

Réunion sur le projet RedCVO

Nouvelles techniques de ML pour la médecine :

**FACTEURS PRÉDICTIFS DE LA
RÉHOSPITALISATION PRÉCOCE DE
DRÉPANOCYTAIRES ADULTES POUR CRISE
VASO-OCCLUSIVE**

Simon Bussy

encadré par

Anne-Sophie Jannot, Stéphane Gaiffas, Agathe Guilloux

Présentation Brigitte Ranque

5 février 2016

Introduction

- Stage de 5 mois : CMAP et équipe INSERM n° 22 de l'HEGP.

Introduction

- Stage de 5 mois : CMAP et équipe INSERM n° 22 de l'HEGP.
- Doctorat de Mathématiques au LSTA.

Introduction

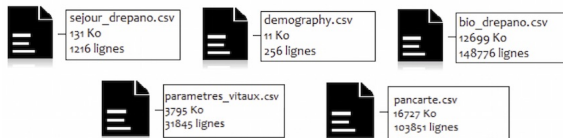
- Stage de 5 mois : CMAP et équipe INSERM n° 22 de l'HEGP.
- Doctorat de Mathématiques au LSTA.
- Travail de modélisation des données pendant le stage.

Introduction

- Stage de 5 mois : CMAP et équipe INSERM n° 22 de l'HEGP.
- Doctorat de Mathématiques au LSTA.
- Travail de modélisation des données pendant le stage.
- Création d'un nouveau modèle d'apprentissage.

Organisation des données

- Données hétérogènes, différentes sources

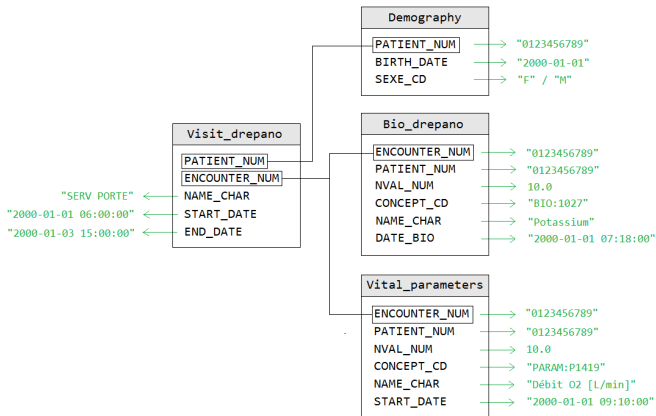


Organisation des données

- Données hétérogènes, différentes sources
- 1 séjour / patient : "choix aléatoire par classe", taux de rechute de 12,84%

Organisation des données

● Réorganisation des données



Organisation des données

● Création d'un fichier JSON

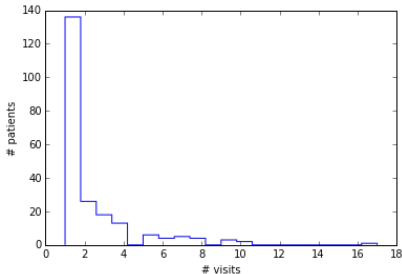
```

Patients = {
  '1' = {
    'patient_num' = '0123456789',
    'sex' = 'F/M',
    'birth_date' = '2000-01-01',
    'visits' = {
      '1' = {
        'encounter_num' = '0123456789',
        'age' = '25',
        'rea' = '0/1',
        'duration' = '3',
        'previous_visit' = 'none/10',
        'next_visit' = 'none/15',
        'details' = {
          '1' = {
            'service' = 'SERV PORTE',
            'start_date' = '2000-01-01 09:00:00',
            'end_date' = '2000-05-01 10:00:00',
            '2' = {...}
          },
          '2' = {...}
        },
        'bio-infos' = {
          '1' = {
            'name_char' = 'Potassium',
            'val' = {
              '1' = {
                'nval_num' = '10.0',
                'concept_cd' = 'BIO:1027',
                'date_bio' = '2000-01-01 11:00:00',
                '2' = {...}
              },
              '2' = {...}
            }
          },
          '2' = {...}
        },
        'vital_parameters' = {
          '1' = {
            'name_char' = 'Débit O2 [L/min]',
            'val' = {
              '1' = {
                'nval_num' = '10.0',
                'concept_cd' = 'PARAM:P1419',
                'date_bio' = '2000-01-01 10:30:00',
                '2' = {...}
              },
              '2' = {...}
            }
          },
          '2' = {...}
        }
      },
      '2' = {...}
    },
    '2' = {...}
  },
  '2' = {...}
}

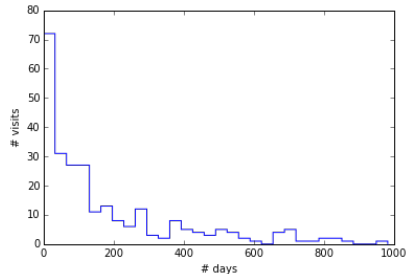
```

