Linear approximation for multivariate categorical functional data analysis

Cristian Preda [1,2] Quentin Grimonprez [3]

[1] Université de Lille, Inria MODAL
e-mail: cristian.preda@univ-lille.fr

[2]ISMMA Romanian Academy

[3]DIAGRAMS Technologies
e-mail: qgrimonprez@diagrams-technologies.c

The 24th Conference of Romanian Society of Probability and Statistics

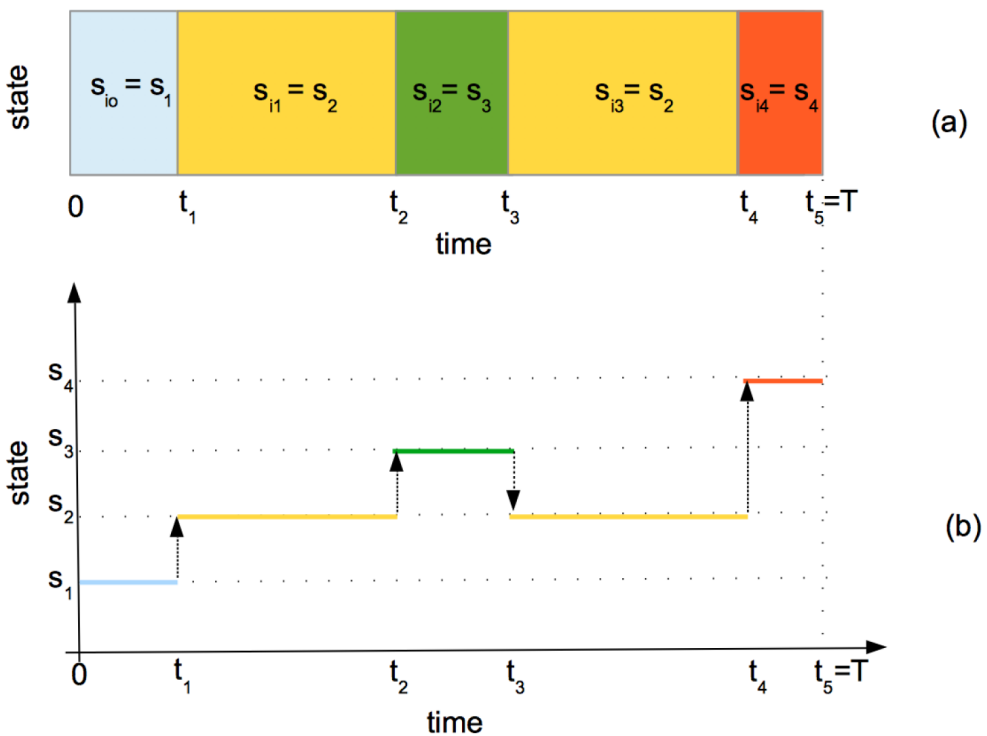
Bucharest, April 21-22 2023

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 s_{i_j} and times points 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

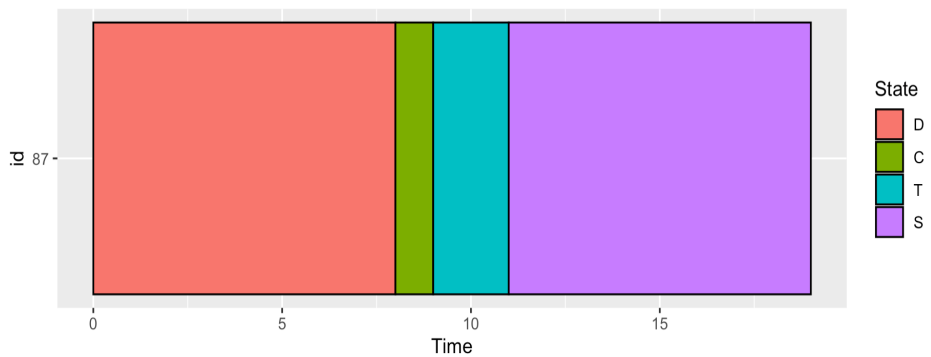
An example : Paths of infected patients.

