Workshop Data Initiative 3

Modelling Patient Time-Series Data from Electronic Health Records using Gaussian Processes.

Marco A.F. Pimentel, David A. Clifton, Lei Clifton, Lionel Tarassenko.
Department of Engineering Science, University of Oxford.

NIPS 2015 Workshop on Machine Learning in Healthcare.

Simon Bussy

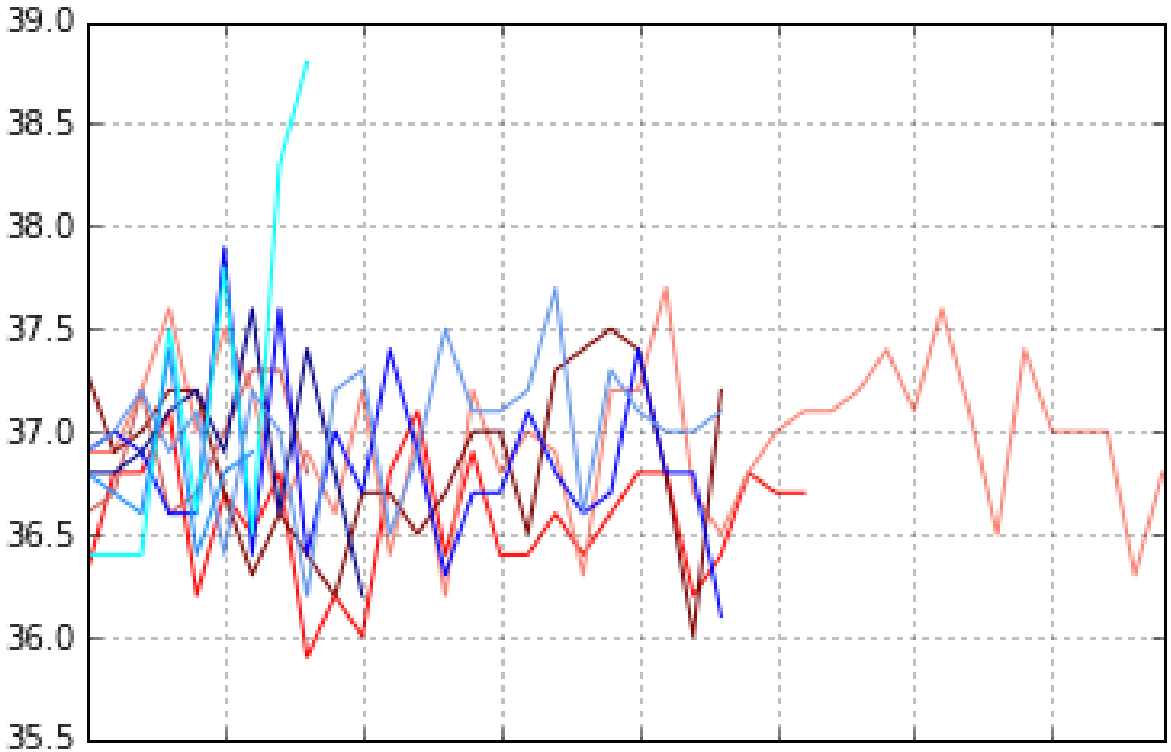
Le 11 Mars 2016

Introduction

- Electronic Health Records : données hétérogènes

Introduction

- Electronic Health Records : données hétérogènes
- Données similaires : problèmes de durée et alignement



Evolution de la température au cours de 10 séjours

Introduction

- Electronic Health Records : données hétérogènes
- Données similaires : problèmes de durée et alignement
- Gaussian Process : représentation des trajectoires de paramètres vitaux

Introduction

- Electronic Health Records : données hétérogènes
- Données similaires : problèmes de durée et alignement
- Gaussian Process : représentation des trajectoires de paramètres vitaux
- But : dissocier les trajectoires ”anormales”

