Présentation des premiers travaux de thèse :

A high dimensional mixture model for time-to-event data with application to risk assessment of sickle-cell anemia.

Simon Bussy

encadré par
Agathe Guilloux, **agathe.guilloux@upmc.fr**
Stéphane Gaiffas, **stephane.gaiffas@cmap.polytechnique.fr**
Anne Sophie Jannot **annesophie.jannot@aphp.fr**

Présentation GTT

Le 27 Janvier 2016

Introduction

- Stage de M2 sur des données de santé ; patients atteints de drépanocytose.

Introduction

- Stage de M2 sur des données de santé ; patients atteints de drépanocytose.
- HEGP : centre de référence pour le traitement des CVO.

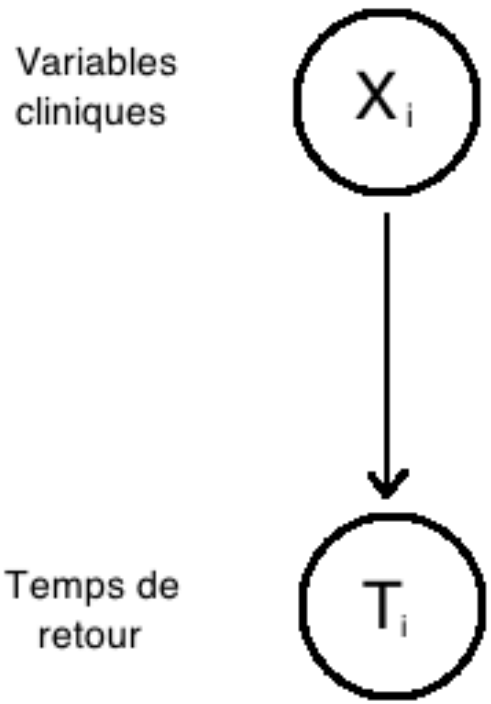
Introduction

- Stage de M2 sur des données de santé ; patients atteints de drépanocytose.
- HEGP : centre de référence pour le traitement des CVO.
- Taux de réhospitalisation élevé ($\sim$ 20%).

