Classification des trajectoires typiques d'évolution : Trajectoires latentes et applications en santé publique

Jules. B Tchatchueng.M

Chargé de Recherche au Centre Pasteur du Cameroun

9 juillet 2019

Introduction

Méthodes

Introduction

Contexte (classification et intelligence artificielle)

Intelligence artificielle

Définition

L'ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine

Les différentes formes d'intelligence artificielle

- **Machines réactives**
- Les machines à mémoire limitée
- Théorie de l'esprit et conscience de soi

Deux types de méthodes de classification ou apprentissage permettent de programmer des machines réactives:

1. Supervisé
   - Exemples:
     - Le trieur de courriels
     - Plateforme de streaming
     - Gestion de la relation client en banque
2. **Non supervisé**
   - Exemples:
     - classification des sous espèces végétales ou animales
     - classification des cellules

Classification non supervisée

Deux grandes approches pour la recherche des partitions :

1. Approche utilisant des critères métriques
   - Exemples
     - Algorithme K-means, K-centroides
     - Classification hiérarchique
2. Approche probabiliste:classification par les classes latentes
   - 2.1 Approche non paramétrique
     - Estimation de la distribution à partir des données
   - 2.2 Approche paramétrique
     - Modèles de mélange
     - Modèles fonctionnels
     - Processus ponctuels

Données longitudinales

Les études de cohorte sont devenues un outil essentiel dans la recherche en épidémiologie étude de cohorte ou suivi de patients: collecte répétée de données sur un ensemble d'individus au cours du temps.

Structure de données longitudinales $\rightarrow$ -Mesures répétées par unité statistique
- Corrélation entre les données de la meme classe
- Variabilité intra et inter classe

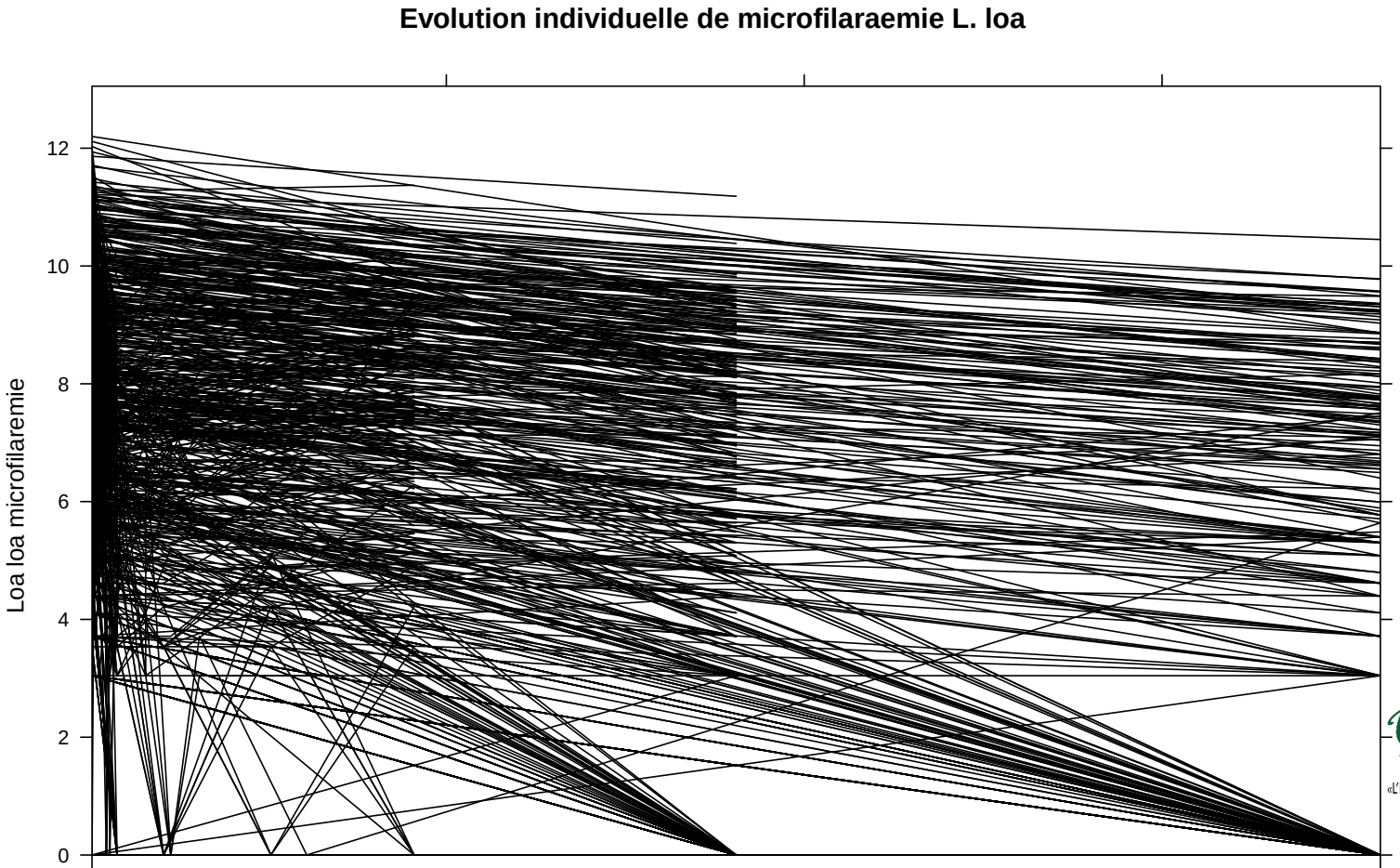
Présentation d'un tableau de données longitudinales (n individus, p mesures).