Présentation des données

● 218 patients, 479 visites



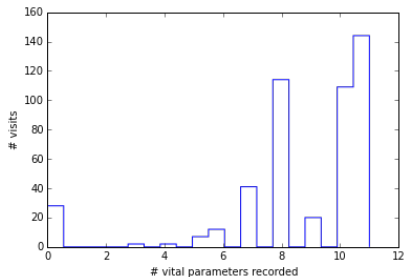
Nombre de visites par patient



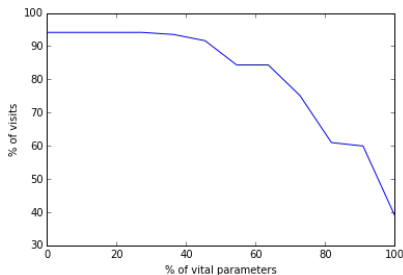
Temps inter-visites

Les paramètres vitaux

● Les paramètres vitaux



Nombre de points
par paramètre vital



Variables biologiques en commun
à tous les séjours

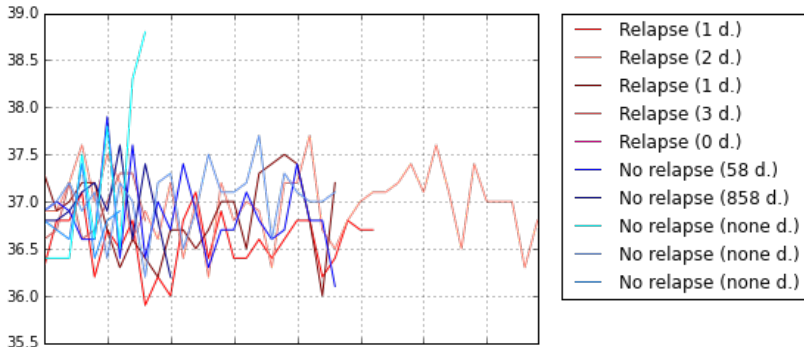
Les paramètres vitaux

- Les paramètres vitaux
- 9 variables gardées

Pression artérielle systolique [mmHg], notée PA max ;
pression artérielle diastolique [mmHg], notée PA min ;
saturation en oxygène [%] ; Douleur EVA [U] ; fréquence
respiratoire [mvt/min] ; température [°C] ; poids [kg] ;
oxygène [L/min] ; fréquence cardiaque [bpm]

Les paramètres vitaux

- Les paramètres vitaux
- 9 variables gardées
- Problème : durées $\neq$ et alignement



Evolution de la température au cours de 10 séjours

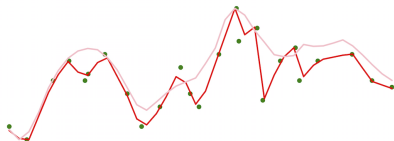
Les paramètres vitaux

- Rescaling entre $[0,1]$

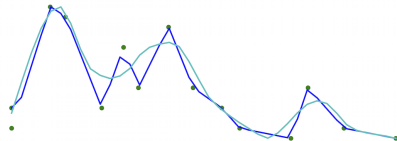
$$\forall j \in \llbracket 1, N \rrbracket, x_j = \frac{x_j - \min_k(x_k)}{\max_k(x_k) - \min_k(x_k)}$$

