Functional Generalized Canonical Correlation Analysis for Studying Multiple Longitudinal Variables

Biopuces

Lucas Sort, Arthur Tenenhaus, Laurent Le Brusquet

Université Paris-Saclay, CentraleSupélec, CNRS, Laboratoire des Signaux et Sytèmes (L2S)

March 13, 2025

Outline

Introduction

Functional Generalized Canonical Correlation Analysis

Perspectives and limitations

References and appendix

Outline

Introduction

Functional Generalized Canonical Correlation Analysis

Perspectives and limitations

References and appendix

Context

Example

Medical study focusing on primary biliary cholangitis. Blood albumin (mg/dl) observed over several years.

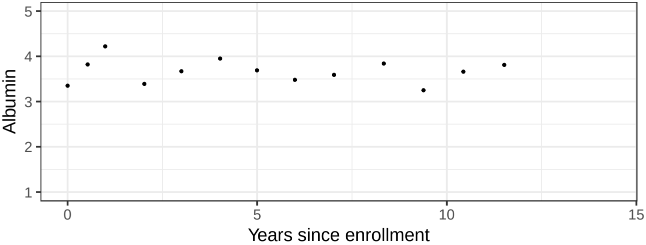

Figure: Albumin observations at multiple time points.

Longitudinal data

Definition

Data obtained from repeated measurements of a variable, typically made over time and, often, on multiple individuals.

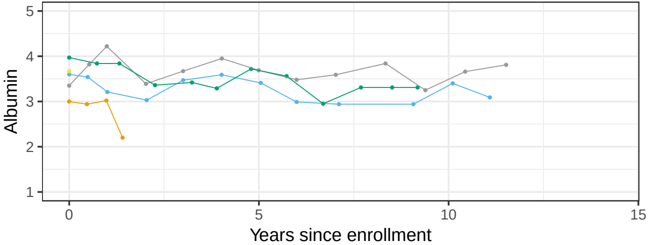

Figure: Multiple albumin trajectories

→ Number of observations and observation time points may differ!

Longitudinal data analysis

We denote

- i the subject number with $i = 1, \dots, I$
- k the observation number $k = 1, \dots, K_i$
- y_{ik} the kth observation of ith subject
- t_{ik} the time point of observation y_{ik}

Longitudinal data analysis

We denote

- i the subject number with $i = 1, \ldots, I$
- k the observation number $k = 1, \ldots, K_i$
- y_{ik} the kth observation of ith subject
- t_{ik} the time point of observation y_{ik}

Various frameworks are used to study longitudinal data:

- Linear mixed models [Laird and Ware, 1982],
- Functional data analysis [Ramsay and Silverman, 2005].

Functional data

Definition

Data that has the form of function, i.e., sampled from some underlying smooth function (defined on $\mathcal{I} \subset \mathbb{R}$), sampled from a random process X

$$y_{ik} = X_i(t_{ik}) + \varepsilon_{ik},\ X_i \in L^2(\mathcal{I}),$$

where we define $L^2(\mathcal{I}) = \left\{ f : \mathcal{I} \to \mathbb{R}, \|f\|^2 = \int_{\mathcal{I}} f^2(s)\mathrm{d}s < \infty \right\}$

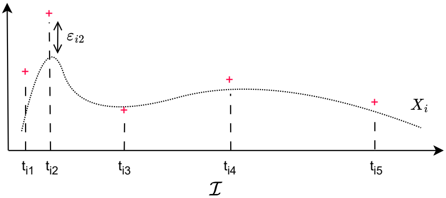

Figure: Observations and underlying function

Functional data analysis

Functional Principal Component Analysis

Considering a random process X

$$\underset{f \in L^2(\mathcal{I}),\ \|f\|=1}{\operatorname{argmax}} \operatorname{var}(\langle f, X \rangle)$$

Functional data analysis

Functional Principal Component Analysis

Considering a random process X

$$\underset{f \in L^2(\mathcal{I}),\ \|f\|=1}{\operatorname{argmax}} \ \operatorname{var}(\langle f, X \rangle)$$

Functional Regression

Considering a random process X and a response Y

$$\underset{f \in L^2(\mathcal{I}),\ \|f\|=1}{\operatorname{argmin}} \ \mathbb{E}\|Y - \langle f, X \rangle\|$$

Functional data analysis

Functional Principal Component Analysis

Considering a random process X

$$\underset{f \in L^2(\mathcal{I}),\ \|f\|=1}{\operatorname{argmax}} \operatorname{var}(\langle f, X \rangle)$$

Functional Regression

Considering a random process X and a response Y

$$\underset{f \in L^2(\mathcal{I}),\ \|f\|=1}{\operatorname{argmin}} \mathbb{E}\|Y - \langle f, X \rangle\|$$

Many rely on $\Sigma_{XX}(s,t) = \operatorname{cov}(X(s), X(t))$ and the associated operator

$$\boldsymbol{\Sigma}_{XX} : L^2(\mathcal{I}) \to L^2(\mathcal{I})\ ,\ f \mapsto g : g(t) = \int_{\mathcal{I}} \Sigma_{XX}(s,t) f(s) \mathrm{d}s.$$

Functional data analysis

Functional Principal Component Analysis

Considering a random process X

$$\underset{f \in L^2(\mathcal{I}),\ \|f\|=1}{\operatorname{argmax}} \ \operatorname{var}(\langle f, X\rangle)$$

$\rightarrow$ Solution derived from eigendecomposition of $\mathbf{\Sigma}_{XX}$

Functional Regression

Considering a random process X and a response Y

$$\underset{f \in L^2(\mathcal{I}),\ \|f\|=1}{\operatorname{argmin}} \ \mathbb{E}\|Y - \langle f, X\rangle\|$$

Many rely on $\Sigma_{XX}(s,t) = \operatorname{cov}(X(s), X(t))$ and the associated operator

$$\mathbf{\Sigma}_{XX} : L^2(\mathcal{I}) \rightarrow L^2(\mathcal{I}) \ , \ f \mapsto g : g(t) = \int_{\mathcal{I}} \Sigma_{XX}(s,t) f(s) \mathrm{d}s.$$