$$\begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,(p-1)} & y_{1,p} \\ y_{2,1} & \vdots & \vdots & \vdots & y_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{(n-1),1} & y_{(n-1),2} & \cdots & \cdots & y_{(n-1),p} \\ y_{n,1} & y_{n,2} & \cdots & \cdots & y_{n,p} \end{pmatrix}$$

On notera $Y_i = (Y_{i,1}, \cdots, Y_{i,p})$ la trajectoire observée de l'individu i, $\quad i = 1, \cdots, n$

Exemple

```
##                LoaJ  LnLoaJ.0 LnLoaJ.2 LnLoaJ.8
## HCY_1.0   58360.83 10.974417 8.530536 6.803405
## HCY_10.0  54827.85 10.911972 9.905103       NA
## HCY_100.0 10865.58  9.293447 7.532345       NA
```



Exemple (suite)

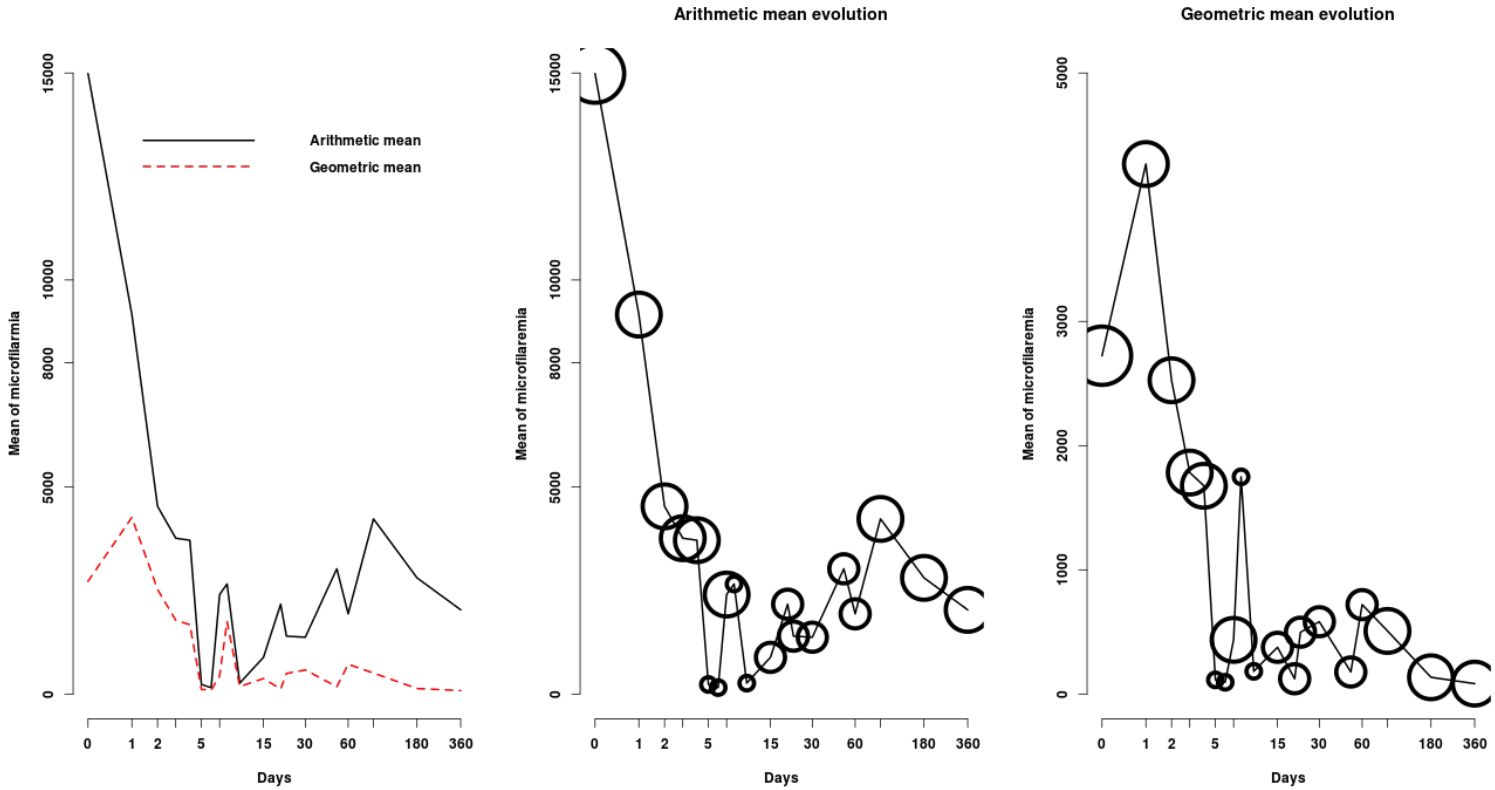


Figure 2: Evolution moyenne

Problème

Comment partitionner les trajectoires Y_i, $\quad i = 1, \cdots, n$?

Méthodes

Méthode Kmeans

Generalisation du Kmeans classique

Idée

0n cherche à partitionner les n trajectoires en k ensembles $C = C_1, C_2, \ldots, C_k$ $(k \leq n)$ en minimisant la distance entre les trajectoire à l'intérieur de chaque partition :

$$\underset{C}{argmin} \sum_{i=1}^{k} \sum_{y_j \in C_i} D(y_j, \mu_i)$$