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

- 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] :

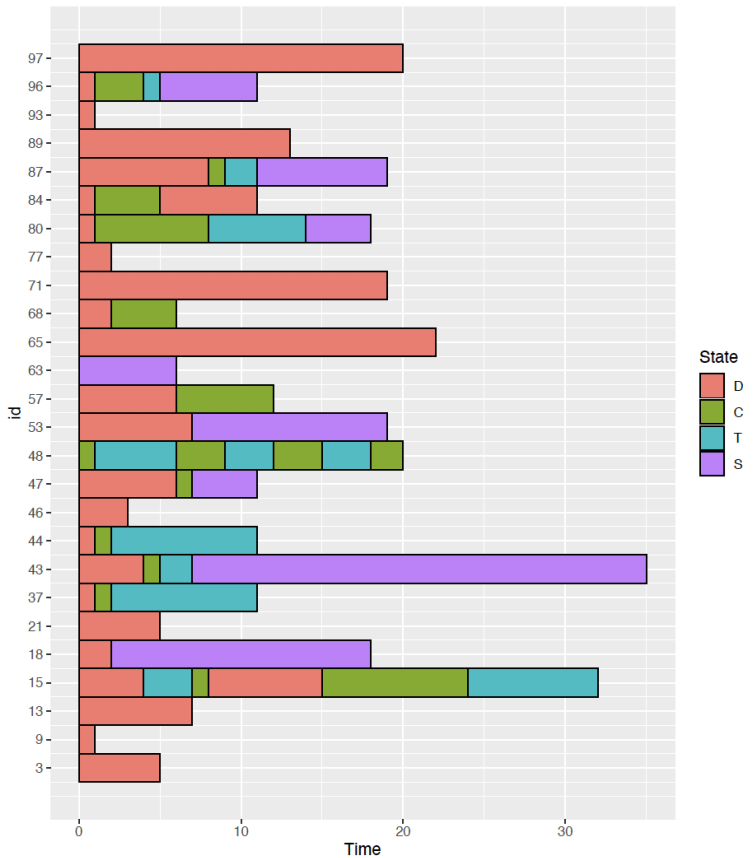
time	state
0	D
8	C
9	T
11	S
19	S



Categorical functional data

An example : Paths of infected patients.

Several paths :



Statistical analysis :dimension reduction for visualisation, clustering and regression (`cfda R package`).

Categorical functional data analysis

Dimension reduction by optimal encoding and principal components

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

$$\int_0^T \eta^2(z; X_t)dt, \quad \text{with} \quad \eta^2(z; X_t) = \frac{var(\mathbb{E}_t(z))}{var(z)} \text{ and } \mathbb{E}_t(z) = \mathbb{E}(z|X_t).$$

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

The *generating eigen-proces* :

Define

$$\xi_t = \mathbb{E}_t(z) = \mathbb{E}(z|X_t) = \sum_{x \in \mathcal{S}} a^x(t) 1_t^x, \quad t \in [0, T]$$

Then, the process $\{\xi_t\}_{t \in [0,T]}$ is such that it maximises the criterion

$$\max \int_0^T \int_0^T Cov(\xi_t \xi_s) dt ds,$$

under the **constraint of unit total variance**.

$\{\xi_t\}_{t \in [0,T]}$ is solution to the following eigen-value problem :

$$\int_0^T \mathbb{E}_t \mathbb{E}_s \xi_s ds = \lambda \xi_t, \quad t \in [0, T].$$

Categorical functional data analysis

Dimension reduction by optimal encoding and principal components

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

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

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

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

and

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

Categorical functional data analysis

Approximation of optimal encoding functions

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

$$\alpha' F\alpha = 1.$$

Categorical functional data analysis

Approximation

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

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

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

- F is defined by

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

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

$$U_{(x,i),(y,j)} = \int_0^T \phi_i(t)\phi_j(t) 1_t^x 1_t^y dt$$

Remark : The matrices F and G are of size $Km \times Km$.