Functional data analysis

Functional Principal Component Analysis

Considering a random process X

$$\underset{f \in L^2(\mathcal{I}),\ \|f\|=1}{\operatorname{argmax}} \operatorname{var}(\langle f, X\rangle)$$

$\rightarrow$ Solution derived from eigendecomposition of $\boldsymbol{\Sigma}_{XX}$

Functional Regression

Considering a random process X and a response Y

$$\underset{f \in L^2(\mathcal{I}),\ \|f\|=1}{\operatorname{argmin}} \mathbb{E}\|Y - \langle f, X\rangle\|$$

$\rightarrow$ Solution involves to $\boldsymbol{\Sigma}_{XX}^{-1}$

Many rely on $\Sigma_{XX}(s,t) = \operatorname{cov}(X(s), X(t))$ and the associated operator

$$\boldsymbol{\Sigma}_{XX} : L^2(\mathcal{I}) \rightarrow L^2(\mathcal{I})\ ,\ f \mapsto g : g(t) = \int_{\mathcal{I}} \Sigma_{XX}(s,t) f(s) \mathrm{d}s.$$

Covariance estimation

- Local linear smoothing estimation [Fan and Gijbels, 2018]
  - Convergence guarantees
  - Robust to sparse and irregular sampling

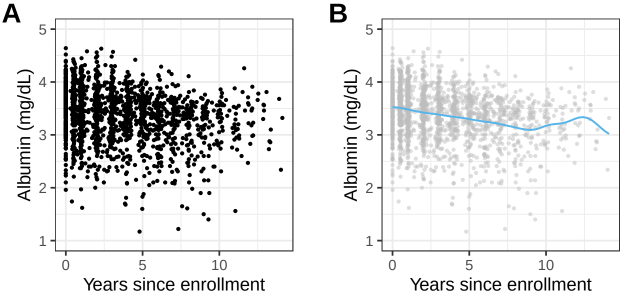

Figure: Mean function $\mu(s) = \mathbb{E}[X(s)]$ estimation with local linear smoothing. (A) aggregated observations; (B) estimated mean function.

Covariance estimation

- Local linear smoothing estimation [Fan and Gijbels, 2018]
  - Convergence guarantees
  - Robust to sparse and irregular sampling

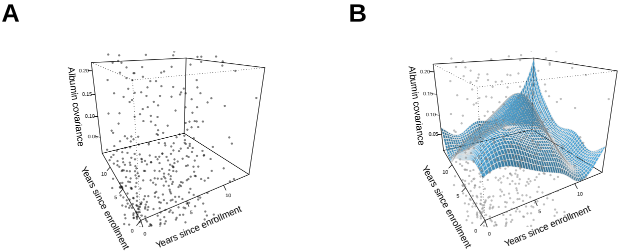

Figure: Covariance surface estimation with local linear smoothing. (A) subset of aggregated raw covariances; (B) estimated covariance surface.

Covariance estimation

- Local linear smoothing estimation [Fan and Gijbels, 2018]
  - Convergence guarantees
  - Robust to sparse and irregular sampling
- Fast covariance estimation [Xiao et al., 2014, Xiao et al., 2017]

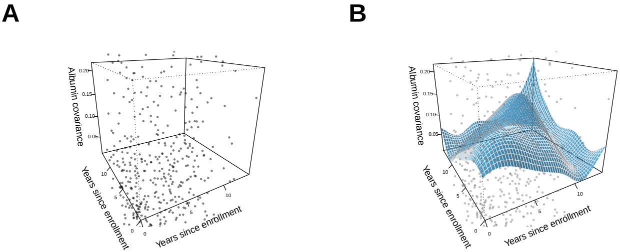

Figure: Covariance surface estimation with local linear smoothing. (A) subset of aggregated raw covariances; (B) estimated covariance surface.

Multiple biomarkers

Consider now not 1 but 3 blood markers: Blood albumin (mg/dl), Prothrombin time (s), and Bilirubin concentration (mg/dl)

Objective

- Investigating relationships between markers.
- Integrating associations to improve characterization.

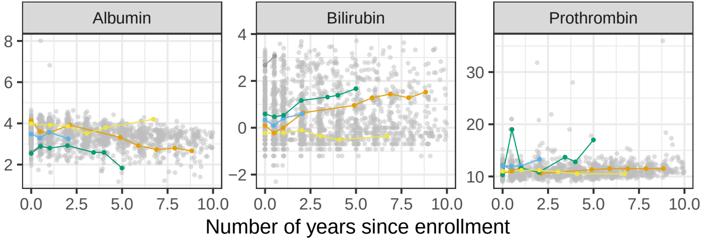

Figure: Observations of blood markers with first trajectories colored.

Outline

Introduction

Functional Generalized Canonical Correlation Analysis

Perspectives and limitations

References and appendix

Canonical Correlation Analysis

Definition

Considering 2 sets of random variables $\mathbf{x}_1 \in \mathbb{R}^{p_1}$, $\mathbf{x}_2 \in \mathbb{R}^{p_2}$. Investigating relationships between $\mathbf{x}_1$ and $\mathbf{x}_2$ with the CCA [Hotelling, 1936]:

$$\begin{aligned} \underset{\mathbf{a}_1, \mathbf{a}_2}{\operatorname{argmax}} \quad & \operatorname{corr}(\mathbf{a}_1^\top \mathbf{x}_1, \mathbf{a}_2^\top \mathbf{x}_2) \\ \text{s.t.} \quad & \operatorname{var}(\mathbf{a}_j^\top \mathbf{x}_j) = 1,\ \forall j \in \{1, 2\} \end{aligned}$$

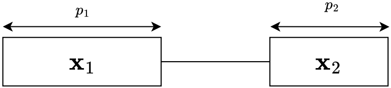

Figure: Canonical correlation analysis

$\mathbf{a}_1, \mathbf{a}_2 \rightarrow$ canonical vectors ; $\mathbf{a}_1^\top \mathbf{x}_1, \mathbf{a}_2^\top \mathbf{x}_2 \rightarrow$ canonical components

Canonical Correlation Analysis for 3 sets

Considering 3 sets $\mathbf{x}_1 \in \mathbb{R}^{p_1}$, $\mathbf{x}_2 \in \mathbb{R}^{p_2}$, $\mathbf{x}_3 \in \mathbb{R}^{p_3}$ and denoting their respective covariance matrices $\mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{22}$ and $\mathbf{\Sigma}_{33}$.