Les paramètres vitaux

- Rescaling entre $[0,1]$
- Interpolation : p valeurs
Puis filtre de Savitzky–Golay : p' valeurs de gradient



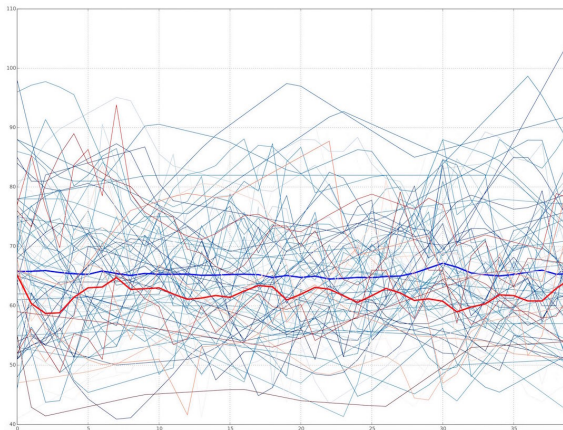
Fréquence respiratoire pour
un patient positif



Fréquence respiratoire pour
un patient négatif

Les paramètres vitaux

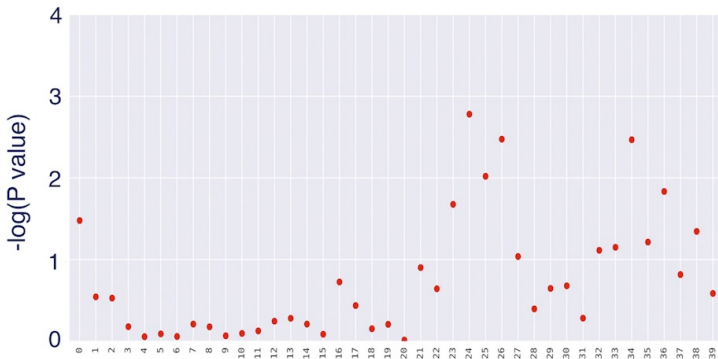
- Comportements différents en moyenne



Pression artérielle diastolique et moyennes par classe

Les paramètres vitaux

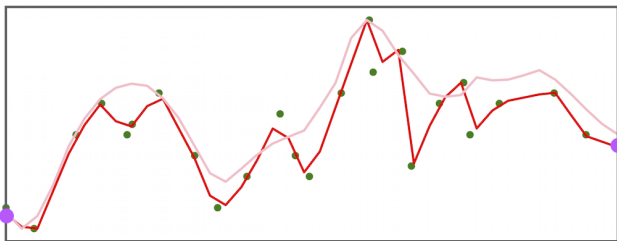
- Comportements différents en moyenne
- Test de Mann-Whitney-Wilcoxon



Test de Mann-Whitney-Wilcoxon pour la température

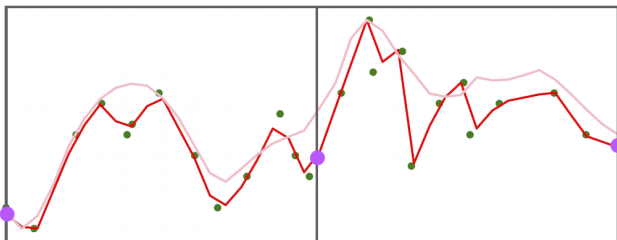
Sélection de p et p'

Procédure suivie :



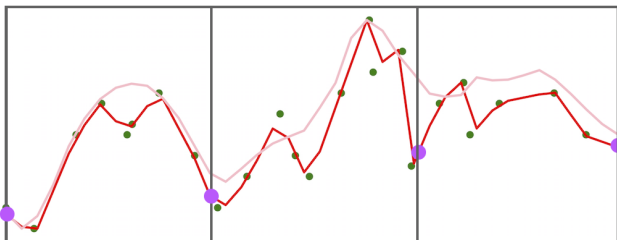
Sélection de p et p'