Gaussian Process

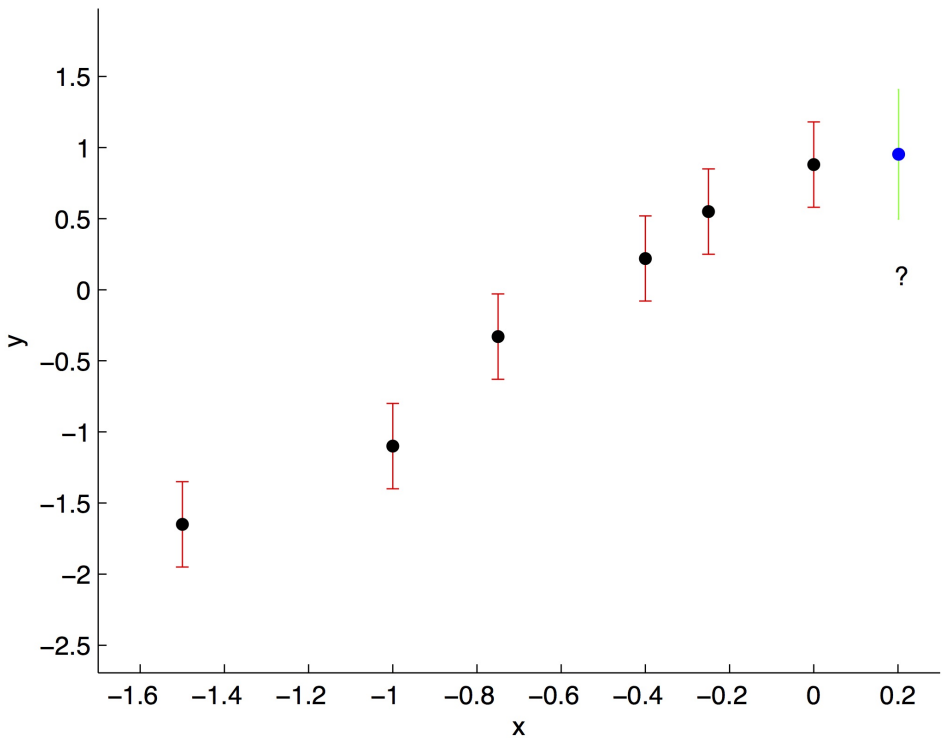
- D paramètres vitaux, N patients : $D \times N$ Gaussian Process à entraîner

Gaussian Process

- D paramètres vitaux, N patients : $D \times N$ Gaussian Process à entraîner
- $y = f(t) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$

Gaussian Process

- D paramètres vitaux, N patients : $D \times N$ Gaussian Process à entraîner
- $y = f(t) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$
- Training set : $\mathbf{X} = (t_i, y_i)_{i \in \{1,\dots,n\}},\ f(t_* = 0.2) = ?$



Gaussian Process

- D paramètres vitaux, N patients : $D \times N$ Gaussian Process à entraîner
- $y = f(t) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$
- Training set : $\mathbf{X} = (t_i, y_i)_{i \in \{1,\dots,n\}},\ f(t_* = 0.2) = ?$
- GP : distribution Gaussienne multivariée de dimension infinie

Gaussian Process

- D paramètres vitaux, N patients : $D \times N$ Gaussian Process à entraîner
- $y = f(t) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$
- Training set : $\mathbf{X} = (t_i, y_i)_{i \in \{1,\dots,n\}},\ f(t_* = 0.2) = ?$
- GP : distribution Gaussienne multivariée de dimension infinie
- $f(.) \sim \mathcal{GP}(m(.), k(.,.)) \Leftrightarrow \forall \mathbf{y} = (y_1, \dots, y_n), \mathbf{y} \sim \mathcal{N}(\mathbf{m}, \mathbf{K} + \sigma^2 \mathbf{I}_n)$

$$\mathbf{m} = \begin{bmatrix} m(t_1) \\ m(t_2) \\ \vdots \\ m(t_n) \end{bmatrix} \qquad \mathbf{K} = \begin{bmatrix} k(t_1,t_1) & k(t_1,t_2) & \dots & k(t_1,t_n) \\ k(t_2,t_1) & k(t_2,t_2) & \dots & k(t_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(t_n,t_1) & k(t_n,t_2) & \dots & k(t_n,t_n) \end{bmatrix}$$