- Extending the analysis to multiple sets.

$$\underset{\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3}{\text{argmax}} \quad \text{corr}(\mathbf{a}_1^\top \mathbf{x}_1, \mathbf{a}_2^\top \mathbf{x}_2) + \text{corr}(\mathbf{a}_1^\top \mathbf{x}_1, \mathbf{a}_3^\top \mathbf{x}_3) + \text{corr}(\mathbf{a}_2 \top \mathbf{x}_2, \mathbf{a}_3^\top \mathbf{x}_3)$$

$$\text{s.t.} \quad \mathbf{a}_j^\top \mathbf{\Sigma}_{jj} \mathbf{a}_j = 1, \ \forall j \in \{1, 2, 3\}$$

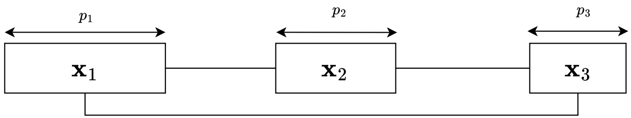

Figure: Canonical correlation analysis for 3 sets

Canonical Correlation Analysis for 3 sets

Considering 3 sets $\mathbf{x}_1 \in \mathbb{R}^{p_1}$, $\mathbf{x}_2 \in \mathbb{R}^{p_2}$, $\mathbf{x}_3 \in \mathbb{R}^{p_3}$ and denoting their respective covariance matrices $\mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{22}$ and $\mathbf{\Sigma}_{33}$.

- Extending the analysis to multiple sets.
- Selecting relationships to investigate.

$$\underset{\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3}{\text{argmax}} \quad \text{corr}(\mathbf{a}_1^\top \mathbf{x}_1, \mathbf{a}_2^\top \mathbf{x}_2) + \text{corr}(\mathbf{a}_1^\top \mathbf{x}_1, \mathbf{a}_3^\top \mathbf{x}_3)$$

$$\text{s.t.} \quad \mathbf{a}_j^\top \mathbf{\Sigma}_{jj} \mathbf{a}_j = 1, \ \forall j \in \{1, 2, 3\}$$

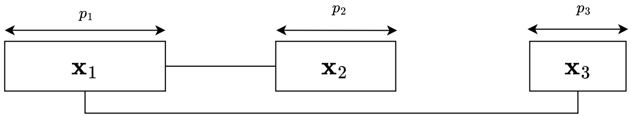

Figure: Hierarchical canonical correlation analysis for 3 sets

Canonical Correlation Analysis for 3 sets

Considering 3 sets $\mathbf{x}_1 \in \mathbb{R}^{p_1}$, $\mathbf{x}_2 \in \mathbb{R}^{p_2}$, $\mathbf{x}_3 \in \mathbb{R}^{p_3}$ and denoting their respective covariance matrices $\mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{22}$ and $\mathbf{\Sigma}_{33}$.

- Extending the analysis to multiple sets.
- Selecting relationships to investigate.
- Adjusting constraints $\mathbf{M}_j = \tau_j \mathbf{I}_{p_j} + (1 - \tau_j)\mathbf{\Sigma}_{jj}$

$$\underset{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3}{\operatorname{argmax}} \quad \operatorname{corr}(\mathbf{a}_1^\top \mathbf{x}_1, \mathbf{a}_2^\top \mathbf{x}_2) + \operatorname{corr}(\mathbf{a}_1^\top \mathbf{x}_1, \mathbf{a}_3^\top \mathbf{x}_3)$$

$$\text{s.t.} \quad \mathbf{a}_j^\top \mathbf{M}_j \mathbf{a}_j = 1,\ \forall j \in \{1, 2, 3\}$$

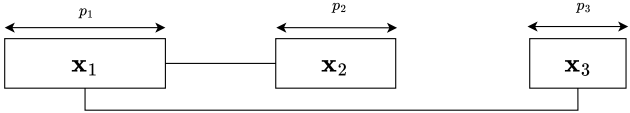

Figure: Hierarchical canonical correlation analysis for 3 sets

Regularized Generalized Canonical Correlation Analysis

Definition

Considering J sets $\mathbf{x}_1 \in \mathbb{R}^{p_1}, \dots, \mathbf{x}_J \in \mathbb{R}^{p_J}$. The RGCCA optimization problem [Tenenhaus et al., 2017] is defined as:

$$\underset{\mathbf{a}_1,\dots,\mathbf{a}_J \in \mathbb{R}^{p_1}\times\cdots\times\mathbb{R}^{p_J}}{\operatorname{argmax}} \sum_{j=1}^{J}\sum_{j'=1}^{J} c_{jj'}\, g(\operatorname{cov}(\mathbf{a}_j^\top \mathbf{x}_j, \mathbf{a}_{j'}^\top \mathbf{x}_{j'}))$$
$$\text{s.t.} \quad \mathbf{a}_j^\top \mathbf{M}_j \mathbf{a}_j = 1, \quad j \in \{1, \dots, J\}$$

where $\mathbf{C} = (c_{jj'}) \in \mathbb{R}^{J\times J}$ is the connection design matrix, g is a convex differentiable function, and $\mathbf{M}_j$ is a positive matrix which if often set to $\mathbf{M}_j = \tau_j \mathbf{I}_{p_j} + (1-\tau_j)\mathbf{\Sigma}_{jj}$ with $\tau_j \in [0, 1]$.