Procédure suivie :



Sélection de p et p'

Procédure suivie :



Sélection de p et p'

● Résultats après 50 itérations

Concept	Nombre de points	Ecart type
Fq cardiaque	28	0,3
▽ Fq cardiaque	26	0
PA max	28	0
▽ PA max	26	0
Température	30	8,4
▽ Température	18	0
Saturation O_2	34	4,8
▽ Saturation O_2	26	0
Douleur EVA	21	0,7
▽ Douleur EVA	26	0
Débit O_2	16	4,6
▽ Débit O_2	26	0
Fq respiratoire	21	5,3
▽ Fq respiratoire	26	0
PA min	28	0
▽ PA min	26	0

Sélection de p et p'

- Résultats après 50 itérations
- Dimension de l'espace : 446

Sélection de p et p'

- Résultats après 50 itérations
- Dimension de l'espace : 446
- Données manquantes remplacées par les moyennes

Benjamini-Hochberg

- Tests de Mann-Whitney-Wilcoxon : 446 *p-values*

Benjamini-Hochberg

- Tests de Mann-Whitney-Wilcoxon : 446 *p-values*
- Procédure de Benjamini-Hochberg

On conserve k features avec

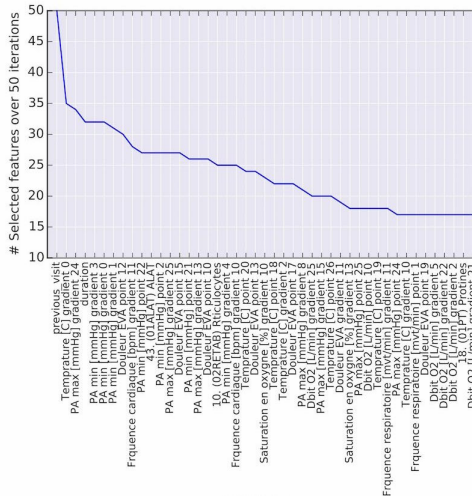
$$k = \operatorname{argmax} \left\{ j : p(j) \leq \frac{\alpha j}{K} \right\}$$

Benjamini-Hochberg

- Tests de Mann-Whitney-Wilcoxon : 446 *p-values*
- Procédure de Benjamini-Hochberg
- $\alpha = 0,9$; 50 itérations

Benjamini-Hochberg

- 45 features sélectionnées ($> \frac{1}{3}$ des cas)



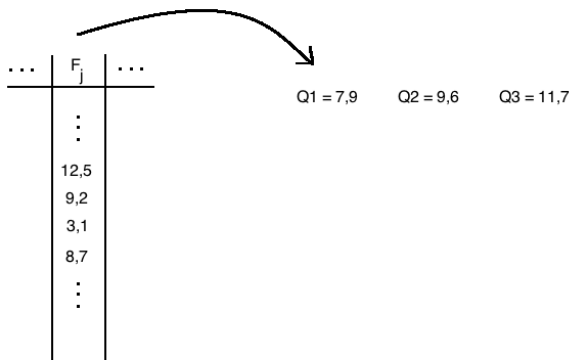
Binarisation

- Binarisation par quartiles ; $dim = 156 \sim 4 \times 45$

...	F_j	...
	:	
	:	
	12,5	
	9,2	
	3,1	
	8,7	
	:	
	:	