Gaussian Process

- D paramètres vitaux, N patients : $D \times N$ Gaussian Process à entraîner
- $y = f(t) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$
- Training set : $\mathbf{X} = (t_i, y_i)_{i \in \{1,\dots,n\}},\ f(t_* = 0.2) = ?$
- GP : distribution Gaussienne multivariée de dimension infinie
- $f(.) \sim \mathcal{GP}(m(.), k(.,.)) \Leftrightarrow \forall \mathbf{y} = (y_1, \dots, y_n), \mathbf{y} \sim \mathcal{N}(\mathbf{m}, \mathbf{K} + \sigma^2 \mathbf{I}_n)$

$$\mathbf{m} = \begin{bmatrix} m(t_1) \\ m(t_2) \\ \vdots \\ m(t_n) \end{bmatrix} \qquad \mathbf{K} = \begin{bmatrix} k(t_1,t_1) & k(t_1,t_2) & \dots & k(t_1,t_n) \\ k(t_2,t_1) & k(t_2,t_2) & \dots & k(t_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(t_n,t_1) & k(t_n,t_2) & \dots & k(t_n,t_n) \end{bmatrix}$$

- $m(.) = 0$ sans perte de généralité

Gaussian Process

- En notant $\mathbf{K}_* = [k(t_*, t_1), \ldots, k(t_*, t_n)]$, $K_{**} = k(t_*, t_*)$

Gaussian Process

- En notant $\mathbf{K}_* = [k(t_*, t_1), \ldots, k(t_*, t_n)]$, $K_{**} = k(t_*, t_*)$
- On a

$$\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix} = \mathcal{N}(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I}_n & \mathbf{K}_*^\top \\ \mathbf{K}_* & K_{**} + \sigma^2 \end{bmatrix})$$

Gaussian Process

- En notant $\mathbf{K}_* = [k(t_*, t_1), \ldots, k(t_*, t_n)]$, $K_{**} = k(t_*, t_*)$
- On a

$$\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix} = \mathcal{N}(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I}_n & \mathbf{K}_*^\top \\ \mathbf{K}_* & K_{**} + \sigma^2 \end{bmatrix})$$

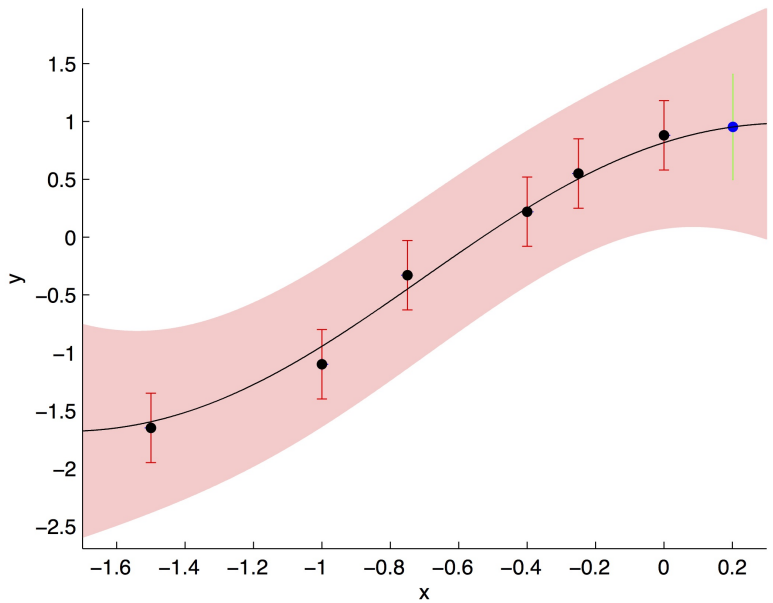
- Et $y_* | t_*, \mathbf{t}, \mathbf{y} \sim \mathcal{N}(\mathbf{K}_*(\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}\mathbf{y}, K_{**} + \sigma^2 - \mathbf{K}_*(\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}\mathbf{K}_*^\top)$