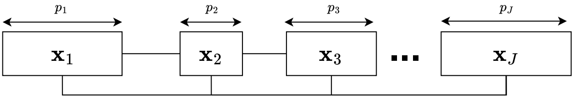

Figure: Regularized Generalized Canonical Correlation Analysis

RGCCA package in R

- `Russett` dataset: 3 blocks of socio-economic data for 47 countries in 1964: *agriculture inequality*, *industrial development*, *political stability*.
- RGCCA applied to find associations between the various modalities.

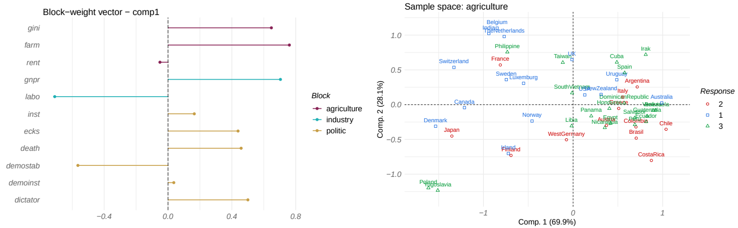

Figure: (right) First block weight vector values for each block (left) First vs. second component for agricultural inequality block.

Adapting to functional data

With multiple longitudinal markers:

- Modeling markers as random processes X_1, X_2, X_3

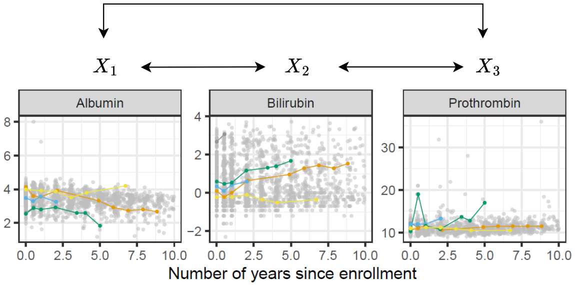

Figure: Investigating relationships between three random processes

Adapting to functional data

With multiple longitudinal markers:

- Modeling markers as random processes X_1, X_2, X_3
- Investigating relationships with RGCCA

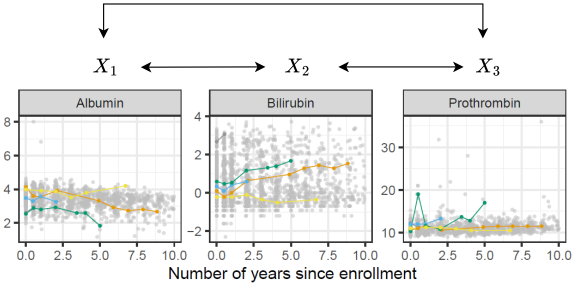

Figure: Investigating relationships between three random processes

Adapting to functional data

With multiple longitudinal markers:

- Modeling markers as random processes X_1, X_2, X_3
- Investigating relationships with RGCCA
- **Adapting RGCCA to functional data**

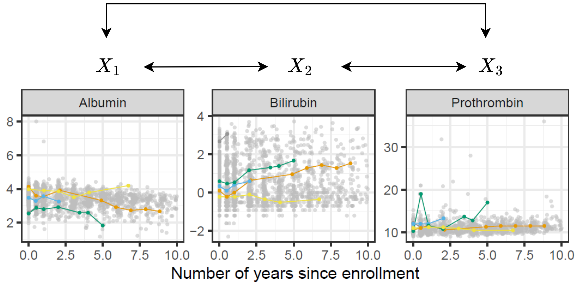

Figure: Investigating relationships between three random processes

Adapting to functional data

Moving the RGCCA framework to the functional setting:

- Random set $\mathbf{x}_j \rightarrow$ random process X_j (defined on $\mathcal{I}_j \subset \mathbb{R}$)
- Vector $\mathbf{a}_j \in \mathbb{R}^{p_j} \rightarrow$ function $f_j \in L^2(\mathcal{I}_j)$
- Dot product $\mathbf{a}_j^\top \mathbf{x}_j \rightarrow L^2(\mathcal{I}_j)$ product $\langle f_j, X_j \rangle = \int_{\mathcal{I}_j} f_j(s) X_j(s) \mathrm{d}s$
- Matrix $\mathbf{M}_j \in \mathbb{R}^{p_j \times p_j} \rightarrow$ Operator $\mathbf{M}_j : L^2(\mathcal{I}_j) \rightarrow L^2(\mathcal{I}_j)$

Functional Generalized Canonical Correlation Analysis

Definition

Considering J random processes $X_1, \dots, X_J$ defined on $\mathcal{I}_1, \dots, \mathcal{I}_J$ respectively. Introducing FGCCA optimization problem as:

$$\underset{f_1,\dots,f_J \in L^2(\mathcal{I}_1)\times\cdots\times L^2(\mathcal{I}_J)}{\text{argmax}} \sum_{j=1}^{J}\sum_{j'=1}^{J} c_{jj'}\, g(\text{cov}(\langle X_j, f_j\rangle, \langle X_{j'}, f_{j'}\rangle))$$

$$\text{s.t.} \quad \langle f_j, \mathbf{M}_j f_j\rangle = 1, \quad \forall j \in \{1, \dots, J\}$$

with usually $\mathbf{M}_j = \tau_j \mathbf{I}_{\mathcal{I}_j} + (1-\tau_j)\mathbf{\Sigma}_{jj}$ with $\tau_j \in]0,1]$.

Functional Generalized Canonical Correlation Analysis

Definition

Considering J random processes $X_1, \ldots, X_J$ defined on $\mathcal{I}_1, \ldots, \mathcal{I}_J$ respectively. Introducing FGCCA optimization problem as:

$$\underset{f_1,\ldots,f_J \in L^2(\mathcal{I}_1)\times\cdots\times L^2(\mathcal{I}_J)}{\operatorname{argmax}} \sum_{j=1}^{J}\sum_{j'=1}^{J} c_{jj'}\, g(\operatorname{cov}(\langle X_j, f_j\rangle, \langle X_{j'}, f_{j'}\rangle))$$
$$\text{s.t.} \quad \langle f_j, \mathbf{M}_j f_j\rangle = 1, \quad \forall j \in \{1, \ldots, J\}$$

with usually $\mathbf{M}_j = \tau_j \mathbf{I}_{\mathcal{I}_j} + (1-\tau_j)\mathbf{\Sigma}_{jj}$ with $\tau_j \in]0,1]$.