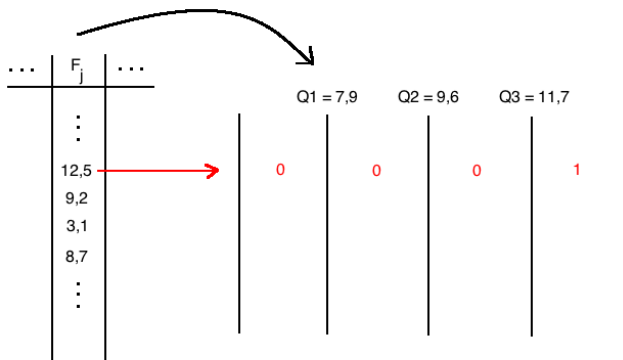
Binarisation

- Binarisation par quartiles ; $dim = 156 \sim 4 \times 45$



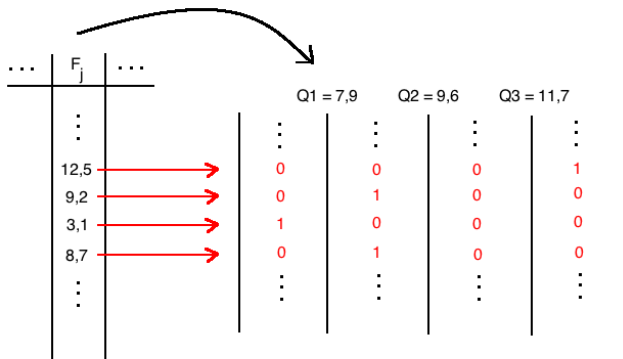
Binarisation

- Binarisation par quartiles ; $dim = 156 \sim 4 \times 45$



Binarisation

- Binarisation par quartiles ; $dim = 156 \sim 4 \times 45$



Association + décorrélation

- Ajout des produits cartésiens

Association + décorrélation

- Ajout des produits cartésiens
- $\dim = 11978 \sim \binom{156}{2}$

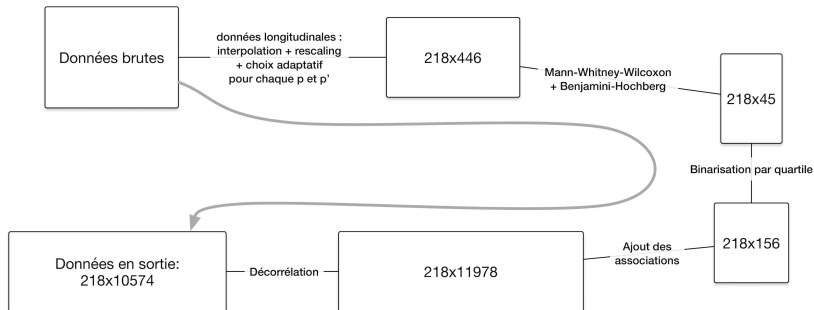
Association + décorrélation

- Ajout des produits cartésiens
- $\dim = 11978 \sim \binom{156}{2}$
- Décorrélation $|\rho| < 0,95$

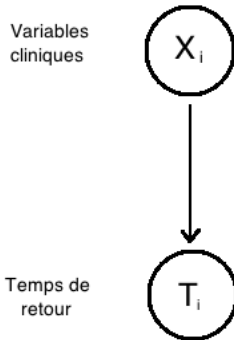
Association + décorrélation

- Ajout des produits cartésiens
- $\dim = 11978 \sim \binom{156}{2}$
- Décorrélation $|\rho| < 0,95$
- Finalement, $\dim = 10574$

