



HAL
open science

The Patient Path: a first approach

Rim Essifi

► **To cite this version:**

Rim Essifi. The Patient Path: a first approach. CMStatistics 2022 - The 15th International Conference of the ERCIM WG on Computational and Methodological Statistics, Dec 2022, Londres, United Kingdom. hal-03942217

HAL Id: hal-03942217

<https://hal.science/hal-03942217>

Submitted on 16 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Patient Path: a first approach

Rim Essifi

INRIA-MODAL, France

CMStatistics 2022, King's College London, 17/12/2022

- 1 Introduction
- 2 Existing functional data clustering approaches
 - Functional approaches
- 3 A first approach for clustering the patient path
 - Model and hypotheses
 - Extension to the multivariate case
- 4 Application to patient path data: the Include dataset
 - Evolution of FcECG as a function of time
 - MagmaClustR applied to 100 patients and 50 measurements
- 5 Research directions

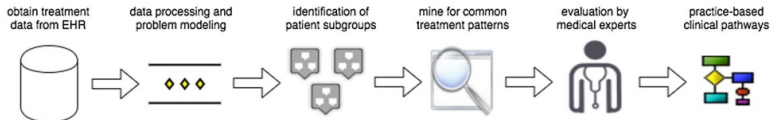
Collaboration between mathematicians and physicians

Patients pathways motivations

- ~> The aging of the population.
- ~> The increase in chronic diseases and patients suffering from multi-morbidity.
- ~> Financial resources and human constraints.
- ~> A considerable proportion of patients who do not receive appropriate evidence-based care because of unwarranted variation.

The quality of care patients is organised in a suboptimal way.

Main goal



Example: pre and post-surgery measures

Dataset: diagnoses, medical acts, UFs, authorization,...

Measures

| ID_INTERVENTION | PARAMETRE | VALEUR | ID_UNITE | DATE_MESURE |
|-----------------|-----------|--------|----------|--------------------|
| 661121 | FcECG | | 116 | 3 03/02/2018 05:45 |
| 661121 | FcECG | | 111 | 3 03/02/2018 05:45 |
| 661121 | FcECG | | 109 | 3 03/02/2018 05:46 |
| 661121 | FcECG | | 114 | 3 03/02/2018 05:47 |
| 661121 | FcECG | | 103 | 3 03/02/2018 05:47 |
| 661121 | FcECG | | 109 | 3 03/02/2018 05:48 |
| 661121 | FcECG | | 104 | 3 03/02/2018 05:48 |
| 661121 | FcECG | | 108 | 3 03/02/2018 05:48 |
| 661121 | FcECG | | 106 | 3 03/02/2018 05:49 |

Patients

| ID_PATIENT | DATE_NAISSANCE | SEXE |
|------------|----------------|------|
| 291743 | 1943-08-17 | F |
| 160413 | 1964-10-07 | F |
| 212991 | 1963-03-06 | F |
| 196615 | 1994-07-30 | F |
| 298214 | 1948-07-31 | M |
| 181364 | 1966-05-22 | F |
| 203297 | 1982-10-23 | F |
| 243583 | 1974-02-20 | F |
| 273458 | 1982-02-28 | F |

Procedures

| ID_PATIENT | ID_INTERVENTION | POIDS | TAILLE | IMC | ASA | URGENCE | SERVICE_NM | SERVICE | DATE_INTERVENTION | ID_SEJOUR |
|------------|-----------------|-------|--------|------|-----|---------|------------|-------------------|-------------------|-----------|
| 4 | 347816 | NA | NA | NA | 1 | 0 | 46 | Chir PÂ@diatrique | 14/05/2011 08:20 | 66127 |
| 4 | 379043 | 39 | 159 | 15,4 | 2 | 0 | 46 | Chir PÂ@diatrique | 03/12/2011 09:03 | 90892 |
| 7 | 659230 | 64 | 157 | 25,9 | 2 | 0 | 14202 | Bloc spe 1280 | 08/02/2016 08:05 | 310497 |
| 16 | 333480 | 90 | 152 | 38,9 | 2 | 0 | 44 | CMCA | 22/02/2011 09:21 | 54772 |
| 20 | 555366 | 67 | 180 | 20,6 | 1 | 0 | 32 | ORL | 05/08/2014 08:10 | 226258 |
| 21 | 288766 | 77 | 178 | 24,3 | 2 | 0 | 31 | Bloc Commun | 18/06/2010 12:31 | 22008 |
| 21 | 330430 | NA | NA | NA | 1 | 0 | 49 | Traumatologie | 12/02/2011 08:04 | 52301 |

