Categorical functional data analysis. The cfda R package

Cristian Preda, Quentin Grimonprez, Vincent Vandewalle

Categorical functional data analysis. The *cfda* R package

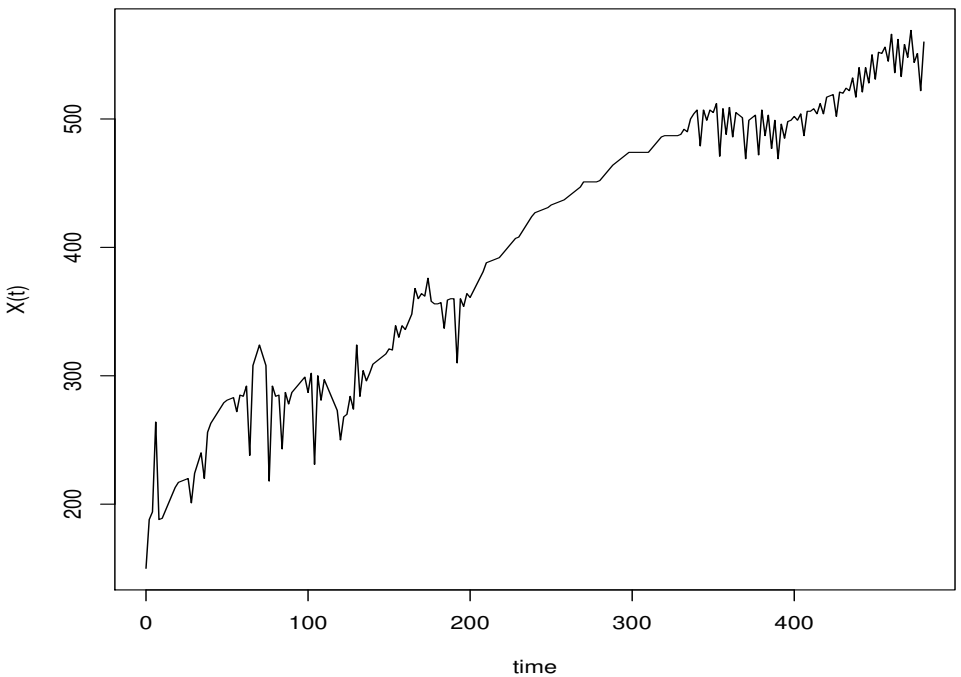
Cristian Preda[1], Quentin Grimonprez[2], Vincent Vandewalle[1]

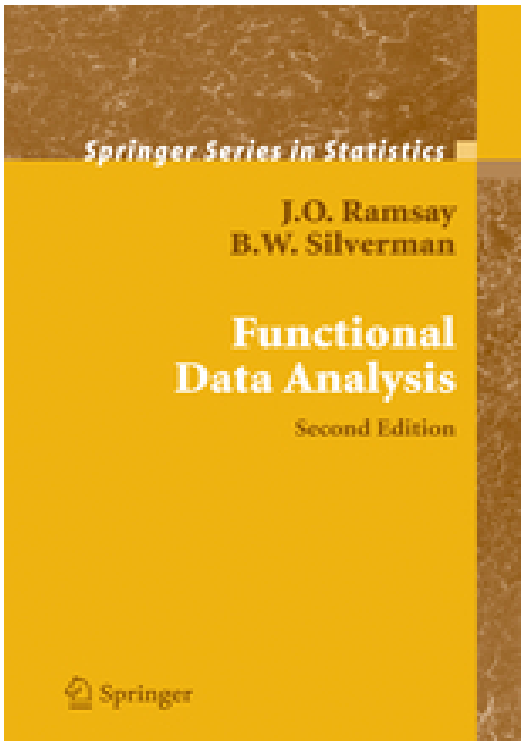
[1]Université de Lille et MODAL/Inria Lille Nord-Europe
[2] DIAGRAMS Technologies

ERCIM 2021
Computational and Methodological Statistics
18th December, 2021

(Categorical) Functional Data

Most of considered functional data is **scalar** (univariate/ multivariate) :