Multivariate categorical functional data

Let $X = \{X_t, t \in [0, T]\}$ be a random process such that

$$X_t = (X_t^1, \ldots, X_t^p),$$

where $X^1, \ldots, X^p$ are p categorical functional random variables such that, for $j = 1, \ldots, p$,

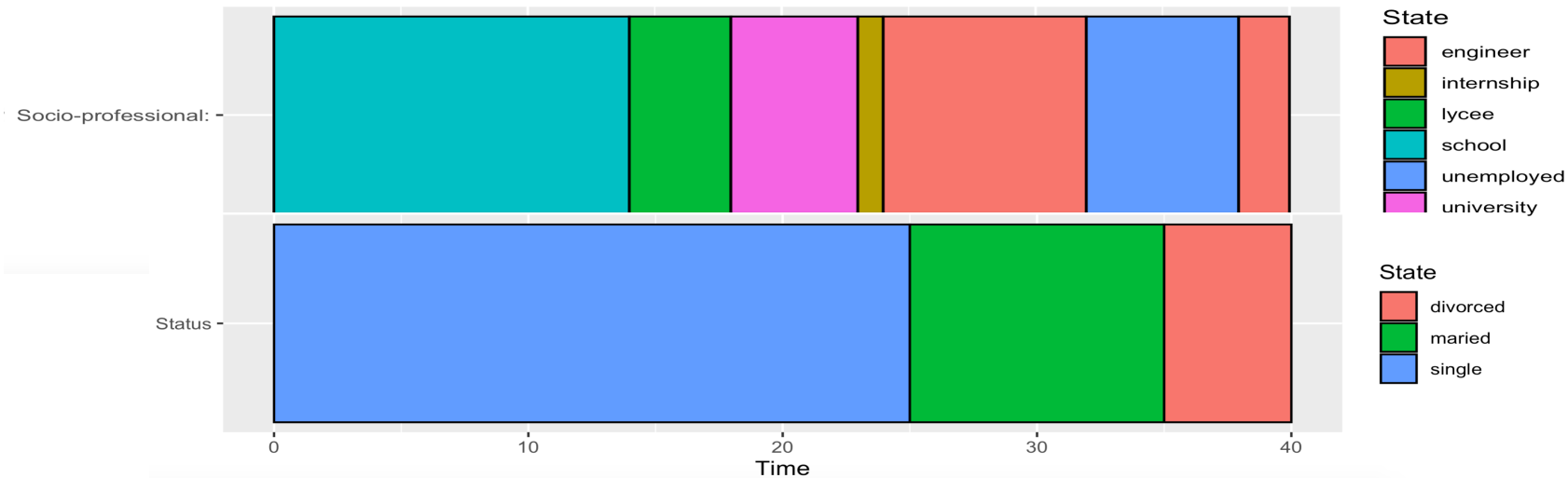
$$X^j = \{X_t^j, t \in [0, T]\},$$

with $X_t^j \in \mathcal{S}^j = \{s_1^j, \ldots, s_{K_j}^j\}$, $K_j \geq 2$.

Multivariate categorical functional data

Example.

Socio-professional (X_1) and marital (X_2) status.
($p = 2$, $K_1 = 6$, $K_2 = 3$).