Stays

| LIB_MODE_ENTREE | LIB_MODE_SORTIE | DATE_ENTREE_SEJOUR | DATE_SORTIE_SEJOUR | ID_SEJOUR |
|-------------------|-----------------|--------------------|--------------------|-----------|
| Domicile urgences | Domicile | 2016-05-11 | 2016-05-12 | 326816 |
| Domicile urgences | Mutation mco | 2012-07-31 | 2012-08-03 | 122966 |
| | | NA | NA | 256870 |
| | | NA | NA | 29320 |
| Domicile | Mutation mco | 2010-02-21 | 2010-03-13 | 8746 |
| | | NA | NA | 299 |
| Domicile urgences | Mutation mco | 2010-01-07 | 2010-01-13 | 2587 |
| Domicile urgences | Mutation mco | 2009-12-10 | 2010-01-07 | 1186 |

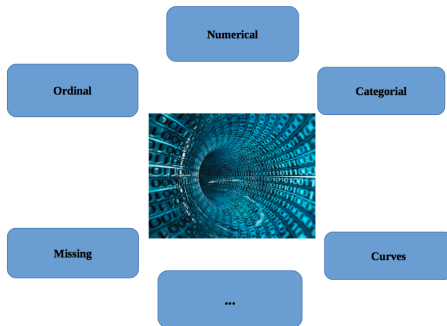
Data from diverse sources

- Data supplemented by those from the EDS of the **Lille University Hospital**; and from the **National Health Data Hub** (SNDS, Health Data Hub)
- Routine data of more than **1.5 million patients over the last 12 years**
- From a software collecting **medical acts, diagnoses, drug administration, the results of medical biology**, visits to the care units (and the severity of the unit), vital signs, the evaluation scales, and all the textual data

Data Extraction

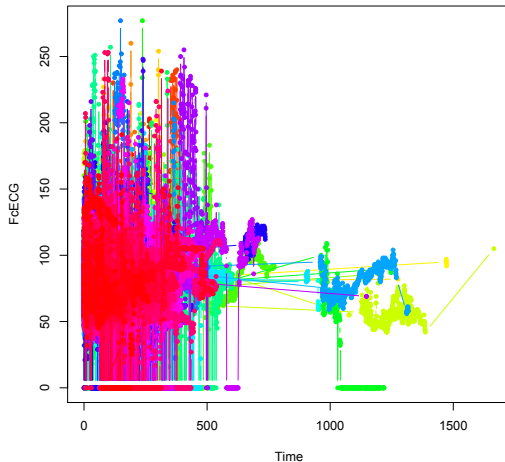
| Track definition | | Feature formalisation | |
|--|--|--|--|
| Raw data | Track | Feature | |
| <p>Hospital stay, admission and discharge</p> | <p>Hospital stay</p> | <p>Duration where $x = 1$ Length of stay : 3 days</p> | |
| <p>Intensive care, admission and discharge</p> | <p>Stay in intensive care unit</p> | <p>Maximum Passage in intensive care unit : 1</p> | |
| <p>Measurement : Mean arterial pressure</p> | <p>Mean arterial pressure < 65 mmHg</p> | <p>Duration where $x = 1$ Total duration of hypotension with mean arterial pressure < 65 mmHg : 25 minutes</p> <p>Count Number of episodes of hypotension with mean arterial pressure < 65 mmHg : 2</p> | |
| <p>Biology : Kalemia</p> | <p>Biology : TODD</p> | <p>TODD : TODD</p> | |
| <p>Drugs : V03AE Kayexalate with dose</p> | <p>V03AE Kayexalate</p> | <p>Maximum Administration of V03AE Kayexalate during hospital stay : 1</p> | |
| | <p>V03AE Kayexalate ≥ 30</p> | <p>Sum Total dose of V03AE Kayexalate during hospital stay : 90 grams</p> | |
| | <p>V03AE Kayexalate ≥ 30</p> | <p>Maximum Administration of V03AE Kayexalate with dose ≥ 30 grams : 1</p> | |
| | <p>V03AE Kayexalate ≥ 30 in ICU</p> | <p>Sum Total dose of V03AE Kayexalate during ICU stay : 30 grams</p> | |

Complex, Multidimensional data: time, space



How to combine different types of the data for a physician decision making?