Model : Stochastic process, $X = \{X_t, \quad t \in \mathcal{T}\}$,

$$X_t \in \mathbb{R}^p, p \geq 1,$$

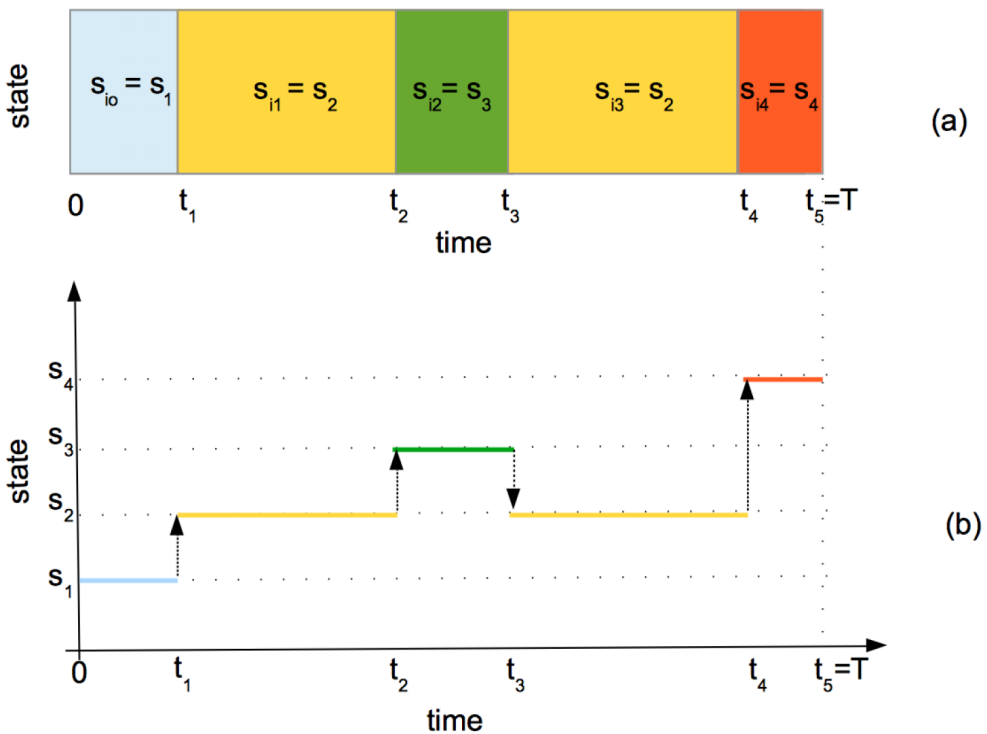
for some continuous index set $\mathcal{T}$

Categorical functional data

Set of states : $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$, $K \geq 2$

$$X_t : \Omega \to \mathcal{S}.$$

A path of X on $[0, T]$ is a sequence of states $\mathbf{s_{i_j}}$ and times points $\mathbf{t_j}$:
$\{(0 = t_0, s_{i_0}), (t_1, s_{i_1}), (t_2, s_{i_2}), \ldots, (t_p = T, s_{i_p})\}$,



Categorical functional data

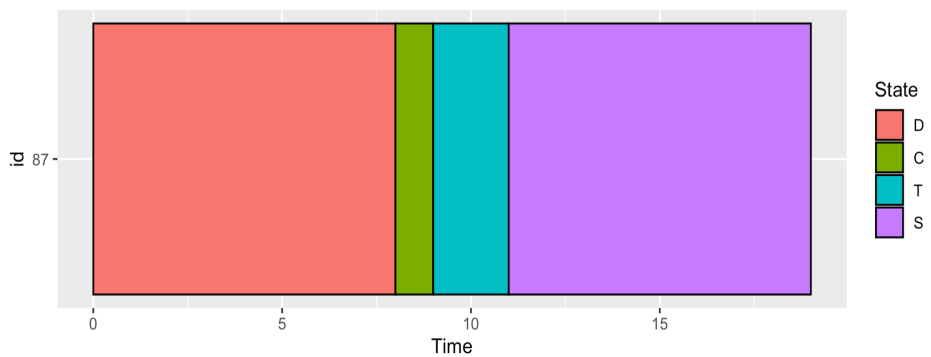
An example : Paths of infected patients.

$$\mathcal{S} = \{D, C, T, S\},$$

- D = patient has no follow-up
- C = patient has a follow-up but no treatment
- T = the patient has a medical follow-up with a treatment but the infection is not suppressed
- S = the patient has a medical follow-up with a treatment and the infection is suppressed.