X can be viewed as a categorical random variable with $K = \prod_{j=1}^{p} K_j$ states,

$$\mathcal{S} = \mathcal{S}^1 \times \ldots \times \mathcal{S}^p$$

Then, `cfda` **can be applied for reasonable** p **and** K_j**'s!**

Multivariate categorical functional data

Linear approximation of generating eigen-processes

The generating eigen-process associated to X is given by :

$$\xi_t = \mathbb{E}_t(z) \quad = \mathbb{E}(z|X_t^1, \ldots, X_t^p)$$

$$\underset{additive}{\approx} f_1(X_t^1) + \ldots f_p(X_t^p)$$

$$\underset{linear}{=} \sum_{i=1}^{K_1} \beta_i^1(t) 1_{X_t^1 = s_i^1} + \ldots + \sum_{i=1}^{K_p} \beta_i^p(t) 1_{X_t^p = s_i^p}$$

$$= \sum_{j=1}^{p} \sum_{i=1}^{K_j} \beta_i^j(t) 1_{X_t^j = s_i^j}.$$

Multivariate categorical functional data

Linear approximation of generating eigen-processes

Then, the encoding function for state $s = (s_{i_1}^1, \ldots, s_{i_p}^p) \in \mathcal{S}$ is given by :

$$a^s(t) = \sum_{j=1}^{p} \beta_{i_j}^j(t), \quad t \in [0, T].$$

If $\{\phi_1, \ldots, \phi_m\}$,

$$\phi_l : [0, T] \to \mathbb{R}, l = 1, \ldots, m,$$

is a basis of functions, then for all $j = 1, \ldots, p$ and $i = 1, \ldots, K_j$, consider the approximation :

$$\beta_i^j(t) \approx \sum_{l=1}^{m} \alpha_l^{(j,i)} \phi_l(t), \quad \forall t \in [0, T].$$

Multivariate categorical functional data

Linear approximation of generating eigen-processes

The coefficient vector

$$\alpha = \left(\alpha_l^{(j,i)}\right)_{j=1:p, i=1:K_j, l=1:m}$$

is solution of the eigen-vector problem

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

under the constraint

$$\alpha' F\alpha = 1.$$

where G and F are of size $m(K_1 + \ldots + K_p) \times m(K_1 + \ldots + K_p)$.

Multivariate categorical functional data

F and G matrices

F and G are indexed by
$(j, i, l), (j', i', l') \in (1 : p) \times (1 : K_j) \times (1 : m)$, with elements

$$G[(j, i, l), (j', i', l')] = Cov(V_l^{(j,i)}, V_{l'}^{(j',i')}),$$

$$V_l^{(j,i)} = \int_0^T \phi_l(t) 1_{X_t^j = s_i^j} dt,$$

and

$$F[(j, i, l), (j', i', l')] = \mathbb{E}(U_{(j,i,l),(j',i',l')}),$$

$$U_{(j,i,l),(j',i',l')}) = \int_0^T \phi_l(t)\phi_{l'}(t) 1_{X_t^j = s_i^j} 1_{X_t^{j'} = s_{i'}^{j'}} dt.$$

Note : F is singular, then $F \to F + \epsilon I$.