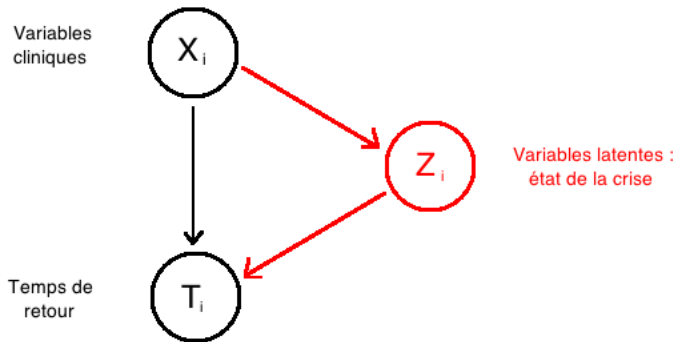
Schéma récapitulatif



Modélisation



Modélisation



Modèle de mélange

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

Modèle de mélange

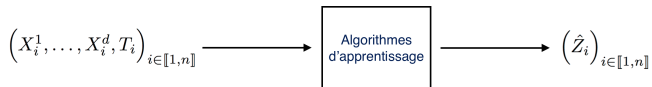
- $Z_i \in \{0, 1\}$ v.a. latente t.q. $\forall i \in \llbracket 1, n \rrbracket, \forall k \in \{0, 1\}, T_i | Z_i = k \sim \ell_k$
- Quantité primordiale : $\pi_0(X_i) = \mathbb{P}(Z_i = 0 | X_i) = \frac{1}{1 + e^{-X_i^\top \beta}}$

Modèle de mélange

- $Z_i \in \{0, 1\}$ v.a. latente t.q. $\forall i \in \llbracket 1, n \rrbracket, \forall k \in \{0, 1\}, T_i | Z_i = k \sim \ell_k$
- Quantité primordiale : $\pi_0(X_i) = \mathbb{P}(Z_i = 0 | X_i) = \frac{1}{1 + e^{-X_i^\top \beta}}$
- $\forall i \in \llbracket 1, n \rrbracket, \forall t \in \mathbb{N}^*, f_{T_i | X_i}(t) = \pi_0(X_i) f_{T_i | X_i, 0}(t) + (1 - \pi_0(X_i)) f_{T_i | X_i, 1}(t)$
- Paramètres à estimer : $\theta = \{p_0, p_1, \beta\} \in \mathbb{R}^{d+3}$

Simulations

- Idée : tester le schéma de prédiction suivant :



Simulations

- Idée : tester le schéma de prédiction suivant :



- Pour la régression logistique : $Z_i \sim Y_i = \mathbb{1}_{\{T_i \leq s\}}$, AUC sur les $\hat{Y}_i$

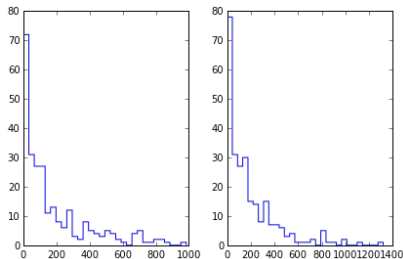
Simulations

- Idée : tester le schéma de prédiction suivant :



- Pour la régression logistique : $Z_i \sim Y_i = \mathbb{1}_{\{T_i \leq s\}}$, AUC sur les $\hat{Y}_i$
- Pour l'EM : AUC sur les $\hat{Z}_i$

Estimation des vrais paramètres

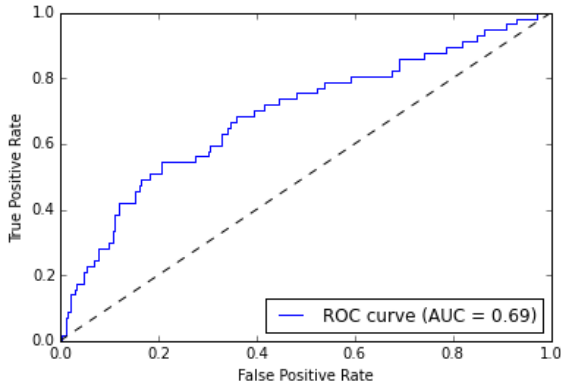


Comparaison des histogrammes de temps de retour réels et simulés après estimation

Paramètres	p_0	p_1	π
$\hat{\theta}$	0,005	0,511	0,872

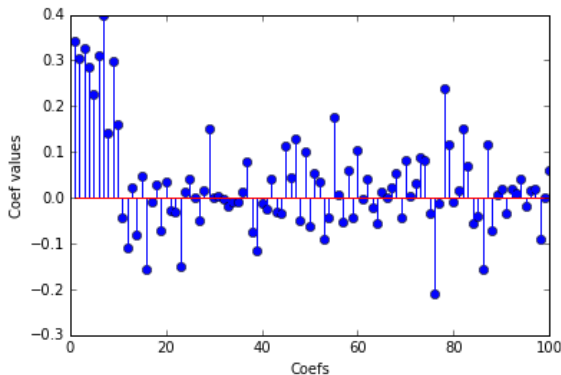
Résultats de prédiction sur les vraies données