A path on [0, 19] :

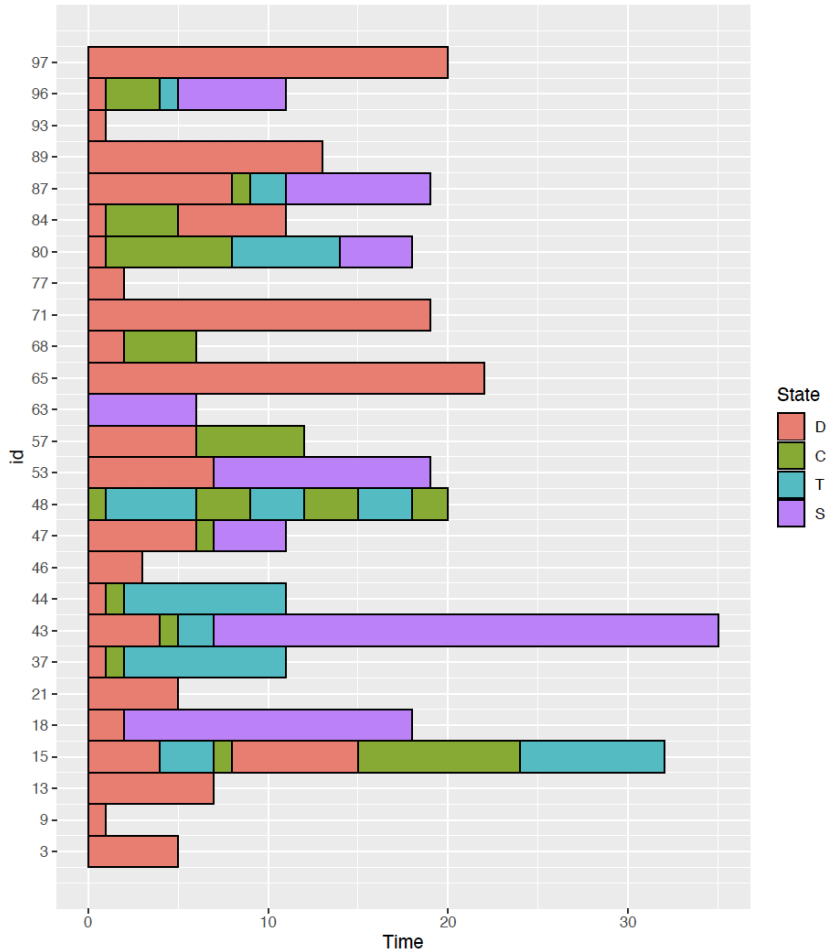
time	state
0	D
8	C
9	T
11	S
19	S



Categorical functional data

An example : Paths of infected patients.