Multivariate categorical functional data

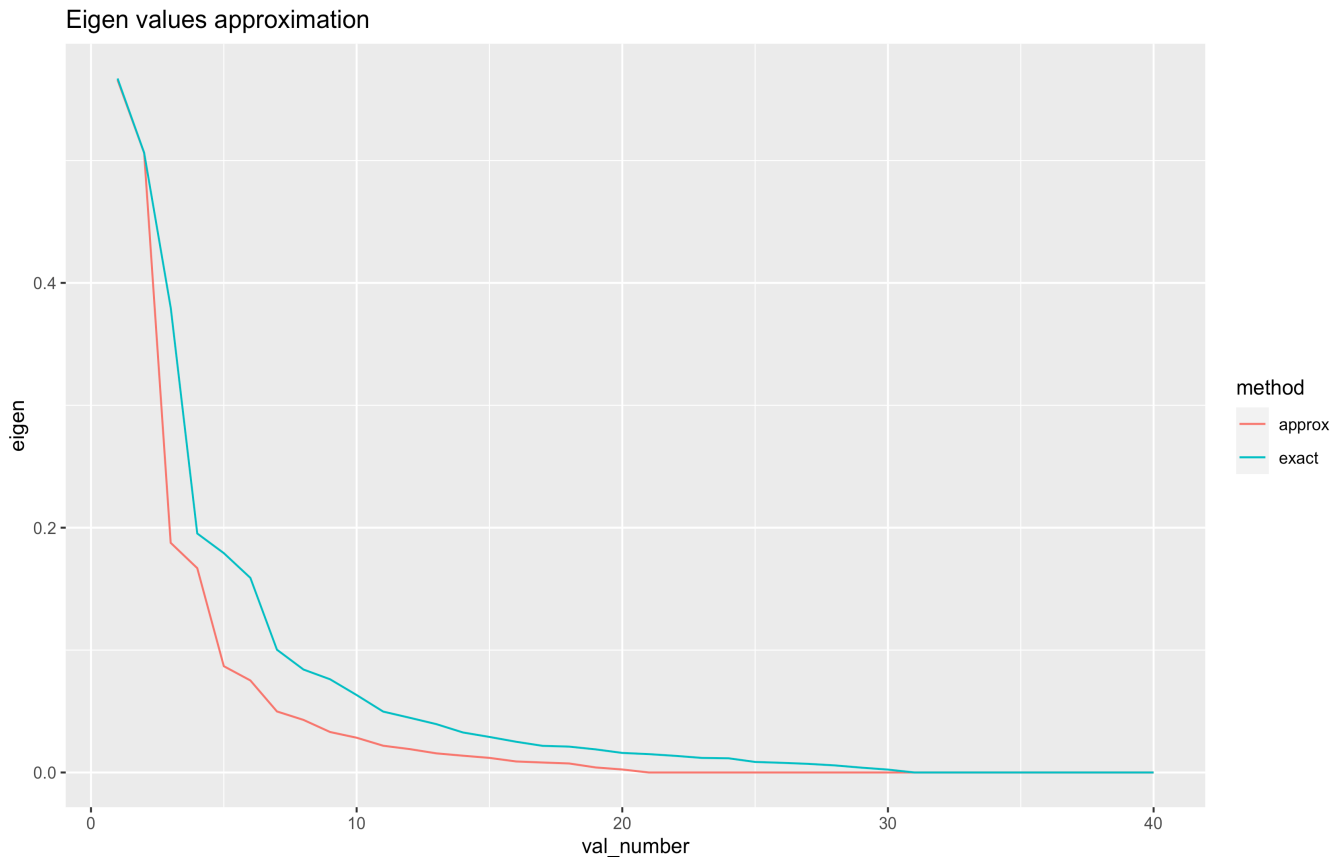
Numerical example : Two independent Markov processes

$X = (X_1, X_2)$:

$X_1 : K_1 = 2,\ \mathcal{S}^1 = \{1, 2\},\ \lambda = (0.5, 1)$

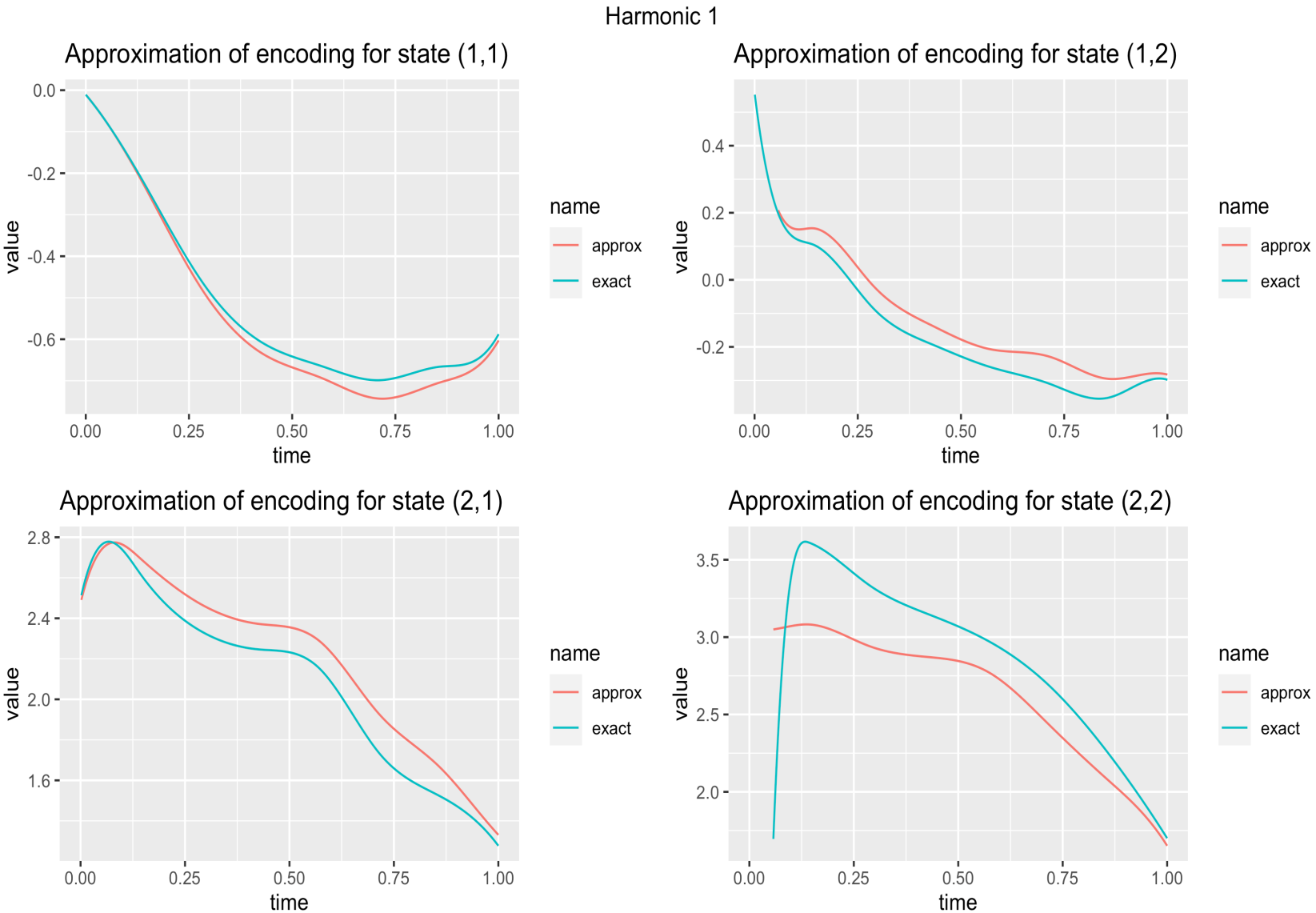
$X_2 : K_2 = 2,\ \mathcal{S}^2 = \{1, 2\},\ \lambda = (1, 1)$

Approximation : $n = 1000,\ m = 10,\ \{\Phi_i\}_{i=1:m}$: cubic B-splines.