ou μ_i est la trajectoire moyenne dans C_i et $D(,)$ une mesure de dissimilarité sur $\mathbb{R}^p$

Exemple de mesure de dissimilarité sur $\mathbb{R}^p$

La distance euclidienne sur $\mathbb{R}^p$

$$D(Y_i, Y_j) = \sqrt{\sum_{k=1}^{p} (Y_{i,k} - Y_{j,k})^2}$$

La distance de Manathan

$$D(Y_i, Y_j) = \frac{1}{p} \sum_{k=1}^{p} |Y_{i,k} - Y_{j,k}|$$

Description de l'algorithme

- **Initialisation**: Choisir k trajectoires qui représentent la position moyenne des partitions $\mu_1, \ldots, \mu_k$
- Répéter jusqu'à ce qu'il y ait convergence :
  - assigner chaque trajectoire à la partition la plus proche

$$C_i^t = \{Y_j; D(Y_j, \mu_i) \leq D(Y_j, \mu_{i^*}, \quad i^* \neq i\}$$

  - mettre à jour la moyenne de chaque cluster :

$$\mu_i^{(t+1)} = \frac{1}{|C_i^t|} \sum_{Y_j \in C_i^t} Y_j$$

ou $|C_i^t|$ est le nombre de trajectoires dans C_i^t

CHOIX DU NOMBRE DE CLUSTER

On note A et B les matrices de variance covariance inter et intra cluster

$$A = \sum_{m=1}^{k} \frac{1}{|C_m|}(\mu_m - \mu)(\mu_m - \mu)'$$

ou μ est la trajectoire moyenne dans C

$$B = \sum_{m=1}^{g} \sum_{Y_i \in C_m} (Y_i - \mu_m)(Y_i - \mu_m)'$$

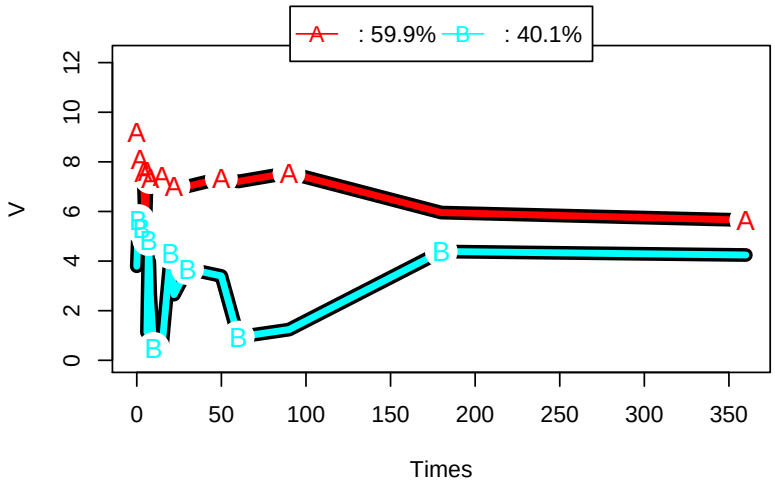
$A + B$ est la covariance totale.