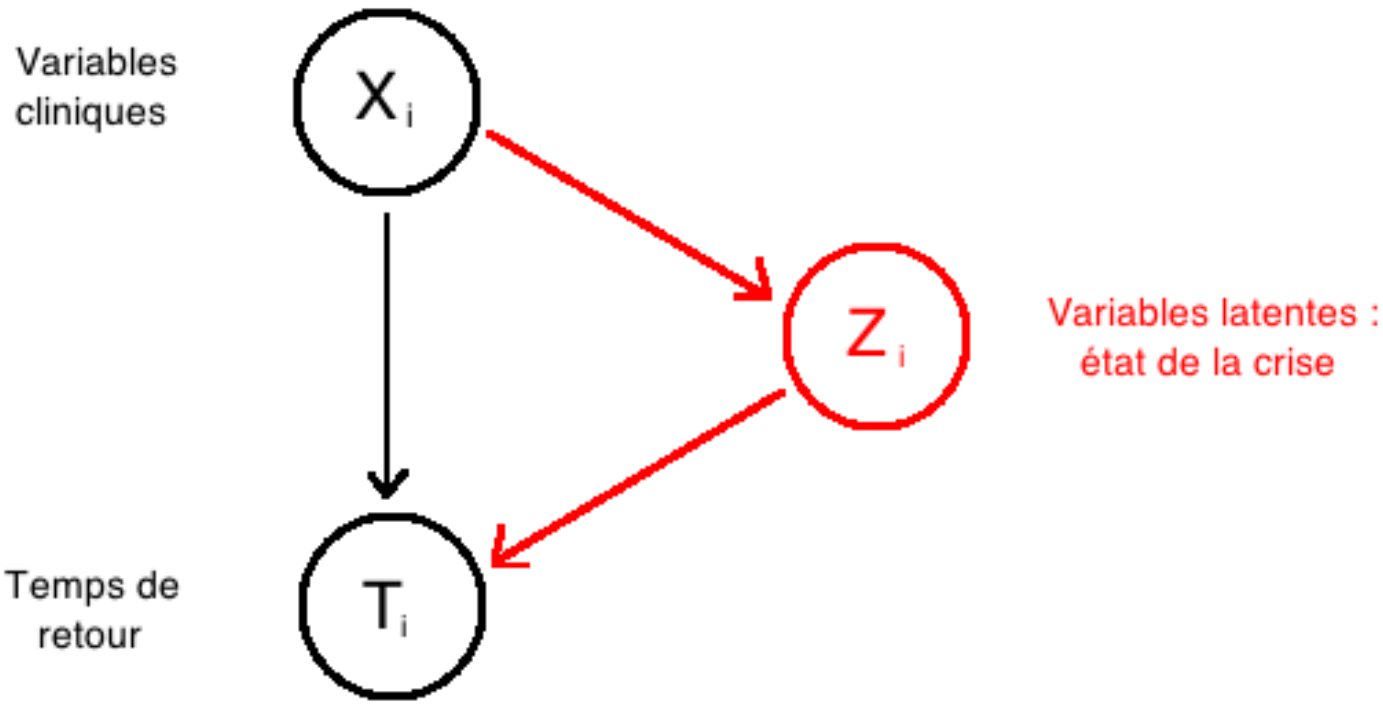
Introduction

- Stage de M2 sur des données de santé ; patients atteints de drépanocytose.
- HEGP : centre de référence pour le traitement des CVO.
- Taux de réhospitalisation élevé ($\sim$ 20%).
- But : construire un modèle prédictif des rechutes.

Modélisation



Modélisation



Modèle de mélange

- $Z_i \in \{0,1\}$ *v.a.* latente t.q. $\forall i \in [\![1,n]\!]$, $\forall k \in \{0,1\}$, $T_i | Z_i = k \sim \ell_k$

Modèle de mélange

- $Z_i \in \{0,1\}$ *v.a.* latente t.q. $\forall i \in [\![1,n]\!]$, $\forall k \in \{0,1\}$, $T_i|Z_i = k \sim \ell_k$
- Quantité primordiale : $\pi_0(X_i) = \mathbb{P}(Z_i = 0|X_i) = \frac{1}{1+e^{-X_i^\top \beta}}$

Modèle de mélange

- $Z_i \in \{0,1\}$ *v.a.* latente t.q. $\forall i \in [\![1,n]\!]$, $\forall k \in \{0,1\}$, $T_i|Z_i = k \sim \ell_k$
- Quantité primordiale : $\pi_0(X_i) = \mathbb{P}(Z_i = 0|X_i) = \frac{1}{1+e^{-X_i^\top \beta}}$
- $\forall i \in [\![1,n]\!], Z_i \sim \mathcal{B}(1-\pi_0(X_i))$

Modèle de mélange

- $Z_i \in \{0,1\}$ *v.a.* latente t.q. $\forall i \in [\![1,n]\!]$, $\forall k \in \{0,1\}$, $T_i|Z_i = k \sim \ell_k$
- Quantité primordiale : $\pi_0(X_i) = \mathbb{P}(Z_i = 0|X_i) = \frac{1}{1+e^{-X_i^\top \beta}}$
- $\forall i \in [\![1,n]\!], Z_i \sim \mathcal{B}(1-\pi_0(X_i))$
- $\forall i \in [\![1,n]\!]$, $\forall t \in \mathbb{N}^*, f_{T_i|X_i}(t) = \pi_0(X_i) f_{T_i|X_i,0}(t) + (1-\pi_0(X_i)) f_{T_i|X_i,1}(t)$

Mélange censuré

- Censure : $T^c = T \wedge C$ et $\delta = \mathbb{1}_{\{T \leq C\}}$

Mélange censuré

- Censure : $T^c = T \wedge C$ et $\delta = \mathbb{1}_{\{T \leq C\}}$
- Echantillon de n patients : $(X_1, T_1^c, \delta_1), \ldots, (X_n, T_n^c, \delta_n) \in \mathbb{R}^d \times \mathbb{N}^* \times \{0; 1\}$

Mélange censuré

- Censure : $T^c = T \wedge C$ et $\delta = \mathbb{1}_{\{T \leq C\}}$
- Echantillon de n patients : $(X_1, T_1^c, \delta_1), \ldots, (X_n, T_n^c, \delta_n) \in \mathbb{R}^d \times \mathbb{N}^* \times \{0; 1\}$
- Alors en notant $\boldsymbol{T}^c = [T_1^c, \cdots, T_n^c]$, $\boldsymbol{\Delta} = [\delta_1, \ldots, \delta_n]$ et $\overline{F} = 1 - F$, la log-vraissemblance du modèle s'écrit

$$\ell_n(\theta\,;\boldsymbol{T}^c, \boldsymbol{\Delta}) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\{\pi_{0,\beta}(X_i) f_{T,0}(T_i^c;\alpha_0) + (1-\pi_{0,\beta}(X_i)) f_{T,1}(T_i^c;\alpha_1)\}\overline{G}(T_i^{c-})\right]^{\delta_i}$$
$$\times \left[\{\pi_{0,\beta}(X_i)\overline{F}_{T,0}(T_i^{c-};\alpha_0) + (1-\pi_{0,\beta}(X_i))\overline{F}_{T,1}(T_i^{c-};\alpha_1)\} g(T_i^c)\right]^{1-\delta_i}$$

- But : estimer $(\alpha_0, \alpha_1, \beta)$

Inférence pour le modèle

- Minimiser $-\ell_n(\theta) + \gamma\left((1-\eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2\right)$

Inférence pour le modèle

- Minimiser $-\ell_n(\theta) + \gamma\left((1-\eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2\right)$
- Algorithme qui s'inspire d'un EM.