Simple example: Time evolution of FcECG during surgery



From original data to observation of patient path random variable?

A Patient path *as* observation of a random variable valued in a (*complex*) space of paths :

$$\mathbf{Y}_t = \left\{ (Y_{t_1}, \dots, Y_{t_d})^\top : t_j \in \mathcal{T}_j, j = 1, \dots, d \right\},$$

$$Y_{t_j} : \mathcal{P}_j \rightarrow \mathcal{S}_j$$

$$\mathcal{T}_j \subseteq \mathbb{R}, \mathcal{S}_j = \mathbb{R} \text{ (curve)}$$

$$\mathcal{T}_j \subseteq \mathbb{R}, \mathcal{S}_j = \{e_1, e_2, \dots, e_K\} \text{ (sequence)}$$

$$\mathcal{T}_j \subseteq \mathbb{R}^2, \mathcal{S}_j = \mathbb{R} \text{ (image/surface)}$$

...

First goal : clustering

How to construct a generic model for irregularly stamped multivariate mixed functional data with missing values to perform clustering. Application: identify subgroups of patients?

Existing functional data clustering approaches

- **The two-stage methods (dimension reduction+multivariate clustering):** Kayano et al. (2010); Peng and Muller (2008); Rossi et al. (2004), Abraham et al. (2003),...
- **Nonparametric methods (with distances or dissimilarities):** Ieva et al (2012); Tokushige et al. (2007); Dabo-Niang et al. (2007); Ferraty and Vieu (2006); Tarpey and Kinateder (2003),...
- **Nonparametric methods (heuristics or geometry criteria):** Hébrail et al (2010), Yamamoto (2012),...
- **Model-based approaches:** Bouveyron and Jacques (2011); Chiou and Li (2007); Jacques and Preda (2013); James and Sugar (2003);...
- **Raw data approach (discretized times):** Bouveyron and Brunet (2012).

Model-based approach for irregularly stamped multivariate and mixed functional data

Recent works using Gaussians processes:

- clustering functional univariate categorical data : Preda et al. (2021)
- clustering functional univariate continuous data : **Leroy et al. (2020)**; Murphy and Murphy (2020)
- supervised learning for continuous multivariate functional data : Constantin et al. (2020)

State of the art

Introduction of Gaussian processes: Wiener, 1949.

~ Study of Gaussian processes: Thompson, 1956; Matheron, 1973; Cressie, 1993.

~ Application to the development of learning methods for regression problems O'Hagan, 1978; Williams and Rasmussen, 1996.

- 1 Introduction
- 2 Existing functional data clustering approaches
 - Functional approaches
- 3 A first approach for clustering the patient path
 - Model and hypotheses
 - Extension to the multivariate case
- 4 Application to patient path data: the Include dataset
 - Evolution of FcECG as a function of time
 - MagmaClustR applied to 100 patients and 50 measurements
- 5 Research directions

First approach for univariate continuous case: Leroy et al (2020)

Fix $n, K \geq 1, T > 0$ and l_1, \dots, l_n .

Data is collected from n different patients. For $1 \leq i \leq n$, set:

$\leadsto t_i = (t_{i,1}, \dots, t_{i,l_i}) \in [0, T]^{l_i}$: the observation times of patient i .

$\leadsto Y_i(t_{i,l_i}) \in \mathbb{R}$.

\leadsto Gaussian process mixture model:

- we associate a latent binary random vector $Z_i = (Z_{i1}, \dots, Z_{iK})$ to each individual, indicating in which cluster it belongs: if the i -th individual comes from the k -th cluster, then $Z_{ik} = 1$ and 0 otherwise.
- Assume that these latent variables are coming from the same multinomial distribution: $Z_i \sim \mathcal{M}(1, \pi)$, $1 \leq i \leq n$, with

$$\pi = (\pi_1, \dots, \pi_K)^t \text{ and } \sum_{k=1}^K \pi_k = 1.$$

First approach for univariate continuous case: Leroy et al (2020)