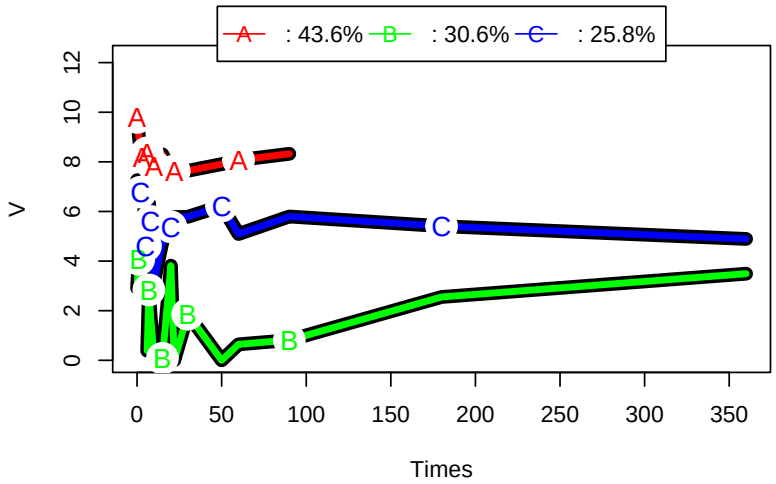
On notre K_{opt} le nombre optimal de cluster

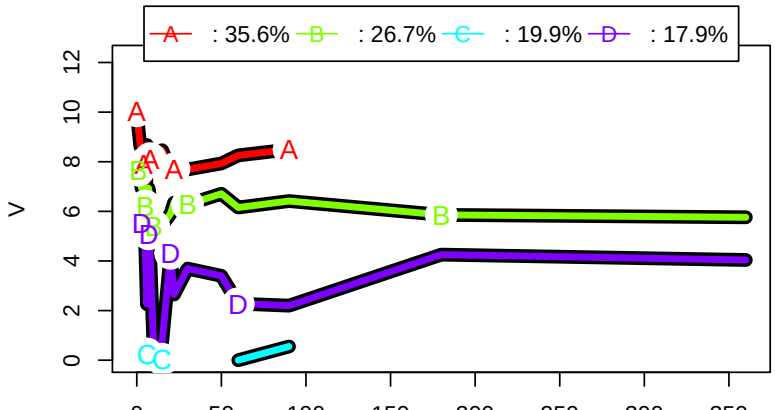
$$K_{opt} = \underset{k \in \{2 \ldots (n-1)\}}{argmax} \left(\frac{Trace(A)}{Trace(B)} \frac{n-k}{k-1} \right)$$

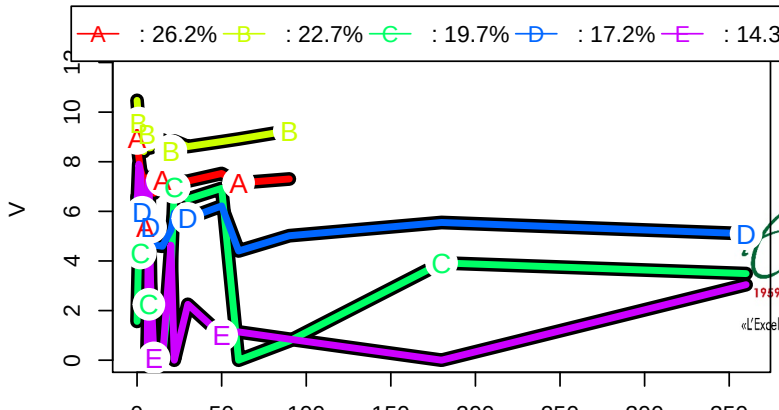
APPLICATION À L'EVOLUTION DES CHARGES MICROFILARIENNES

```
##  ~ Fast KmL ~
## ***********************************************************************************
## 100 S
```









