_j
  $(\because\ max(rank(\boldsymbol{X})) = N-1)$
- $\boldsymbol{\xi}_j$ satisfied maximization problem and the orthogonal constraints $(\xi_j \perp (\xi_1, \cdots, \xi_{j-1}))$ for $\forall j$
  $\Rightarrow \boldsymbol{\xi}$ is a solution of PCA

PCA and eigenanalysis

Functional PCA

- Assumtion : $x_i(t)$ is centerized. $(x_i(t) - \frac{1}{N}\sum_i x_i(t))$
- Covariance function

$$v(s,t) = \frac{1}{N}\sum_{i=1}^{N} x_i(s)x_i(t)$$

PCA and eigenanalysis

Functional PCA

- Each of PC weight functions $\xi_j(s)$ satisfies

$$\int v(s,t)\xi(t)dt = \rho\xi(s)$$

where LHS is an $integral\ transform\ V$ of the weight function ξ defined by

$$V\xi = \int v(\cdot,t)\xi(t)dt\ (covariance\ operator\ V)$$

- Eigen equation using $covariance\ operator\ V$

$$V\xi = \rho\xi$$

where ξ is an eigen function

PCA and eigenanalysis

Difference between multivariate and functional eigen analysis problems

- max{# of different eigen pairs}가 다르다
  - multivariate : # of variables $= p$
  - functional : # of functions $= \infty$ ($\because$ smoothed)
    but if x_i are linearly independent, $rank(V) = N - 1$ and only $N - 1$ nonzero eigen values exist.

Outline

Summary

Comparison between MPCA and FPCA

Multivariate PCA

- PC score
  $f_i = \sum_{j=1}^{p} \xi_j x_{ij}$
- Objective function
  $Var(\xi_j) = \frac{1}{N} \sum_i f_{ij}^2$
- Constraints
  $\|\xi_i\|^2 = \sum_j \xi_{ji}^2 = 1$
  $\sum_j \xi_{jk}\xi_{jm} = \boldsymbol{\xi}_k^{'}\boldsymbol{\xi}_m = 0$

Functional PCA

- PC score
  $f_i = \int \xi(s) x_i(s) ds$
- Objective function
  $\frac{1}{N} \sum_i f_{ij}^2 = \frac{1}{N} \sum_i (\int \xi_j x_i)^2$
- Constarints
  $\|\xi_i\|^2 = \int \xi_i(s)^2 ds = 1$
  $\int \xi_k \xi_m = 0$

Reference


J.O. Ramsay, B.W. Silverman.
Functional Data Analysis 2nd edition.
Springer, 2005.