Assume that the i -th individual belongs to the k -th group, let the data observed at time t be as observation of a sum of a cluster-specific mean process and an individual-specific centered process:

$$Y_i(t) = \mu_k(t) + f_i(t) + \varepsilon_i(t), \quad t \in [0, T].$$

Hypotheses

For $1 \leq i \leq n$, and $1 \leq k \leq K$,

$\rightsquigarrow \mu_k(\cdot) \sim \mathcal{GP}(m_k(\cdot), c_{\gamma_k}(\cdot, \cdot))$: common mean process of the k - cluster,

$\rightsquigarrow f_i(\cdot) \sim \mathcal{GP}(0, A_{\theta_i(\cdot, \cdot)})$: specific process of the i - individual,

$\rightsquigarrow \varepsilon_i(\cdot) \sim \mathcal{GP}(0, \sigma_i^2 I)$: specific noise the i -individual,

$\rightsquigarrow \Theta = \{(\gamma_k)_{1 \leq k \leq K}, (\theta_i)_{1 \leq i \leq n}, (\sigma_i^2)_{1 \leq i \leq n}, \pi\}$: the set of all parameters of the model.

Hypotheses

- $\{\mu_k\}_{1 \leq k \leq K}$ are independent,
- $\{f_i\}_{1 \leq i \leq n}$ are independent,
- $\{Z_i\}_{1 \leq i \leq n}$ are independent,
- $\{\varepsilon_i\}_{1 \leq i \leq n}$ are independent,
- For all $1 \leq i \leq n$, $1 \leq k \leq K$, μ_k , f_i , Z_i are independent.

Inference: a variational version of the EM algorithm is used (MagmaClustR, Leroy et al (2020)).

Extensions

- **The multivariate continuous case:** $Y_i(t) \in \mathbb{R}^d$, $1 \leq i \leq n$, $t \in [0, T]$, with $d \geq 1$.
- **The categorical case:** $Y_i(t) \in C$, $1 \leq i \leq n$, $t \in [0, T]$, where C is a finite set.

A more general extension: the multivariate mixed case

Fix $n, d, p, q, K \geq 1, T > 0$, with $p + q = d$, $(l_{i,j})_{1 \leq i \leq n, 1 \leq j \leq d}$ positive integers. Set p positive integer c_1, \dots, c_p .

Consider p finite sets $C_j = \{1, \dots, c_j\}$ for $1 \leq j \leq p$, n different patients.

For $1 \leq i \leq n$, set

$\rightsquigarrow t_{i,l}^{(j)} \in [0, T]$ $1 \leq j \leq d, 1 \leq l \leq l_{i,j}$: the observation times.

$\rightsquigarrow (Y_i^{(j)}(t_{i,l}^{(j)}))_{1 \leq j \leq d, 1 \leq l \leq l_{i,j}} \in C_1 \times \dots \times C_p \times \mathbb{R}^q$.

$\rightsquigarrow p_{c,l}^{(j)}(t_{i,l}^{(j)}) = \mathbb{P}(Y_i^{(j)}(t_{i,l}^{(j)}) = c_j)$, for $c \in C_j$, $1 \leq j \leq p$ and $1 \leq l \leq l_{i,j}$.

Gaussian process mixture model

- We associate a latent binary random vector $Z_i = (Z_{i1}, \dots, Z_{iK})$ to each i -individual, indicating in which cluster it belongs: if the i -th individual comes from the k -th cluster, then $Z_{ik} = 1$ and 0 otherwise.
- Assume that these latent variables are coming from the same multinomial distribution: $Z_i \sim \mathcal{M}(1, \pi)$, $1 \leq i \leq n$, with

$$\pi = (\pi_1, \dots, \pi_K)^t \text{ and } \sum_{k=1}^K \pi_k = 1.$$

The model

Assuming that the i -th individual belongs to the k -th group, we can define its functional expression as the sum of a cluster-specific mean process and an individual-specific centred process:

$$\begin{cases} p_{c,i}^{(j)}(t^{(j)}) = \frac{\exp(m_{k,c}^{(j)}(t^{(j)}))}{1 + \sum_{c=2}^{c_j} \exp(m_{k,c}^{(j)}(t^{(j)}))}, & 2 \leq c \leq c_j, 1 \leq j \leq p, \\ Y_i^{(j)}(t^{(j)}) = \mu_k^{(j)}(t^{(j)}) + f_i^{(j)}(t^{(j)}) + \varepsilon_i^{(j)}(t^{(j)}), & p+1 \leq j \leq d, \end{cases}$$

$$t^{(j)} \in [0, T], 1 \leq j \leq d.$$

Hypotheses

For $1 \leq j \leq d$, $1 \leq i \leq n$, and $1 \leq k \leq K$,

$\rightsquigarrow \mu_k^{(j)}(\cdot) \sim \mathcal{GP}(m_k^{(j)}(\cdot), c_{\gamma_k^{(j)}}(\cdot, \cdot))$: common j -mean process of the k -cluster,