Kmeans basé sur une distance de forme

Exemple

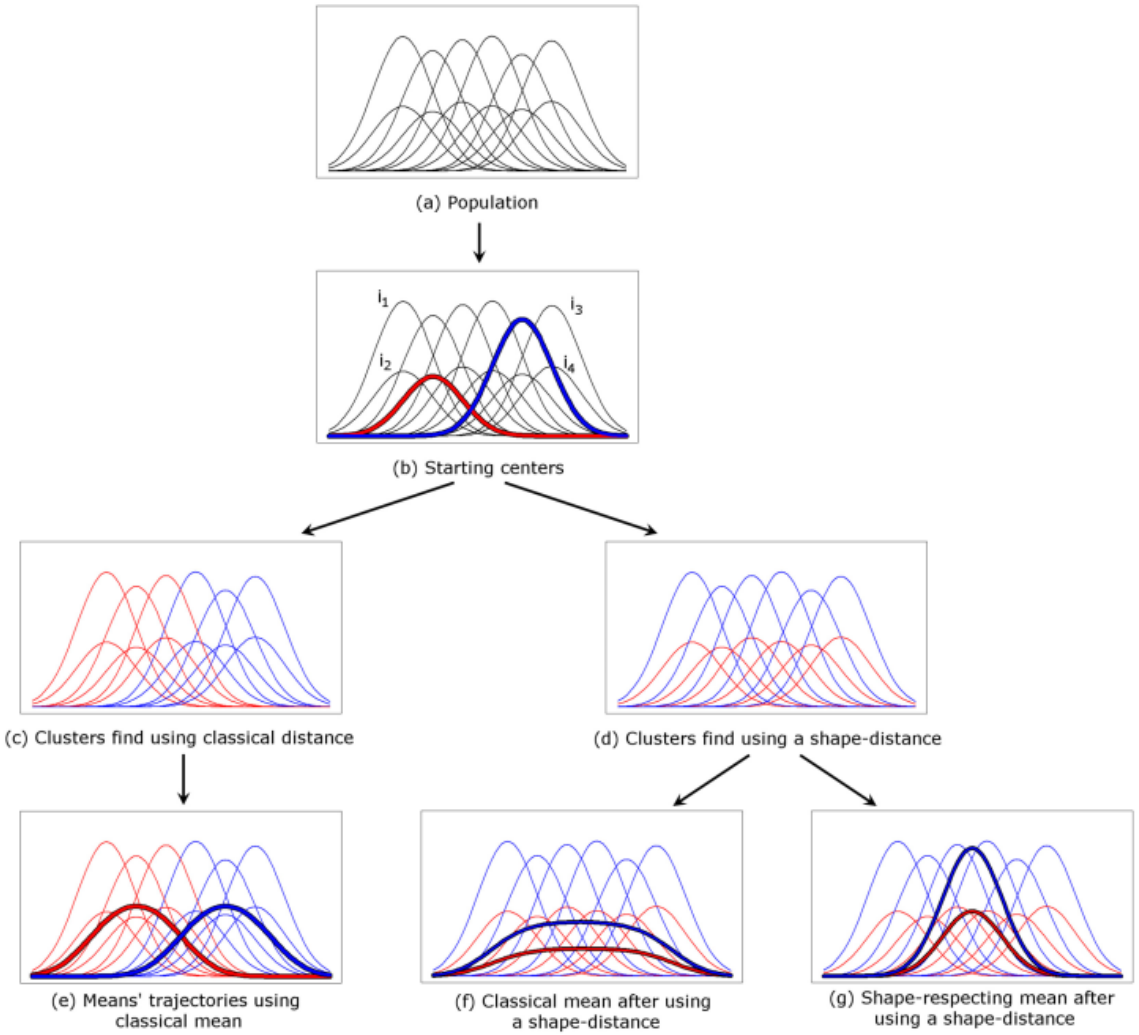


Figure 4: Evolution moyenne

Distance de Frechet

Soient P et Q deux courbes reparamétrées[1] par α et β.

La distance entre les courbes P et Q avec les reparamétrisation α et β ou point $s \in [0, T]$ est définie par:

$$d_{\alpha,\beta,s} = d\left(\begin{pmatrix}\alpha(s)\\ P(\alpha(s))\end{pmatrix}, \begin{pmatrix}\beta(s)\\ Q(\beta(s))\end{pmatrix}\right)$$

ou $d(,)$ la distance euclidienne sur $\mathbb{R}^2$. \ On définit la distance de fréchet entre P et Q reparametrées par α et β sur $[0, t]$ par:

$$F(P, Q) = \underset{\alpha,\beta}{Inf}\left[\underset{s\in[0,t]}{max}\left(d_{\alpha,\beta,s}(P, Q)\right)\right]$$

[1] une reparamétrisation sur $[0, t]$ est une fonction continue croissante et surjective de $[0, t]$ vers $[0, t]$

Algorithme Kmeans pour données longitudinales

Data: Population: n individuals $Y_1, ..., Y_n$

Result: Partition: Cluster vector of size n taking values in $[1...k]$

#Step 0: Initialization

k individuals $C_1, C_2, ..., C_k$ are randomly chosen in $Y_1, ..., Y_n$

s <-0

$Cluster_0$ <- (0, 0,. . . , 0) ## vector of size n

repeat . s<- s+1

. # Step s.1, phase maximization

. **for** i in 1. . . n **do**

. **for** j in 1. . . k **do**

. Compute $DistF_{i,j}$ (The Frechet distance between Y_i and C_j)