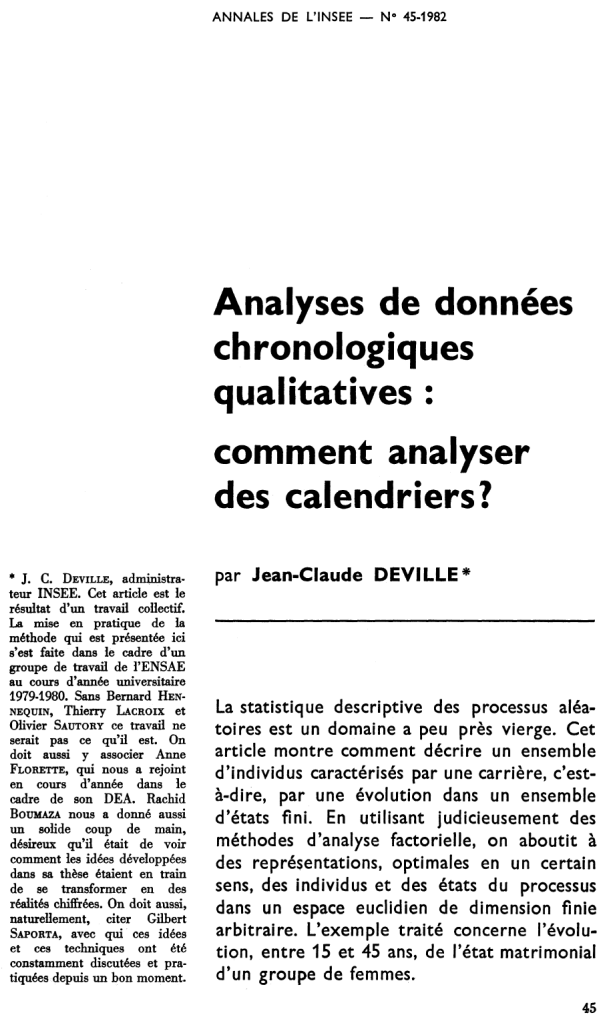
Several paths :



Categorical functional data

The seminal work of Jean-Claude Deville (1982)

Analyse de données chronologiques qualitatives : comment analyser des calendriers ?, Annales de l'INSEE, No. 45, p. 45-104.



ANNALES DE L'INSEE — N° 45-1982

Analyses de données chronologiques qualitatives :

comment analyser des calendriers?

par **Jean-Claude DEVILLE** *

* J. C. DEVILLE, administrateur INSEE. Cet article est le résultat d'un travail collectif. La mise en pratique de la méthode qui est présentée ici s'est faite dans le cadre d'un groupe de travail de l'ENSAE au cours d'année universitaire 1979-1980. Sans Bernard HENNEQUIN, Thierry LACROIX et Olivier SAUTORY ce travail ne serait pas ce qu'il est. On doit aussi y associer Anne FLORETTE, qui nous a rejoint en cours d'année dans le cadre de son DEA. Rachid BOUMAZA nous a donné aussi un solide coup de main, désireux qu'il était de voir comment les idées développées dans sa thèse étaient en train de se transformer en des réalités chiffrées. On doit aussi, naturellement, citer Gilbert SAPORTA, avec qui ces idées et ces techniques ont été constamment discutées et pratiquées depuis un bon moment.

La statistique descriptive des processus aléatoires est un domaine a peu près vierge. Cet article montre comment décrire un ensemble d'individus caractérisés par une carrière, c'est-à-dire, par une évolution dans un ensemble d'états fini. En utilisant judicieusement des méthodes d'analyse factorielle, on aboutit à des représentations, optimales en un certain sens, des individus et des états du processus dans un espace euclidien de dimension finie arbitraire. L'exemple traité concerne l'évolution, entre 15 et 45 ans, de l'état matrimonial d'un groupe de femmes.

45

Categorical functional data

Other pioneer works

- Boumaza Rachid (1980). Contribution à l'Étude Descriptive d'une Fonction Aléatoire Qualitative. Ph.D. thesis, Université Paul Sabatier, Toulouse.

- Deville Jean-Claude, Saporta Gilbert (1983). Correspondence Analysis with an Extension towards Nominal Time Series. Journal of Econometrics, 22, pages 169–189.

Categorical functional data analysis

Dimension reduction by optimal encoding and principal components

Idea : find $z \in L_2(\Omega)$ that maximizes

$$\int_0^T \eta^2(z; X_t)dt,$$

with $\eta^2(z; X_t) = \frac{var(\mathbb{E}_t(z))}{var(z)}$ and $\mathbb{E}_t(z) = \mathbb{E}(z|X_t)$.