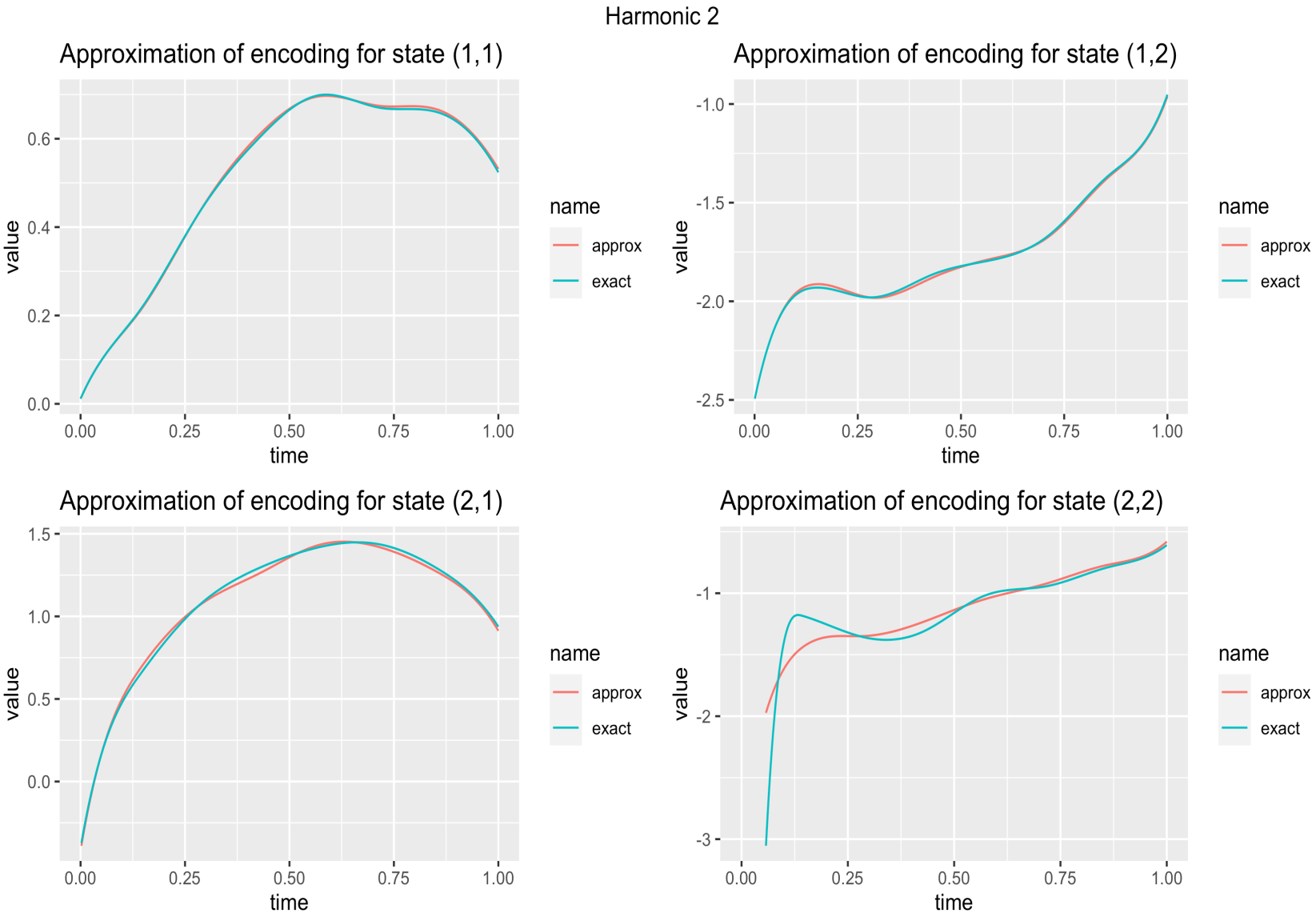
Multivariate categorical functional data

Numerical example : Two independent Markov processes



Multivariate categorical functional data

Numerical example : Two independent Markov processes



1. Saporta G. (1981) *Méthodes exploratoires d'analyse de données temporelles*, Cahiers du B.U.R.O, Université Pierre et Marie Curie,37-38, Paris.

2. Deville J.C. (1982) *Analyse de données chronologiques qualitatives : comment analyser des calendriers ?*, Annales de l'INSEE, No 45, p. 45-104.

3. Preda C., Grimonprez Q., Vandewalle V., *Categorical Functional Data Analysis. The cfda R Package*. Mathematics. 2021 ; 9(23) :3074. https ://doi.org/10.3390/math9233074