. Clusters $S(i)$<- j such that $DistF_{i,j}$ is smaller than $DistF_{i,j'}$ for $j' \neq j$

. **end**

. **end**

Algorithme Kmeans pour données longitudinales (suite)

. # Step s.2, phase expectation

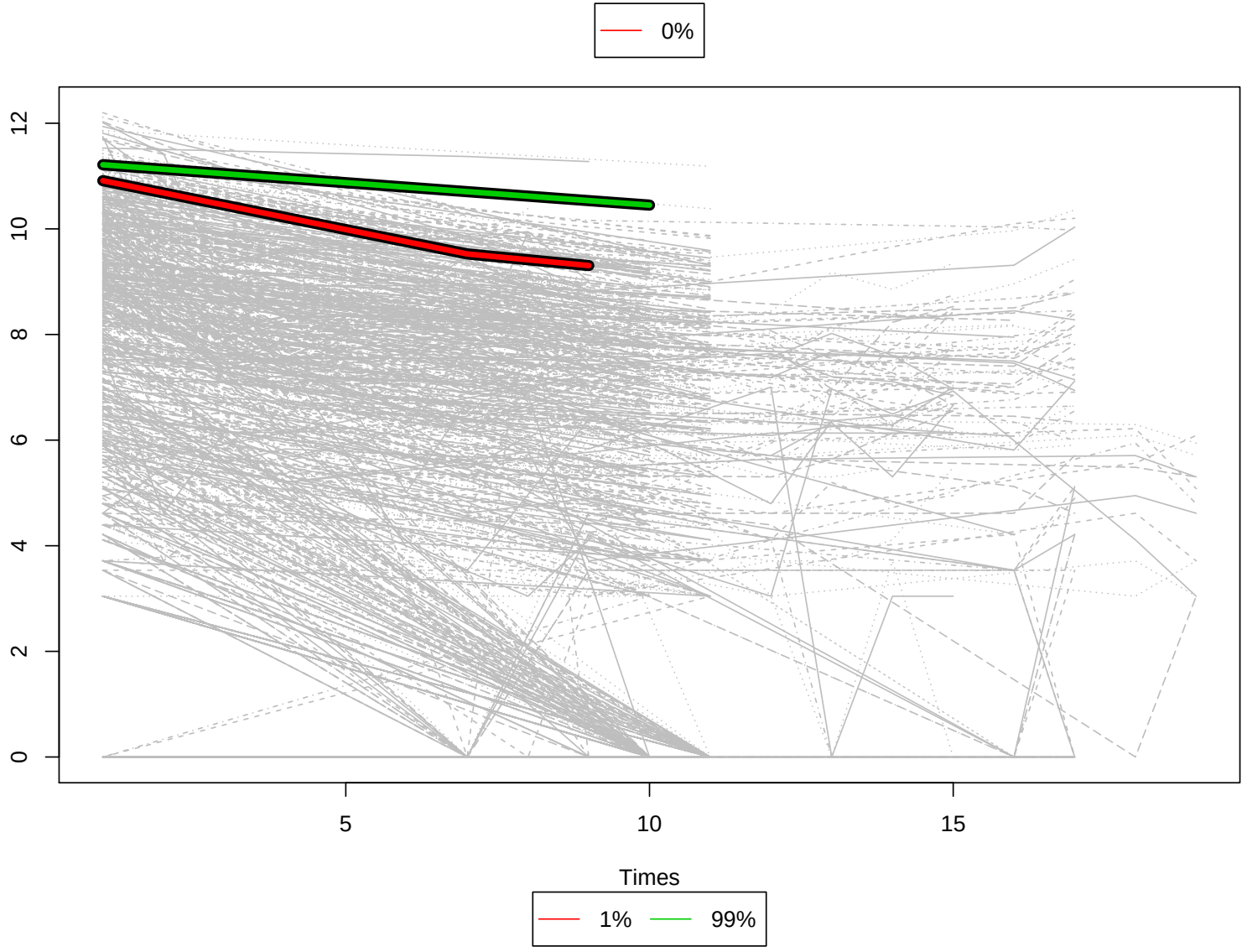
. **for** j in $1...k$ **do**

. Compute M_j , the Fréchet mean of clusters j (that is the Fréchet mean of all the Y_i such that Cluster $S(i) == j$)