Defining the cross-covariance surface $\Sigma_{jj'}(s,t) = \operatorname{cov}(X_j(s), X_{j'}(t))$ between processes j and j' and the associated operator:

$$\mathbf{\Sigma}_{jj'} : L^2(\mathcal{I}_j) \to L^2(\mathcal{I}_{j'})\ ,\ f \mapsto g : g(t) = \int_{\mathcal{I}_j} \Sigma_{jj'}(s,t) f(s) \mathrm{d}s.$$

Functional Generalized Canonical Correlation Analysis

Definition

Considering J random processes $X_1, \dots, X_J$ defined on $\mathcal{I}_1, \dots, \mathcal{I}_J$ respectively. Introducing FGCCA optimization problem as:

$$\underset{f_1,\dots,f_J \in L^2(\mathcal{I}_1)\times\dots\times L^2(\mathcal{I}_J)}{\operatorname{argmax}} \sum_{j=1}^{J}\sum_{j'=1}^{J} c_{jj'}\, g(\langle f_j, \boldsymbol{\Sigma}_{jj'} f_{j'}\rangle))$$

$$\text{s.t.} \quad \langle f_j, \mathbf{M}_j f_j \rangle = 1, \quad \forall j \in \{1, \dots, J\}$$

with usually $\mathbf{M}_j = \tau_j \mathbf{I}_{\mathcal{I}_j} + (1-\tau_j)\boldsymbol{\Sigma}_{jj}$ with $\tau_j \in]0,1]$.

Defining the cross-covariance surface $\Sigma_{jj'}(s,t) = \mathrm{cov}(X_j(s), X_{j'}(t))$ between processes j and j' and the associated operator:

$$\boldsymbol{\Sigma}_{jj'} : L^2(\mathcal{I}_j) \to L^2(\mathcal{I}_{j'})\ ,\ f \mapsto g : g(t) = \int_{\mathcal{I}_j} \Sigma_{jj'}(s,t) f(s) \mathrm{d}s.$$

$\rightarrow$ Very convenient! $\Sigma_{jj'}$ can be estimated for sparse and irregular data.

Solving procedure

- For each f_j individually, when fixing f_j for $j \neq j'$, maximization of criterion Ψ is easily achieved using the convexity.
- Block Relaxation [de Leeuw, 1994] procedure:

Result: Estimation of $\mathbf{f} = (f_1, \ldots, f_J)$
Initialize $f_1^{(0)}, \ldots, f_J^{(0)}$ randomly
repeat
  for $j = 1$ **to** J **do**

$$f_j^{(s+1)} = \underset{\substack{f \in L^2(\mathcal{I}_j) \\ \langle f, \mathbf{M}_j f \rangle = 1}}{\operatorname{argmax}} \Psi(f_1^{(s+1)}, \ldots, f_{j-1}^{(s+1)}, f, f_{j+1}^{(s)}, \ldots, f_J^{(s)})$$

  end
until $|\Psi(\mathbf{f}^{(s+1)}) - \Psi(\mathbf{f}^{(s)})| < \varepsilon$;

Retrieving higher-order functions

- Multivariate setting: projection/deflation of sets of variables $\mathbf{x}_j$:

$$\mathbf{x}_j' = \mathbf{x}_j - (\mathbf{a}_j^\top \mathbf{a}_j)^{-1} \mathbf{a}_j \mathbf{a}_j^\top \mathbf{x}_j.$$

Retrieving higher-order functions

- Multivariate setting: projection/deflation of sets of variables $\mathbf{x}_j$:

$$\mathbf{x}_j' = \mathbf{x}_j - (\mathbf{a}_j^\top \mathbf{a}_j)^{-1} \mathbf{a}_j \mathbf{a}_j^\top \mathbf{x}_j.$$

- Functional setting: limited access to X_j, deflating cross-covariance:

Retrieving higher-order functions

- Multivariate setting: projection/deflation of sets of variables $\mathbf{x}_j$:

$$\mathbf{x}_j' = \mathbf{x}_j - (\mathbf{a}_j^\top \mathbf{a}_j)^{-1} \mathbf{a}_j \mathbf{a}_j^\top \mathbf{x}_j.$$

- Functional setting: limited access to X_j, deflating cross-covariance:

Proposition

Defining $\boldsymbol{\Phi_j} : L^2(\mathcal{I}_j) \to L^2(\mathcal{I}_j)\ ,\ f \mapsto \langle f_j, f \rangle f_j$

$$\boldsymbol{\Sigma}_{jj'}' = (\mathbf{I}_{\mathcal{I}_j} - \boldsymbol{\Phi}_j)\boldsymbol{\Sigma}_{jj'}(\mathbf{I}_{\mathcal{I}_{j'}} - \boldsymbol{\Phi}_{j'}),$$

Retrieving higher-order functions

- Multivariate setting: projection/deflation of sets of variables $\mathbf{x}_j$:

$$\mathbf{x}_j' = \mathbf{x}_j - (\mathbf{a}_j^\top \mathbf{a}_j)^{-1} \mathbf{a}_j \mathbf{a}_j^\top \mathbf{x}_j.$$

- Functional setting: limited access to X_j, deflating cross-covariance:

Proposition

Defining $\boldsymbol{\Phi_j} : L^2(\mathcal{I}_j) \to L^2(\mathcal{I}_j)\ ,\ f \mapsto \langle f_j, f \rangle f_j$

$$\boldsymbol{\Sigma}_{jj'}' = (\mathbf{I}_{\mathcal{I}_j} - \boldsymbol{\Phi}_j)\boldsymbol{\Sigma}_{jj'}(\mathbf{I}_{\mathcal{I}_{j'}} - \boldsymbol{\Phi}_{j'}),$$

$\rightarrow$ We can derive a similar result for uncorrelated components deflation.