Gaussian Process

- En notant $\mathbf{K}_* = [k(t_*, t_1), \ldots, k(t_*, t_n)]$, $K_{**} = k(t_*, t_*)$
- On a

$$\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix} = \mathcal{N}(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I}_n & \mathbf{K}_*^\top \\ \mathbf{K}_* & K_{**} + \sigma^2 \end{bmatrix})$$

- Et $y_* | t_*, \mathbf{t}, \mathbf{y} \sim \mathcal{N}(\mathbf{K}_*(\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}\mathbf{y}, K_{**} + \sigma^2 - \mathbf{K}_*(\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}\mathbf{K}_*^\top)$
- Prédiction avec IC à 95% : $\overline{y}_* \pm 1.96\sqrt{\mathrm{Var}(y_*)}$



Design de k

- $k(t_i, t_j|\theta) = k_L(t_i, t_j|\sigma_L, \delta_L) + k_S(t_i, t_j|\sigma_S, \delta_S, P_L)$

Design de k

- $k(t_i, t_j|\theta) = k_L(t_i, t_j|\sigma_L, \delta_L) + k_S(t_i, t_j|\sigma_S, \delta_S, P_L)$
- Changements physiologiques d'un jour à l'autre :

$$k_L(t_i, t_j|\sigma_L, \delta_L) = \sigma_L^2 \exp\Big\{ - \frac{\|t_i - t_j\|_2^2}{2\delta_L^2} \Big\}$$

Design de k

- $k(t_i, t_j|\theta) = k_L(t_i, t_j|\sigma_L, \delta_L) + k_S(t_i, t_j|\sigma_S, \delta_S, P_L)$
- Changements physiologiques d'un jour à l'autre :

  $$k_L(t_i, t_j|\sigma_L, \delta_L) = \sigma_L^2 \exp\Big\{ - \frac{\|t_i - t_j\|_2^2}{2\delta_L^2} \Big\}$$

- Périodicité (rythme circadien : veille-sommeil, etc.) :

  $$k_S(t_i, t_j|\sigma_S, \delta_S, P_L) = \sigma_S^2 \exp\Big\{ - \frac{\|t_i - t_j\|_2^2}{2\delta_S^2} \Big\} \exp\Big\{ - \frac{\sin^2[(2\pi/P_L)\|t_i - t_j\|_2^2]}{2} \Big\}$$

Design de k

- $k(t_i, t_j|\theta) = k_L(t_i, t_j|\sigma_L, \delta_L) + k_S(t_i, t_j|\sigma_S, \delta_S, P_L)$
- Changements physiologiques d'un jour à l'autre :

  $$k_L(t_i, t_j|\sigma_L, \delta_L) = \sigma_L^2 \exp\Big\{ - \frac{\|t_i - t_j\|_2^2}{2\delta_L^2} \Big\}$$

- Périodicité (rythme circadien : veille-sommeil, etc.) :

  $$k_S(t_i, t_j|\sigma_S, \delta_S, P_L) = \sigma_S^2 \exp\Big\{ - \frac{\|t_i - t_j\|_2^2}{2\delta_S^2} \Big\} \exp\Big\{ - \frac{\sin^2[(2\pi/P_L)\|t_i - t_j\|_2^2]}{2} \Big\}$$

- Hyper-paramètre $\theta = (\sigma_L, \delta_L, \sigma_S, \delta_S, P_L)$

Estimation de θ et σ^2

- Marginal likelihood : $p(\mathbf{y}|\mathbf{t}, \theta, \sigma^2) = \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I}_n)$

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{t}, \theta, \sigma^2) &= \int p(\mathbf{y}|\mathbf{f}, \mathbf{t}, \theta, \sigma^2) p(\mathbf{f}|\mathbf{t}, \theta) \mathrm{d}\mathbf{f} \\
&= \int \mathcal{N}(\mathbf{f}, \sigma^2 \mathbf{I}_n) \mathcal{N}(\mathbf{0}, \mathbf{K}) \mathrm{d}\mathbf{f} \\
&= \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I}_n)
\end{aligned}
$$