. **end**

until $ClusterS == Cluster \quad S - 1$ or $s > MaxIteration$ }

APPLICATION À L'EVOLUTION DES CHARGES MICROFILARIENNES



APPLICATION À L'EVOLUTION DES CHARGES MICROFILARIENNES

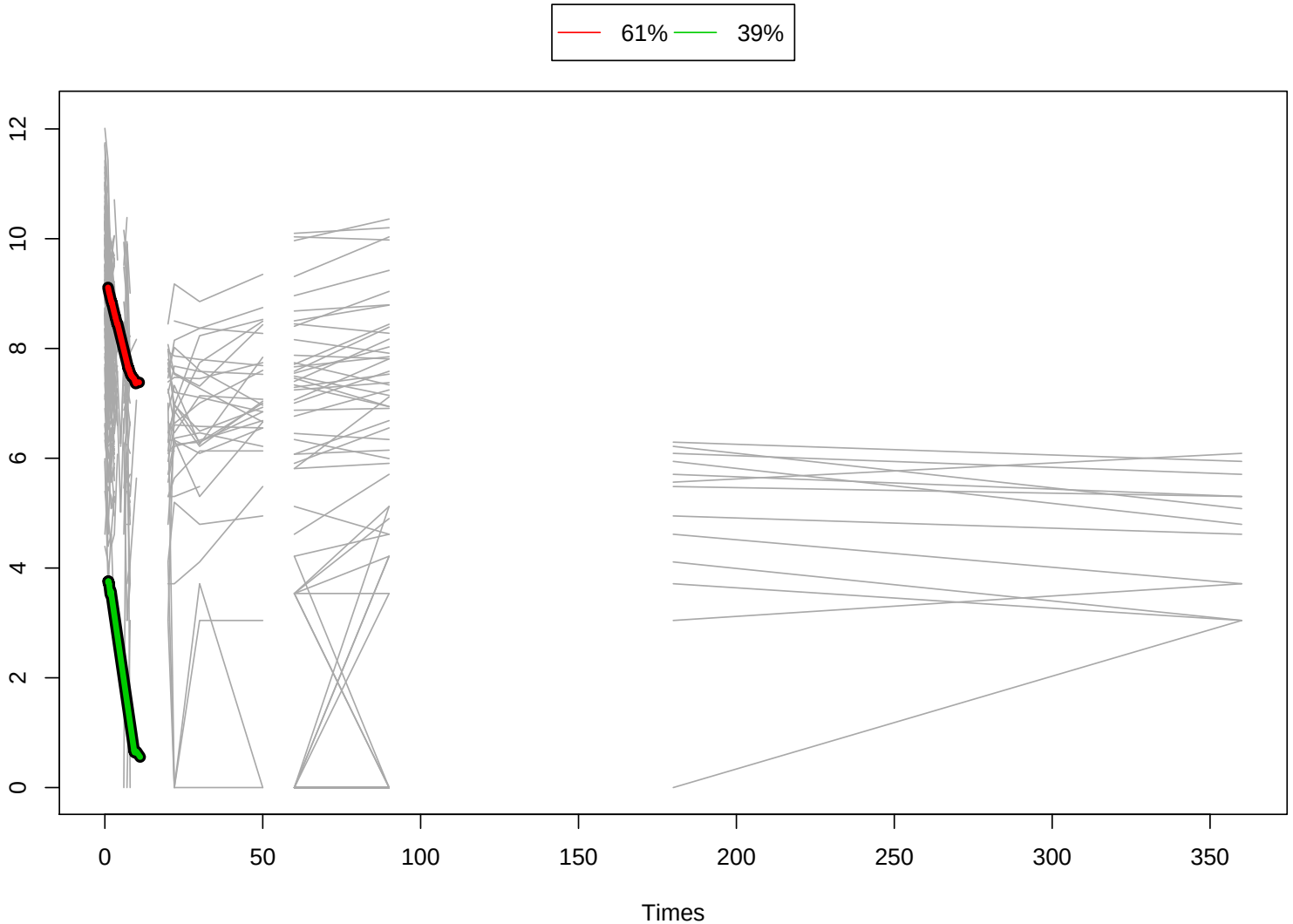


Figure 5: classification avec la distance de fréchet

CLASSIFICATION PAR LES CLASSES LATENTES

Variable Latente

Définition

- Variable hypothétique (dont on fait l'hypothèse)
- Variable non-mesurable: L'intelligence
- Variable qui synthétise plusieurs variables observées: le vieillissement

- Variable qui rend les variables observées indépendantes si l'on en tient compte.
  $\rightarrow$ Indépendance locale/indépendance conditionnelle
- Variable dont les valeurs correspondent àla valeur attendue si l'on avait répétée la mesure un nombre infini de fois
  $\rightarrow$ Espérance
- Variable qu'on ne peut calculer à l'aide de variables observées.
- Variable que l'on a pas observée dans notre échantillon