Component estimation

- Multivariate setting:
  - Denoting $\mathbf{x}_{ij}$ the ith sample of set $\mathbf{x}_j$
  - Component of ith sample of $\mathbf{x}_j$ is computed by $u_{ij} = \mathbf{a}_j^\top \mathbf{x}_{ij}$

Component estimation

- Multivariate setting:
  - Denoting $\mathbf{x}_{ij}$ the ith sample of set $\mathbf{x}_j$
  - Component of ith sample of $\mathbf{x}_j$ is computed by $u_{ij} = \mathbf{a}_j^\top \mathbf{x}_{ij}$
- Functional setting:
  - Denoting X_{ij} the ith sample of process X_j
  - Limited view of X_{ij}: component $u_{ij} = \int_{\mathcal{I}_j} f_j(t) X_{ij}(t) \mathrm{d}t$ approximated

Component estimation

- Multivariate setting:
  - Denoting $\mathbf{x}_{ij}$ the ith sample of set $\mathbf{x}_j$
  - Component of ith sample of $\mathbf{x}_j$ is computed by $u_{ij} = \mathbf{a}_j^\top \mathbf{x}_{ij}$
- Functional setting:
  - Denoting X_{ij} the ith sample of process X_j
  - Limited view of X_{ij}: component $u_{ij} = \int_{\mathcal{I}_j} f_j(t) X_{ij}(t) \mathrm{d}t$ approximated

Proposition

Assuming $\mathbf{u} = (u_1, \ldots, u_J)$ and model errors are jointly Gaussian, the empirical Bayes estimator for $\mathbf{u}_i$ is

$$\tilde{\mathbf{u}}_i = \mathbb{E}(\mathbf{u}_i | \mathbf{y}_i)$$

which has a closed-form expression depending on $\boldsymbol{\Sigma}_{jj'}$ and $f_1, \ldots, f_J$.

Simulations

- Generating data for $J = 2$ random processes with shared statistical structure using $M = 6$ basis functions on a grid of size $K = 30$ of $\mathcal{I}_1 = \mathcal{I}_2 = [0, 1]$. Considering
  - No Sparsity/Dense ($\text{NA}_{\%} = 0.0$)
  - Low Sparsity ($\text{NA}_{\%} = 0.2$)
  - Medium Sparsity ($\text{NA}_{\%} = 0.5$)
  - High Sparsity ($\text{NA}_{\%} = 0.8$)
- Comparing:
  - **Functional Generalized Canonical Correlation Analysis (FGCCA)**
  - Functional Principal Component Analysis (FPCA) [Yao et al., 2005]
  - Functional Singular Value Decomposition (FSVD) [Yang et al., 2011]

Performances

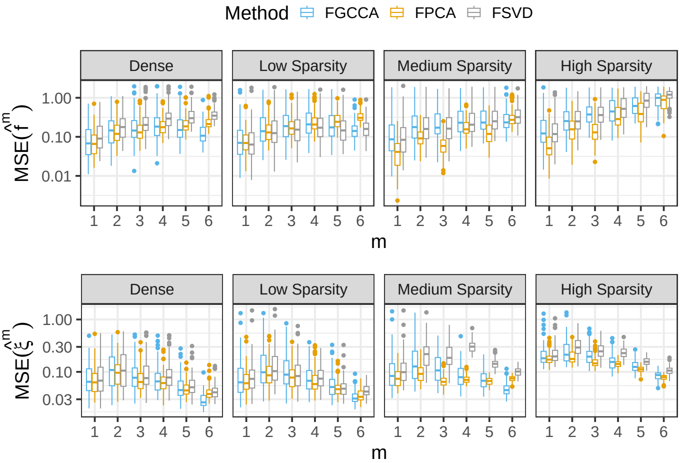

Figure: Performances with $\mathrm{MSE}(\hat{\mathbf{f}}^m) = \frac{1}{J}\sum_{j=1}^{J} \|f_j^m - \hat{f}_j^m\|^2 \mathrm{d}t$ and $\mathrm{MSE}(\hat{\xi}^m) = \frac{1}{IJ}\sum_{j=1}^{J}\sum_{i=1}^{I}(\xi_{ij}^m - \hat{\xi}_{ij}^m)^2$

Reminder: primary biliary cholangitis

Objectives

- Investigating relationships between markers.
- Integrating associations to improve characterization.

Details

Data of $J = 3$ markers on $I = 282$ subjects with PBC: 140 dead and 142 alive at the end. Considering the first 10 years.

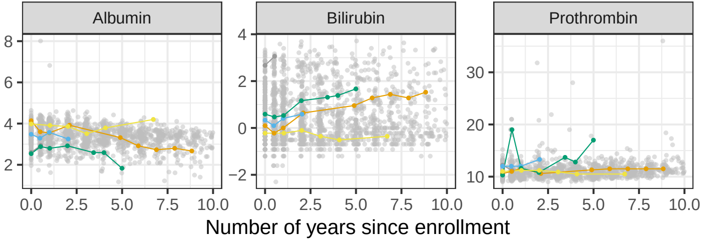

Figure: Observations of blood markers with first trajectories colored.

Results: primary biliary cholangitis

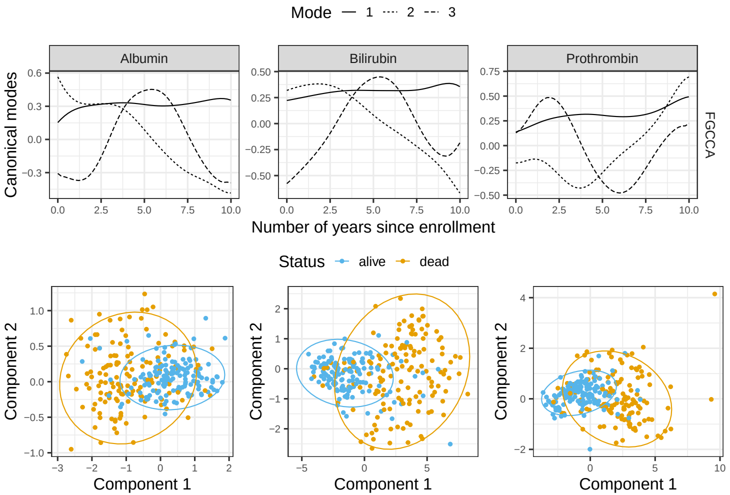

Figure: Canonical functions (top) Canonical component (bottom)