Résultats : régression logistique



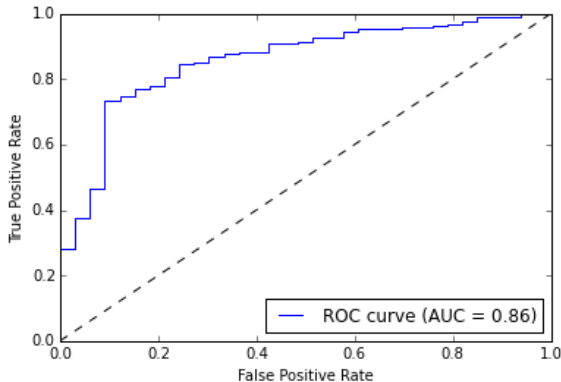
Courbe ROC pour la prédiction de la régression logistique

Résultats : régression logistique



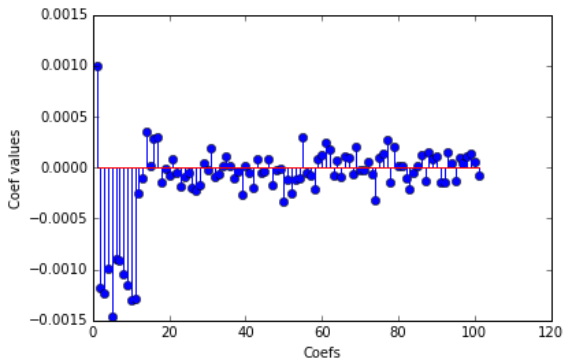
$\hat{\beta}$ appris par la régression logistique