Types de modèles à variable latente

	Variables Latentes	
Variables Observée	qualitative	quantita
qualitative	Modèles de profils latent	Modèles de répo
quantitative	**Modèles à classe latente/ modèles de mélange**	Modèles à équation

Modèles à classe latente/ modèles de mélange

Notations

π_k: $0 < \pi_k \leq 1 \quad \forall k \in \{1, ..., K\}$ et $\sum_{k=1}^{K} \pi_k = 1$
$f_k, \quad k = 1, ..., K$ les densité de probabilité sur un espace $\mathbb{E}$

On défini la densité de mélange fini à K composantes:

$$g = \sum_{k=1}^{K} \pi_k f_k$$

Si les f_k sont issues de la même famille paramétrée par α_k alors:

$$g(., \theta) = \sum_{k=1}^{K} \pi_k f(., \alpha_k)$$

ou $\theta = (\pi_1, ..., \pi_K, \alpha_1, ..., \alpha_K$ est le paramètre du mélange.

Un individu issu d'un mélange de population est caractérisé par (Y, Z) ou Y est la variable observée Z la variable latente indiquant la classe latente.

Lorsque Y est une trajectoire le modèle est dit à **trajectoire latente**

Modèle de mélange pour la classification

- La variable observée à partir de laquelle on construit les classes.
- Comprend une variable latente à K catégories.
- Une Probabilité à posteriori d'appartenance à chaque classe pour chaque individu.
- Chaque classe latente regroupe les individus ayant la même trajectoire typique.
- Des indicateurs associés aux probabilités à posteriori.
- Une equation de scoring permettant de calculer les probabilités à posteriori à partir des indicateurs.

Vraisemblance du modèle

$$L(y_i, \theta) = \prod_{i=1}^{n} g(x_i, \theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(x_i, \alpha_k)$$

Estimation des paramètres du modèle

vraisemblance complète

Soit $z = (z_{ik})_{i=1..n,k=1..K}$ la matrice de classification $z_{ik} = 1$ si l'individu i est dans la classe k

$$L(x, z, \theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k f(x_i, \alpha_k)]^{z_{ik}}$$

La log-vraisemblance complète est:

$$l_c(z, \theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} log(\pi_k f(x_i, \alpha_k))$$

$$\theta^* = \underset{\theta \in \Theta}{argmax}(l_c(z, \theta))$$

Maximisation de la vraisemblance (EM)

L'algorithme EM maximise la fonction :

$$\begin{aligned} Q(\theta, \theta') &= \mathbb{E}(l_c(z, \theta)|x, \theta') \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} log(\pi'_k f(x_i, \alpha'_k)) \end{aligned}$$

t_{ik} est la probabilité à posteriori d'appartenance de l'individu i à la classe k lorsque $\theta = \theta'$:

$$\begin{aligned} t_{ik} = \mathbb{E}(z_{ik}|x, \theta') &= P(z_{ik} = 1|x, \theta') \\ &= \frac{\pi'_k f(x_i, \alpha'_k)}{\sum_{q=1}^{K} \pi'_q f(x_i, \alpha'_q)} \end{aligned}$$

Etape de l'algorithme EM

- **Initialisation**: choix arbitraire d'une solution initiale $\theta^{(0)}$
- **Etape E**(espérance): calcul des probabilité d'appartenance aux classes conditionnellement au paramètres courant:

$$t_{ik}^{(c)} = \frac{\pi_k^{(c)} f(x_i, \alpha_k^{(c)})}{\sum_{q=1}^{K} \pi_q^{(c)} f(x_i, \alpha_q^{(c)})}$$

- **Etape E** (Maximisation): On calcule les proportions des differentes classes

$$\pi_k^{(c+1)} = \frac{1}{n} \sum_{i=1}^{n} t_{ik}^{(c)}$$

Choix du modèle et du nombre de classes

Objectifs

- l'adequation du modèle aux données
- Obtenir la meilleure partition des données

Soit $\mathcal{M}$ la famille de modèles selectionnés. Cela revient à trouver le meilleur coupe (M, K) ou $M \in \mathcal{M}$ et $1 \leq K \leq K_{max}$ avec $K_{max} = n^{0.3}$.

Critères de sélection

- $AIC(M, K) = log(g(x|M, K, \theta^*)) + \eta_{M,K}$ ou $\eta_{M,K}$ est le nombre de paramètres libres du modèle (M, K)
- $BIC(M, K) = -log(g(x|M, K, \theta^*)) - \frac{\eta_{M,K}}{2} log(n)$
- $ICL(M, K) = BIC(M, K) + \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} log(t_{ik})$