Inférence pour le modèle

- Minimiser $-\ell_n(\theta) + \gamma\left((1-\eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2\right)$
- Algorithme qui s'inspire d'un EM.
- Au pas l, on calcule $Q_n(\theta, \theta^{(l)}) = \mathbb{E}_{\theta^{(l)}}[\ell_n^{comp}(\theta; \boldsymbol{T}^c, \boldsymbol{\Delta}, \boldsymbol{Z}) | \boldsymbol{T}^c, \boldsymbol{\Delta}]$

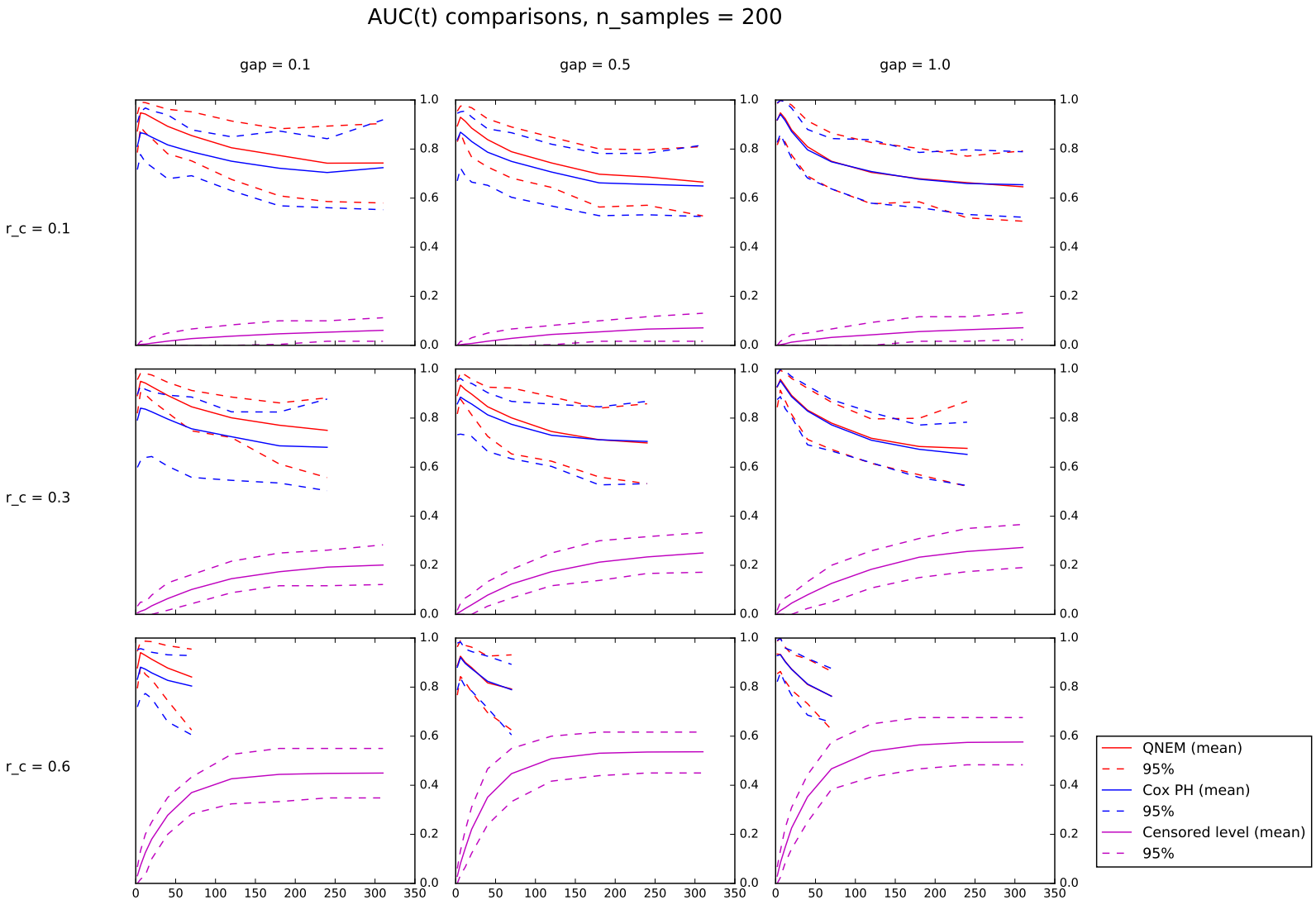
Inférence pour le modèle

- Minimiser $-\ell_n(\theta) + \gamma\left((1-\eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2\right)$
- Algorithme qui s'inspire d'un EM.
- Au pas l, on calcule $Q_n(\theta, \theta^{(l)}) = \mathbb{E}_{\theta^{(l)}}[\ell_n^{comp}(\theta; \boldsymbol{T}^c, \boldsymbol{\Delta}, \boldsymbol{Z})|\boldsymbol{T}^c, \boldsymbol{\Delta}]$
- On maximise la quantité précédente pour mettre à jour les paramètres.