Estimation de θ et σ^2

- Marginal likelihood : $p(\mathbf{y}|\mathbf{t}, \theta, \sigma^2) = \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I}_n)$
- Dans l'article, maximisation de la marginal log-lik (grid search)

$$\log p(\mathbf{y}|\mathbf{t}, \theta, \sigma^2) = -\frac{1}{2}\log|\mathbf{K} + \sigma^2 \mathbf{I}_n| - \frac{1}{2}\mathbf{y}(\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}\mathbf{y}^\top + \text{cste}$$

Estimation de θ et σ^2

- Marginal likelihood : $p(\mathbf{y}|\mathbf{t}, \theta, \sigma^2) = \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I}_n)$
- Dans l'article, maximisation de la marginal log-lik (grid search)

$$\log p(\mathbf{y}|\mathbf{t}, \theta, \sigma^2) = -\frac{1}{2}\log|\mathbf{K} + \sigma^2 \mathbf{I}_n| - \frac{1}{2}\mathbf{y}(\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}\mathbf{y}^\top + \text{cste}$$

- Alternative courante : MAP

$$p(\theta, \sigma^2|\mathbf{y}, \mathbf{t}) \propto p(\mathbf{y}|\mathbf{t}, \theta, \sigma^2)p(\theta, \sigma^2)$$

Estimation de θ et σ^2

- Marginal likelihood : $p(\mathbf{y}|\mathbf{t}, \theta, \sigma^2) = \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I}_n)$
- Dans l'article, maximisation de la marginal log-lik (grid search)

$$\log p(\mathbf{y}|\mathbf{t}, \theta, \sigma^2) = -\frac{1}{2}\log|\mathbf{K} + \sigma^2 \mathbf{I}_n| - \frac{1}{2}\mathbf{y}(\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}\mathbf{y}^\top + \text{cste}$$

- Alternative courante : MAP

$$p(\theta, \sigma^2|\mathbf{y}, \mathbf{t}) \propto p(\mathbf{y}|\mathbf{t}, \theta, \sigma^2)p(\theta, \sigma^2)$$

- Rque : (θ, σ^2) optimisé pour chaque patient et pour chaque paramètre vital ($D \times N$ fois)

Indice de similarité

- Pour le patient $k \in \{1, \ldots, N\}$, on a entraîné D gaussian process $\mathbf{X}_k$

Indice de similarité

- Pour le patient $k \in \{1, \ldots, N\}$, on a entraîné D gaussian process $\mathbf{X}_k$
- Si on a un nouveau patient $\mathbf{X}^*(\mathbf{t}, \mathbf{y})$ (D time series), on peut le comparer au patient k en calculant la log-vraissemblance négative locale au point i :

$$p(\mathbf{X}_i^*(\mathbf{t}_i, \mathbf{y}_i)|\mathbf{x}_k) = -\log \prod_{j=1}^{D} p(\mathbf{y}_i^j|\mathbf{t}_i, \mathbf{X}_k^j)$$

Indice de similarité

- Pour le patient $k \in \{1, \ldots, N\}$, on a entraîné D gaussian process $\mathbf{X}_k$
- Si on a un nouveau patient $\mathbf{X}^*(\mathbf{t}, \mathbf{y})$ (D time series), on peut le comparer au patient k en calculant la log-vraissemblance négative locale au point i :

$$p(\mathbf{X}_i^*(\mathbf{t}_i, \mathbf{y}_i)|\mathbf{x}_k) = -\log \prod_{j=1}^{D} p(\mathbf{y}_i^j|\mathbf{t}_i, \mathbf{X}_k^j)$$

- Vraissemblance globale (pour tous les points) :

$$\mathcal{L}_k(\mathbf{X}^*) = n^{-1} \sum_i [p(\mathbf{X}_i^*(\mathbf{t}_i, \mathbf{y}_i)|\mathbf{x}_k)]$$