Solution :

$$\int_0^T \mathbb{E}_t(z)dt = \lambda z.$$

$\{(\lambda_i, z_i)\}_{i\geq 1}$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and

$$\int_0^T \mathbb{E}_t(z_i)dt = \lambda_i z_i.$$

Categorical functional data analysis

Dimension reduction by optimal encoding and principal components

Optimal encoding functions : a^x, $x \in \mathcal{S}$:

$$z = \int_0^T \sum_{x \in \mathcal{S}} a^x(t) \mathbf{1}_t^x dt,$$

with

$$\int_0^T \sum_{y \in \mathcal{S}} p^{x,y}(t,s) a^y(s) ds = \lambda a^x(t) p^x(t), \quad \forall t \in [0,T], \forall x \in \mathcal{S},$$

where $p^x(t) = \mathbb{P}(X_t = x)$ and $p^{x,y}(t,s) = \mathbb{P}(X_t = x, X_s = y)$.
under the constraint :

$$\int_0^T \sum_{x \in \mathcal{S}} [a^x(t)]^2 p^x(t) dt = 1.$$

Categorical functional data analysis

Two expansion formulas

$$\mathbf{1}_t^x = \sum_{i \geq 1} z_i a_i^x(t) \frac{1}{p^x(t)}, \quad \forall x \in \mathcal{S}.$$

and

$$p^{x,y}(t,s) = p^x(t)p^y(s) \sum_{i \geq 1} \lambda_i a_i^x(t) a_i^y(s),$$

$\forall t, s \in [0, T], \forall x, y \in \mathcal{S}$.

In particular, for $x = y$ and $s = t$ we obtain

$$p^x(t) = \left\{ \sum_{i \geq 1} \lambda_i \left[a_i^x(t) \right]^2 \right\}^{-1}, \ \forall t \in [0, T], \forall x \in \mathcal{S}.$$

A two-states model example (Saporta (1981))

Let $[0, T] = [0, 1]$ and $\theta \sim U[0, 1]$.

$$X_t(\omega) = \begin{cases} 0, & \textit{if } t < \theta(\omega), \\ 1, & \textit{otherwise}. \end{cases}$$

$$P(t, s) = \begin{pmatrix} 1 - s & s - t \\ 0 & t \end{pmatrix} \text{ for } s < t,$$

$$P(t, s) = \begin{pmatrix} 1 - t & 0 \\ t - s & s \end{pmatrix} \text{ for } t \geq s.$$

A two-states model example (Saporta (1981) –cont. 1

$$\lambda \begin{pmatrix} a_0(t) \\ a_1(t) \end{pmatrix} = \int_0^t \begin{pmatrix} 1 & 0 \\ 1-\frac{s}{t} & \frac{s}{t} \end{pmatrix} \begin{pmatrix} a_0(s) \\ a_1(s) \end{pmatrix} ds$$

$$+ \int_t^1 \begin{pmatrix} \frac{1-s}{1-t} & \frac{s-t}{1-t} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a_0(s) \\ a_1(s) \end{pmatrix} ds$$

$\lambda_i = \frac{1}{i(i+1)}$, $i \geq 1$.

$$\lambda_1 = \tfrac{1}{2}: \quad \begin{cases} a_0(t) = \sqrt{6}t, \\ a_1(t) = \sqrt{6}(t-1). \end{cases}$$

$$\lambda_2 = \tfrac{1}{6}: \quad \begin{cases} a_0(t) = \sqrt{\frac{120}{11}}(t^2 - 0.5t), \\ a_1(t) = \sqrt{\frac{120}{11}}(t^2 - 1.5t + 0.5). \end{cases}$$