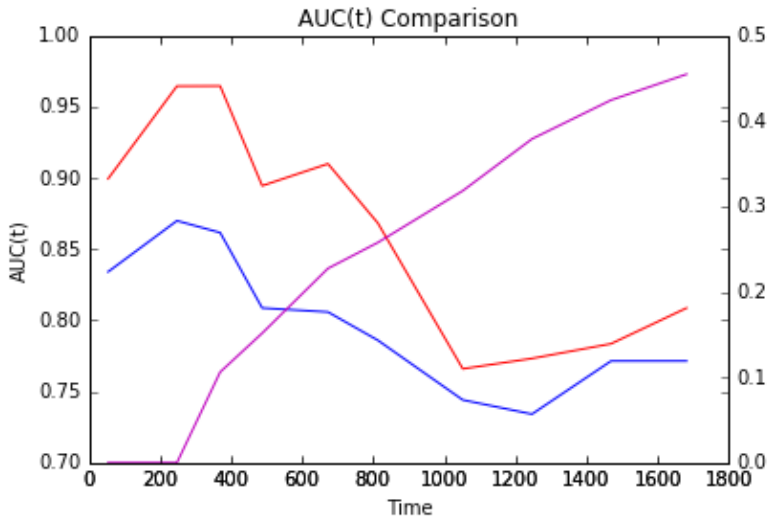
Inférence pour le modèle

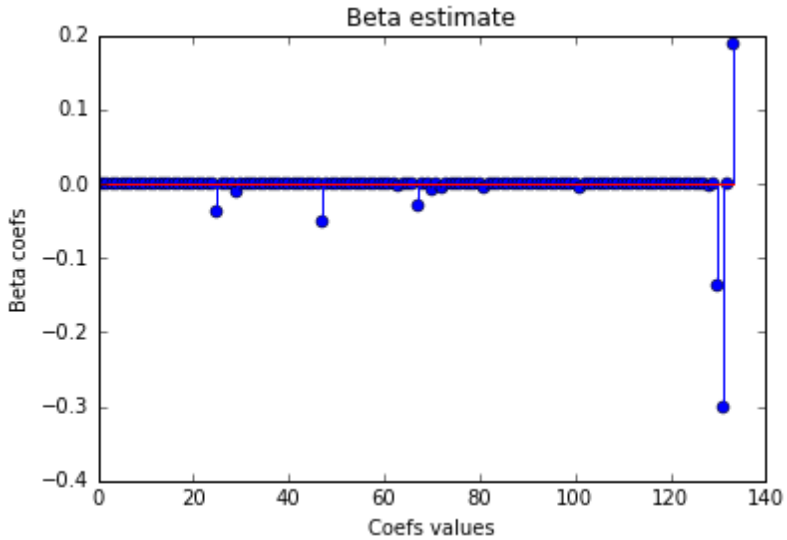
- Minimiser $-\ell_n(\theta) + \gamma\left((1-\eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2\right)$
- Algorithme qui s'inspire d'un EM.
- Au pas l, on calcule $Q_n(\theta, \theta^{(l)}) = \mathbb{E}_{\theta^{(l)}}[\ell_n^{comp}(\theta; \boldsymbol{T}^c, \boldsymbol{\Delta}, \boldsymbol{Z})|\boldsymbol{T}^c, \boldsymbol{\Delta}]$
- On maximise la quantité précédente pour mettre à jour les paramètres.
- On utilise l'algorithme L-BGFS-B pour la mise à jour de β.

Performance en simulation



Performance sur données réelles





Conclusion

- Objectif : publication pour présenter le modèle, puis publications cliniques applicatives.

Conclusion

- Objectif : publication pour présenter le modèle, puis publications cliniques applicatives.
- Pour la suite, travail sur une BDD beaucoup plus imposante, avec des données longitudinales et un suivi patient sur + de 10 ans.

Conclusion

- Objectif : publication pour présenter le modèle, puis publications cliniques applicatives.
- Pour la suite, travail sur une BDD beaucoup plus imposante, avec des données longitudinales et un suivi patient sur + de 10 ans.
- Apprentissage et modélisation dans le cadre de données fonctionnelles.