Outline

Introduction

Functional Generalized Canonical Correlation Analysis

Perspectives and limitations

References and appendix

Extensions

- Integrating multivariate response in FGCCA:

$$\underset{\substack{f_1,\dots,f_J,\in\Omega_1\times\cdots\times\Omega_J \\ \|\mathbf{a}\|_2=1}}{\operatorname{argmax}} \sum_{j\neq j'} c_{jj'}\, g(\langle f_j, \mathbf{\Sigma}_{jj'} f_{j'}\rangle) + 2\sum_j g(\langle f_j, \mathbf{\Sigma}_{j\mathbf{y}}\mathbf{a}\rangle). \qquad (1)$$

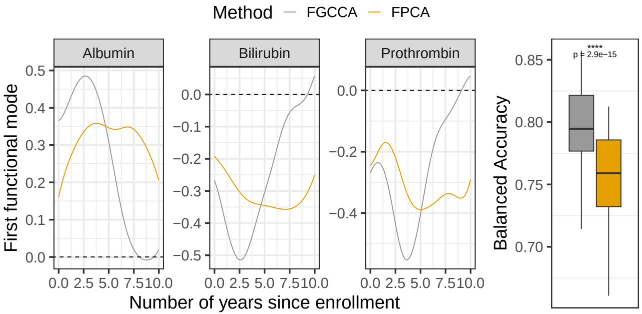

Figure: Comparing FPCA and FGCCA with status integration status. (left) principal and canonical function; (right) balanced accuracy, p-value, and significance level of difference derived from a t-test.

Extensions

- Integrating survival information and joint modeling with FGCCA:

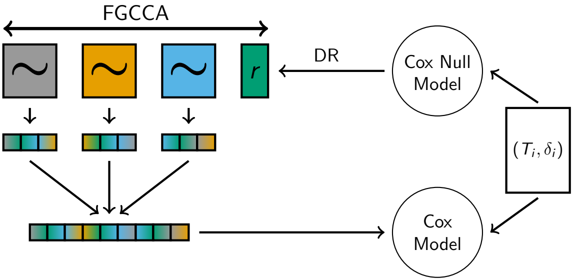

Figure: Joint modeling using FGCCA.

Joint modeling

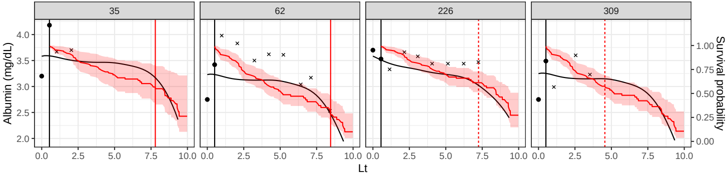

Figure: Dynamic prediction. Observations (points), reconstructed trajectories (black solid lines), survival functions (red solid lines) for 4 subjects. Censoring (dashed red vertical lines) or death (solid red vertical lines) times are indicated along with observations not considered (crosses).

Joint modeling

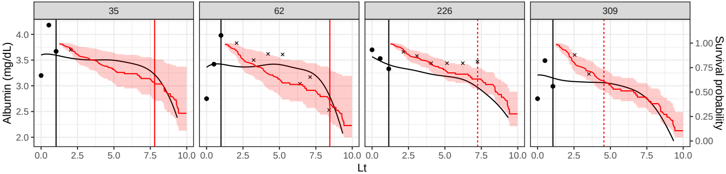

Figure: Dynamic prediction. Observations (points), reconstructed trajectories (black solid lines), survival functions (red solid lines) for 4 subjects. Censoring (dashed red vertical lines) or death (solid red vertical lines) times are indicated along with observations not considered (crosses).

Joint modeling

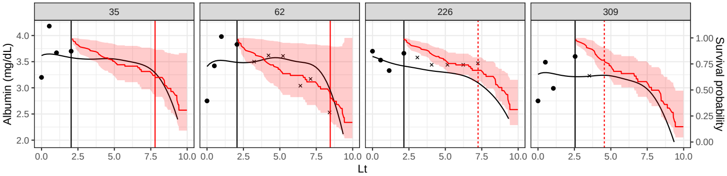

Figure: Dynamic prediction. Observations (points), reconstructed trajectories (black solid lines), survival functions (red solid lines) for 4 subjects. Censoring (dashed red vertical lines) or death (solid red vertical lines) times are indicated along with observations not considered (crosses).

Numerous blocks

Example

Alzheimer Disease Neuroimaging Initiative (ADNI) study. Six neurocognitive markers observed over several years

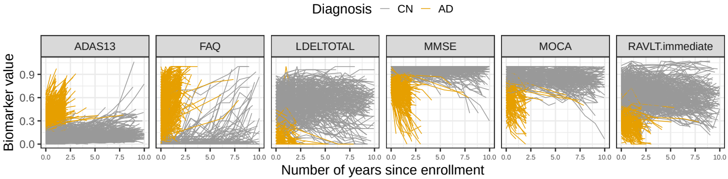

Figure: Trajectories of six neurocognitive markers in the ADNI study colored by baseline diagnosis: Cognitively Normal (CN), Alzheimer's Disease (AD)

Numerous blocks

Example

Alzheimer Disease Neuroimaging Initiative (ADNI) study. Six neurocognitive markers observed over several years

Limitations

- No separable characterization of subject, time, and marker
- Do not integrate potential higher-dimensional structure

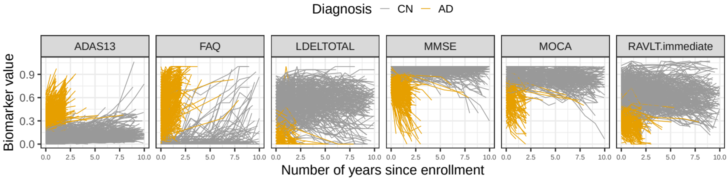

Figure: Trajectories of six neurocognitive markers in the ADNI study colored by baseline diagnosis: Cognitively Normal (CN), Alzheimer's Disease (AD)

Thank you! Questions?