$\rightsquigarrow f_i^{(j)}(\cdot) \sim \mathcal{GP}(0, A_{\theta_i^{(j)}}^{(j)}(\cdot, \cdot))$: specific j -process of the i -individual,

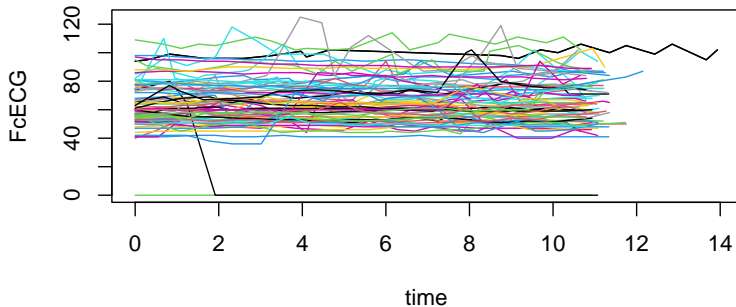
$\rightsquigarrow \varepsilon_i^{(j)}(\cdot) \sim \mathcal{GP}(0, \sigma_i^{2(j)} I)$: specific j -noise of the i -individual,

$\rightsquigarrow \Theta = \{(\gamma_k^{(j)})_{1 \leq j \leq d, 1 \leq k \leq K}, (\theta_i^{(j)})_{1 \leq j \leq d, 1 \leq i \leq n}, (\sigma_i^{2(j)})_{1 \leq j \leq d, 1 \leq i \leq n}, \pi\}$: the set of all parameters of the model.

Hypotheses

- $\{\mu_k\}_{1 \leq k \leq K}$ are independent,
- $\{f_i\}_{1 \leq i \leq n}$ are independent,
- $\{Z_i\}_{1 \leq i \leq n}$ are independent,
- $\{\varepsilon_i\}_{1 \leq i \leq n}$ are independent,
- For all $1 \leq i \leq n$, $1 \leq k \leq K$, μ_k, f_i, Z_i are independent,
- For all $1 \leq i \leq n$, $\{Y_i^{(j)} | Z_i\}$, $1 \leq j \leq d$ are independent.

Evolution of FcECG as a function of time

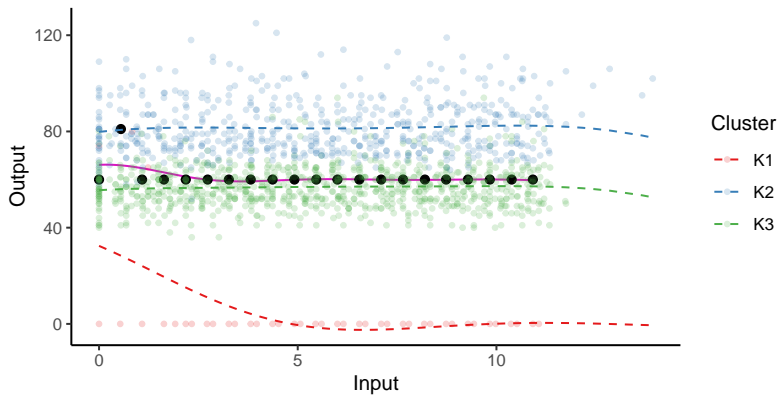


Selection of the clusters number



Clustering

Mixture of GP predictions



- Perform clustering using the Gaussian processes mixture model described above in the multivariate continuous case.
- Perform clustering using the Gaussian processes mixture described above in the categorical case.
- Perform clustering using the Gaussian processes mixture described above in the multivariate mixed case.
- Take into account the possibility of having missing data.
- Predict the future of a patient path in term of covariates.
- Answer clinicians' questions based on our generic model.
- Develop a package aimed at clinicians.

Thank you for your attention!



Sidi Bou said, Tunisia