Indice de similarité

- Pour le patient $k \in \{1, \ldots, N\}$, on a entraîné D gaussian process $\mathbf{X}_k$
- Si on a un nouveau patient $\mathbf{X}^*(\mathbf{t}, \mathbf{y})$ (D time series), on peut le comparer au patient k en calculant la log-vraissemblance négative locale au point i :

$$p(\mathbf{X}_i^*(\mathbf{t}_i, \mathbf{y}_i)|\mathbf{x}_k) = -\log \prod_{j=1}^{D} p(\mathbf{y}_i^j|\mathbf{t}_i, \mathbf{X}_k^j)$$

- Vraissemblance globale (pour tous les points) :

$$\mathcal{L}_k(\mathbf{X}^*) = n^{-1} \sum_i [p(\mathbf{X}_i^*(\mathbf{t}_i, \mathbf{y}_i)|\mathbf{x}_k)]$$

- Indice de similarité : $S_k(\mathbf{X}^*) = \mathcal{L}_*/\mathcal{L}_k(\mathbf{X}^*)$,

$$\text{avec } \mathcal{L}_* = n^{-1} \sum_i [p(\mathbf{X}_i^*(\mathbf{t}_i, \mathbf{y}_i)|\mathbf{x}^*)] \leq \mathcal{L}_k(\mathbf{X}^*)$$

Indice de similarité

- Pour le patient $k \in \{1, \ldots, N\}$, on a entraîné D gaussian process $\mathbf{X}_k$
- Si on a un nouveau patient $\mathbf{X}^*(\mathbf{t}, \mathbf{y})$ (D time series), on peut le comparer au patient k en calculant la log-vraissemblance négative locale au point i :

$$p(\mathbf{X}_i^*(\mathbf{t}_i, \mathbf{y}_i)|\mathbf{x}_k) = -\log \prod_{j=1}^{D} p(\mathbf{y}_i^j|\mathbf{t}_i, \mathbf{X}_k^j)$$

- Vraissemblance globale (pour tous les points) :

$$\mathcal{L}_k(\mathbf{X}^*) = n^{-1} \sum_i [p(\mathbf{X}_i^*(\mathbf{t}_i, \mathbf{y}_i)|\mathbf{x}_k)]$$

- Indice de similarité : $S_k(\mathbf{X}^*) = \mathcal{L}_* / \mathcal{L}_k(\mathbf{X}^*)$,

$$\text{avec } \mathcal{L}_* = n^{-1} \sum_i [p(\mathbf{X}_i^*(\mathbf{t}_i, \mathbf{y}_i)|\mathbf{x}^*)] \leq \mathcal{L}_k(\mathbf{X}^*)$$

- Si $\mathbf{X}^*$ similaire à $\mathbf{X}_k$, alors $S_k(\mathbf{X}^*) \to 1$, et $S_k(\mathbf{X}^*) \to 0$ sinon

Clustering hiérarchique

- Matrice ce similarité $S \in [0,1]^{n \times n}$

Clustering hiérarchique

- Matrice ce similarité $S \in [0,1]^{n \times n}$
- On construit un arbre de clustering et on choisit un seuil pour former les clusters

Clustering hiérarchique

- Matrice ce similarité $S \in [0,1]^{n \times n}$
- On construit un arbre de clustering et on choisit un seuil pour former les clusters
- Quand vient un nouvel individu (D time series), on calcule sa similarité avec chacun des clusters, et on peut le classifier comme ”anormal” si le max est inférieur à un certain seuil.