Résultats : EM



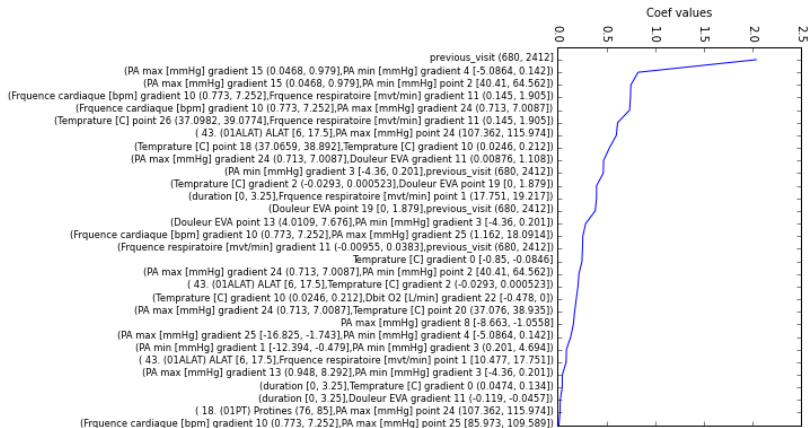
Courbe ROC pour la prédiction de l'EM

Résultats : EM



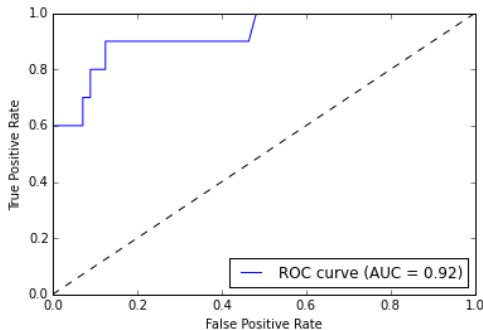
$\hat{\beta}$ appris par l'EM

Résultats : Support du lasso



Coefficients de $|\hat{\beta}_{lasso}|$ non nuls triés

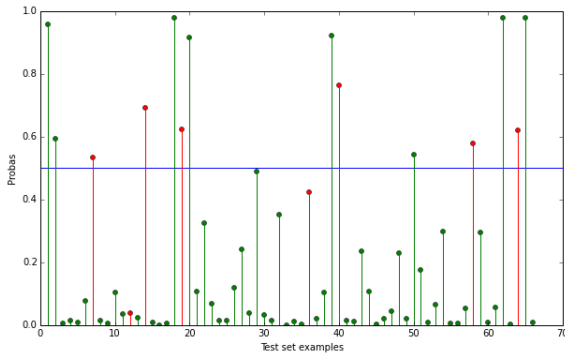
Résultats : régression ridge (dim = 30)



Courbe ROC pour la régression ℓ_2 dans l'espace de dimension 30

Classe	Précision	Recall	Support
0	0,96	0,89	56
1	0,57	0,80	10
Moyenne/Total	0,90	0,88	66

Résultats : régression ridge (dim = 30)



Probabilités prédites sur le jeu de test

Mélange censuré

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

Mélange censuré

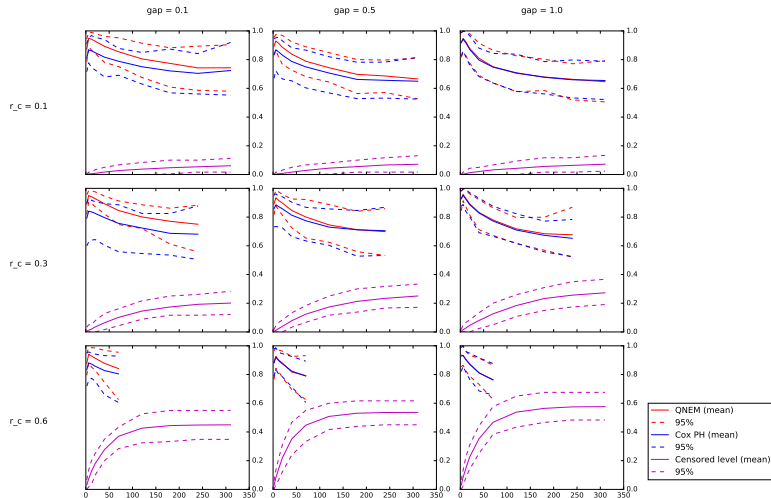
- Censure : $T^c = T \wedge C$ et $\delta = \mathbb{1}_{\{T \leq C\}}$
- Minimiser $-\ell_n(\theta) + \gamma((1 - \eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2)$

Mélange censuré

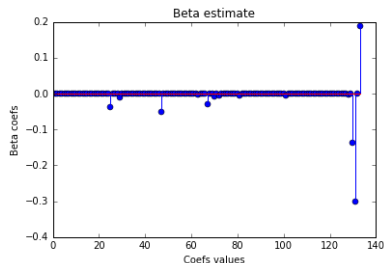
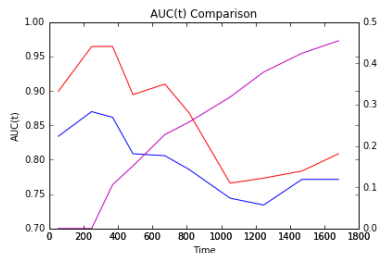
- Censure : $T^c = T \wedge C$ et $\delta = \mathbb{1}_{\{T \leq C\}}$
- Minimiser $-\ell_n(\theta) + \gamma((1 - \eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2)$
- Algorithme qui s'inspire d'un EM mais qui diffère sur plusieurs points, notamment avec la présence d'une régularisation.

Performance en simulation

AUC(t) comparisons, $n_{\text{samples}} = 200$



Performance sur données réelles



Conclusion

- Objectif : publication pour présenter le modèle mathématique.

Conclusion

- Objectif : publication pour présenter le modèle mathématique.
- Puis publication clinique applicative sur les données de RedCVO.