$$\frac{\lambda_1 + \lambda_2}{(m-1)T} = 67\%.$$

A two-states model example (Saporta (1981)) – cont. 2

The principal components z_i are, up to a constant,

$$z_i = P_i(2\theta - 1),$$

where P_i is the i-th Legendre polynomial,

$$P_i(x) = \frac{\partial (x^2 - 1)^i}{\partial^i x}$$

Thus,

$$z_1 = \frac{\sqrt{6}}{2}(2\theta - 1),$$

$$z_2 = \sqrt{\frac{5}{24}}\left[3(2\theta - 1)^2 - 1\right]$$

Categorical functional data analysis

Approximation

Let $\{\phi_1, \ldots, \phi_m\}$, $\phi_i : [0, T] \to \mathbb{R}$, $i = 1, \ldots, m$, be a basis of functions (Fourier, B-splines, monomial, etc.) and for each $x \in \mathcal{S}$ consider the approximation :

$$a^x(t) \approx \alpha_{(x,1)}\phi_1(t) + \alpha_{(x,2)}\phi_2(t) + \ldots + \alpha_{(x,m)}\phi_m(t), \quad \forall t \in [0, T],$$

where $\alpha_x = \left(\alpha_{(x,1)}, \alpha_{(x,2)}, \ldots, \alpha_{(x,m)}\right)' \in \mathbb{R}^m$.

Let $\alpha \in \mathbb{R}^{m \times K}$ be the column vector obtained by the concatenation of the vectors $\{\alpha_x\}_{x \in \mathcal{S}}$

Categorical functional data analysis

Approximation

Then,

$$G\alpha = \lambda F\alpha,$$

under the constraint

$$\alpha' F\alpha = 1,$$

where the matrix G is the covariance matrix of the random variables $\{V_{(x,i)}, x \in \mathcal{S}, i \in 1, \ldots m\}$, defined as

$$V_{(x,i)} = \int_0^T \phi_i(t)\mathbf{1}_t^x dt, \quad \forall x \in \mathcal{S},$$

$$G = \left\{ G_{(x,i),(y,j)} = cov\left(V_{(x,i)}, V_{(y,j)}\right), \quad x, y \in \mathcal{S}, i, j = 1, \ldots, m \right\},$$

Categorical functional data analysis

Approximation

The matrix F is defined by

$$F = \left\{ F_{(x,i),(y,j)} = \mathbb{E}\left(U_{(x,i),(y,j)} \right), \quad x, y \in \mathcal{S}, i, j = 1, \ldots, m \right\}, \quad (1)$$

where $U_{(x,i),(y,j)}$ is the random variable

$$U_{(x,i),(y,j)} = \int_0^T \phi_i(t)\phi_j(t)\mathbf{1}_t^x \mathbf{1}_t^y dt = \begin{cases} \displaystyle\int_0^T \phi_i(t)\phi_j(t)\mathbf{1}_t^x dt & \text{if } x = y, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, F is a block diagonal matrix, each block being a square matrix of size $m \times m$ corresponding to each x in $\mathcal{S}$, $\{U_{(x,i),(x,j)}, i, j = 1, \ldots, m\}$.

Categorical functional data analysis

Estimation

Let $\{X_1, \ldots, X_n\}$ be a sample of n paths of X corresponding to a random sample $(\omega_1, \ldots, \omega_n) \in \Omega^n$. Then,

- the V data set with n rows and Km columns for the V's random variables,

$$\mathsf{V} = \begin{array}{c|cccccccc} \omega & V_{(s_{\mathbf{1}},\mathbf{1})} & \cdots & V_{(s_{\mathbf{1}},m)} & \cdots & V_{(x,i)} & \cdots & V_{(s_K,m)} \\ \hline \omega_{\mathbf{1}} & V_{(s_{\mathbf{1}},\mathbf{1})}(\omega_{\mathbf{1}}) & \cdots & V_{(s_{\mathbf{1}},m)}(\omega_{\mathbf{1}}) & \cdots & V_{(x,i)}(\omega_{\mathbf{1}}) & \cdots & V_{(s_K,m)}(\omega_{\mathbf{1}}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \omega_n & V_{(s_{\mathbf{1}},\mathbf{1})}(\omega_n) & \cdots & V_{(s_{\mathbf{1}},m)}(\omega_n) & \cdots & V_{(x,i)}(\omega_n) & \cdots & V_{(s_K,m)}(\omega_n) \end{array}$$