Résultats

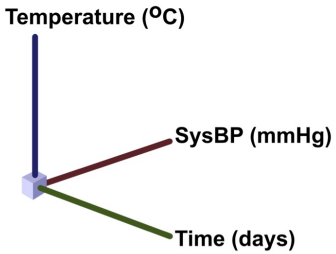
- Séjours de patients atteints du cancer après une opération

Résultats

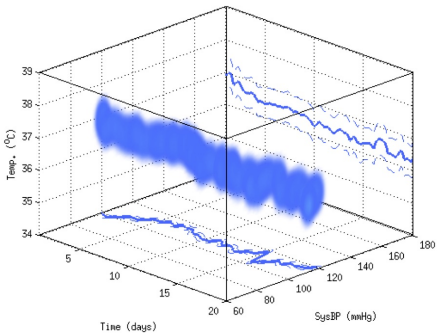
- Séjours de patients atteints du cancer après une opération
- N=100 patients ”normaux”, D=2 : PA systolique et température

Résultats

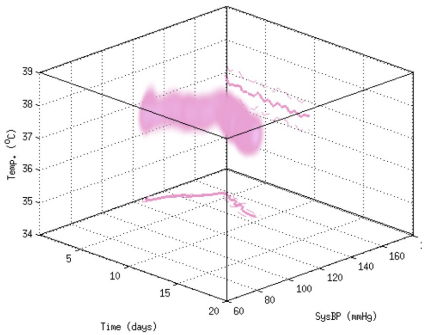
- Séjours de patients atteints du cancer après une opération
- N=100 patients ”normaux”, D=2 : PA systolique et température
- 4 clusters, moyennes des fonctions moyennes apprises :