- **Functional Generalized Canonical Correlation Analysis for studying multiple longitudinal variables**, Lucas SORT, Laurent LE BRUSQUET, Arthur TENENHAUS *Biometrics*, 2024, vol. 80, no 4.
- **Development of new statistical approaches for the investigation of multiblock and tensor longitudinal data**, Lucas SORT, Thèse de doctorat, Université Paris-Saclay, 2025.

Outline

Introduction

Functional Generalized Canonical Correlation Analysis

Perspectives and limitations

References and appendix

References I

de Leeuw, J. (1994).
Block-relaxation Algorithms in Statistics, page 308–324.
Springer Berlin Heidelberg.

Fan, J. and Gijbels, I. (2018).
Local Polynomial Modelling and Its Applications.
Routledge.

Hotelling, H. (1936).
Relations Between Two Sets of Variates.
Biometrika, 28(3/4):321.

Laird, N. M. and Ware, J. H. (1982).
Random-effects models for longitudinal data.
Biometrics, 38(4):963.

Ramsay, J. O. and Silverman, B. W. (2005).
Functional Data Analysis.
Springer New York.

References II

Tenenhaus, M., Tenenhaus, A., and Groenen, P. J. F. (2017).
Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods.
Psychometrika, 82(3):737–777.

Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. (2017).
Fast covariance estimation for sparse functional data.
Statistics and Computing, 28(3):511–522.

Xiao, L., Zipunnikov, V., Ruppert, D., and Crainiceanu, C. (2014).
Fast covariance estimation for high-dimensional functional data.
Statistics and Computing, 26(1–2):409–421.

Yang, W., Müller, H.-G., and Stadtmüller, U. (2011).
Functional singular component analysis.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 73(3):303–324.

References III

Yao, F., Müller, H.-G., and Wang, J.-L. (2005).
Functional data analysis for sparse longitudinal data.
Journal of the American Statistical Association, 100(470):577–590.

Local linear smoothing details

- **Mean estimation**: Aggregating all observations, for each estimation point $s \in \mathcal{I}$, we consider

$$\underset{\beta_0(s),\beta_1(s)}{\operatorname{argmin}} \sum_{i=1}^{I} \sum_{k=1}^{n_i} K_1\left(\frac{t_{ik}-s}{h}\right)(y_{ik}-\beta_0(s)-\beta_1(s)(s-t_{ik}))^2,$$

  and use $\hat{\mu}(s) = \beta_0(s)$.

- **Covariance estimation**: Aggregating "raw" covariance $C_{XX}(t_{ik}, t_{il}) = (y_{ik} - \hat{\mu}(t_{ik}))(y_{il} - \hat{\mu}(t_{il}))$ and, similar to before, at any covariance estimation point (s, t)

$$\underset{\beta_0(s,t),\beta_1(s,t),\beta_2(s,t)}{\operatorname{argmin}} \sum_{i=1}^{I} \sum_{k=1}^{n_i} \sum_{l=1;l\neq k}^{n_i} K_2\left(\frac{t_{ik}-s}{h}, \frac{t_{il}-t}{h}\right) \times (C_{XX}(t_{ik}, t_{il}) - \beta_0(s,t) - \beta_1(s,t)(s-t_{ik}) - \beta_2(s,t)(t-t_{il}))^2,$$

  and use $\hat{\Sigma}(s,t) = \beta_0(s,t)$.

ADNI application

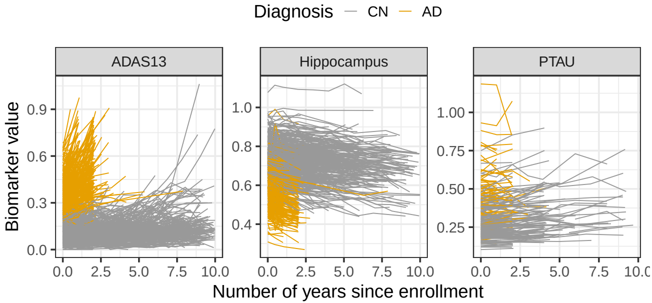

Figure: FGCCA applied on 3 longitudinal markers observed in the ADNI study: ADAS score, hippocampus volume, and protein tau concentration.

ADNI application

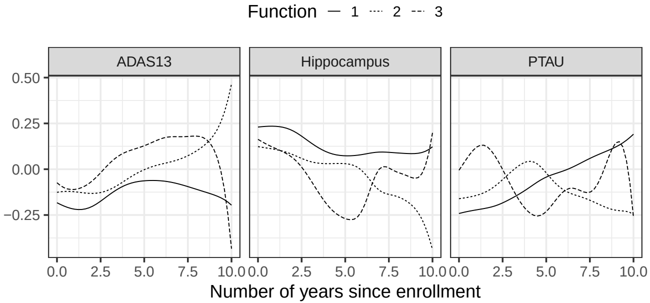

Figure: FGCCA applied on 3 longitudinal markers observed in the ADNI study: ADAS score, hippocampus volume, and protein tau concentration.

ADNI application

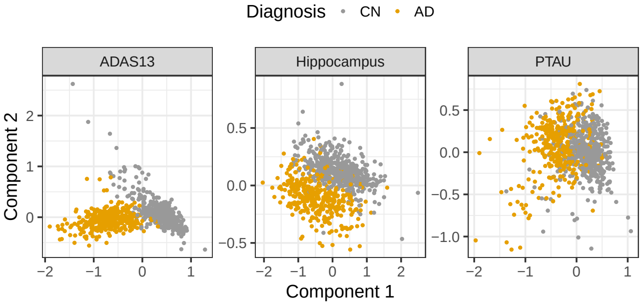

Figure: FGCCA applied on 3 longitudinal markers observed in the ADNI study: ADAS score, hippocampus volume, and protein tau concentration.