- and the U dataset with n rows and Km^2 columns for the U's random variables, respectively :

$$\mathsf{U} = \begin{array}{c|cccccc} \omega & \cdots & U_{(x,i),(x,\mathbf{1})} & \cdots & U_{(x,i),(x,m)} & \cdots & U_{(s_K,m),(s_K,m)} \\ \hline \omega_{\mathbf{1}} & \cdots & U_{(x,i),(x,\mathbf{1})}(\omega_{\mathbf{1}}) & \cdots & U_{(x,i),(x,m)}(\omega_{\mathbf{1}}) & \cdots & U_{(s_K,m),(s_K,m)}(\omega_{\mathbf{1}}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \omega_n & \cdots & U_{(x,i),(x,\mathbf{1})}(\omega_n) & \cdots & U_{(x,i),(x,m)}(\omega_n) & \cdots & U_{(s_K,m),(s_K,m)}(\omega_n) \end{array}$$

Categorical functional data analysis

Estimation

For each i and j in $\{1, \ldots, m\}$ and x and y in $\mathcal{S}$ one has :

$$\widehat{G}_{(x,i),(y,j)} = \widehat{cov}\left(V_{(x,i)}, V_{(y,j)}\right) = \frac{1}{n-1}\left(\sum_{h=1}^{n} V_{(x,i)}(\omega_h) V_{(y,j)}(\omega_h) - n\bar{V}_{(x,i)}\bar{V}_{(y,j)}\right)$$

and

$$\widehat{F}_{(x,i),(y,j)} = \begin{cases} \bar{U}_{(x,i),(y,j)} = \dfrac{1}{n}\displaystyle\sum_{h=1}^{n} U_{(x,i),(y,j)}(\omega_h) & \text{if } x = y, \\ 0 & \text{otherwise.} \end{cases}$$

An estimate of i-th eigenvector $\alpha^{(i)}$, $i \geq 1$, is the i-th eigenvector $\widehat{\alpha}_i$ of the eigen-equation,

$$\widehat{G}\widehat{\alpha} = \widehat{\lambda}\widehat{F}\widehat{\alpha},$$

under the constraint

$$\widehat{\alpha}'\widehat{F}\widehat{\alpha} = 1.$$

Categorical functional data analysis

Estimation

Then, for each $x \in \mathcal{S}$, the i-th encoding eigen-function a_i^x is estimated by

$$\widehat{a}_i^x = \sum_{j=1}^{m} \widehat{\alpha}_{i,(x,j)} \phi_j, \quad i \geq 1.$$

The estimates for the principal components :

$$\widehat{z}_i(\omega) = \int_0^T \sum_{x \in \mathcal{S}} \widehat{a}_i^x(t) \mathbf{1}_t^x(\omega) dt = \sum_{x \in \mathcal{S}} \sum_{j=1}^{m} \widehat{\alpha}_{i,(x,j)} V_{(x,j)}(\omega), \quad i \geq 1.$$

Categorical functional data analysis

Estimation

Confidence bounds.
Bootstrapping from the V and U datasets,

$$\widehat{var(a_i^x(t))} = \phi(t)'\widehat{\Sigma}_{(i,x)}\phi(t),$$

where $\phi(t)$ is the column vector $\phi(t) = (\phi_1(t), \ldots, \phi_m(t))'$ and $\widehat{\Sigma}_{(i,x)}$ is the covariance matrix of $\widehat{\alpha}_{i,x} = \left(\widehat{\alpha}_{i,(x,1)}, \ldots, \widehat{\alpha}_{i,(x,m)}\right)$.

Then, for a confidence level $1-u$, $u \in [0,1]$, a confidence interval for $a_i^x(t)$ is obtained as

$$\mathtt{CI}^{1-u}(a_i^x(t)) = \widehat{a_i^x(t)} \pm \zeta_{1-\frac{u}{2}} \sqrt{\widehat{var(a_i^x(t))}},$$

where $\zeta_{1-\frac{u}{2}}$ is the quantile of order $1-\frac{u}{2}$ of the standard normal distribution.