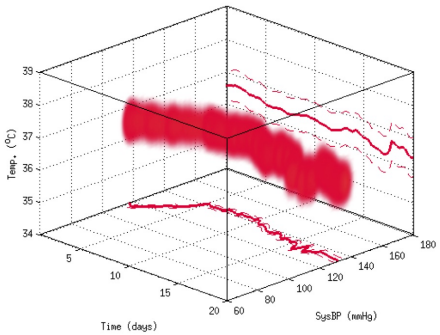
(a) Axis labels



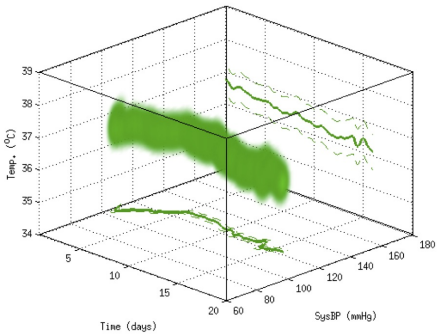
(b) Prototype A



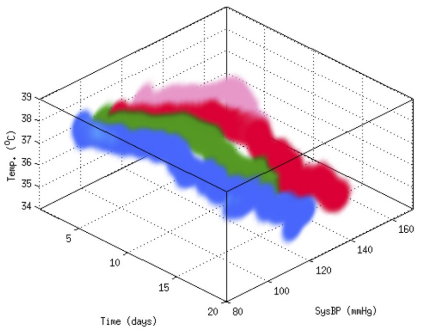
(c) Prototype B



(d) Prototype C



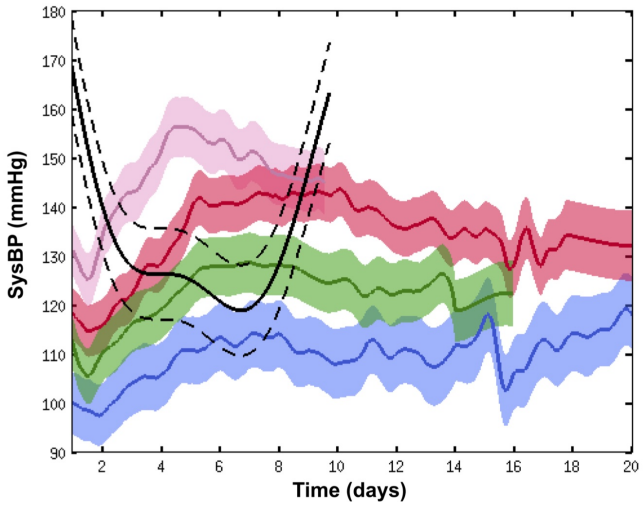
(e) Prototype D



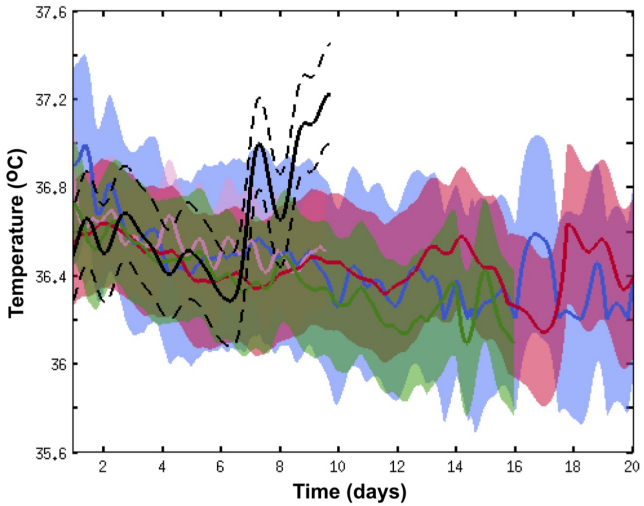
(f) All prototypes

Résultats

- En noir, un individu ”anormal” (soins intensifs après 9 jours)



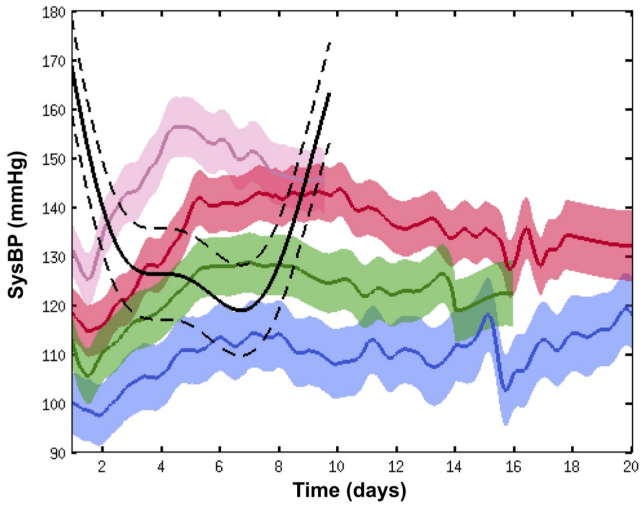
(g)



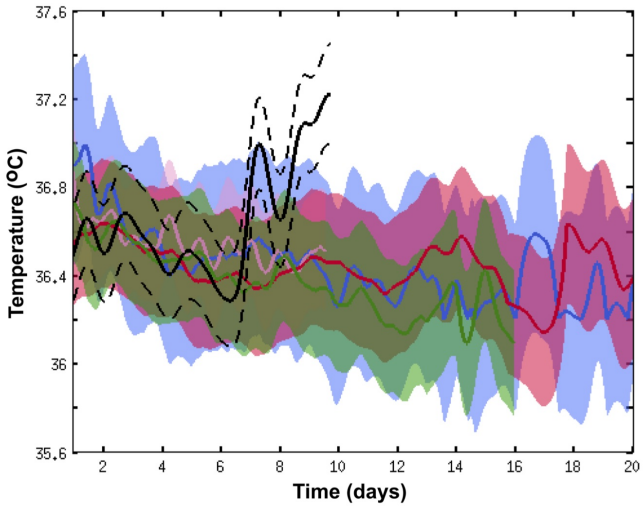
(h)

Résultats

- En noir, un individu ”anormal” (soins intensifs après 9 jours)



(g)



(h)

- Influence des autres paramètres vitaux (D=5)

Cluster	$D = 2$ (*)	$D = 5$ (**)	Overlap
Prototype A	12	14	8
Prototype B	20	19	12
Prototype C	34	35	26
Prototype D	34	35	24

Conclusion/Questions

- Rien n'empêche de faire dépendre n de (j, k) ?

Conclusion/Questions

- Rien n'empêche de faire dépendre n de (j, k) ?
- Utiliser ce procédé sur les données brutes de l'HEGP, puis prendre θ comme features et superviser par nos temps de retour T^c avec QNEM par ex ?