Prix de thèse Daniel Schwartz

présentée le 02/10/2020 par

Simon Bussy [1]

https://simonbussy.com/

Introduction de modèles de machine learning interprétables en grande dimension et leurs applications

sous la direction de

Pr. Stéphane Gaïffas [1,2]
Pr. Agathe Guilloux [3]
Dr. Anne-Sophie Jannot [4,5]

[1] LPSM, UMR 8001, Sorbonne Université, Paris, France. [2] CMAP, UMR 7641, École Polytechnique CNRS, Paris, France. [3] LAMME, Université Evry, CNRS, Université Paris-Saclay, Paris, France. [4] APHP, Département d'Informatique Biomédicale et de Santé Publique, HEGP, Paris, France. [5] INSERM, UMRS 1138, Eq22, Centre de Recherche des Cordeliers, Université Paris Descartes, Paris, France.

Introduction

Cadre

- Apprentissage supervisé

Introduction

Cadre

- Apprentissage supervisé
- Analyse de survie

Introduction

Cadre

- Apprentissage supervisé
- Analyse de survie
- Statistique en grande dimension

Introduction

Cadre

- Apprentissage supervisé
- Analyse de survie
- Statistique en grande dimension
- Données longitudinales

Introduction

Cadre

- Apprentissage supervisé
- Analyse de survie
- Statistique en grande dimension
- Données longitudinales

Principaux projets entrepris

I. Trajectories of biological values and vital parameters, a retrospective cohort study on non-complicated vaso-occlusive crises

Introduction

Cadre

- Apprentissage supervisé
- Analyse de survie
- Statistique en grande dimension
- Données longitudinales

Principaux projets entrepris

I. Trajectories of biological values and vital parameters, a retrospective cohort study on non-complicated vaso-occlusive crises

II. Early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework

Introduction

Cadre

- Apprentissage supervisé
- Analyse de survie
- Statistique en grande dimension
- Données longitudinales

Principaux projets entrepris

I. Trajectories of biological values and vital parameters, a retrospective cohort study on non-complicated vaso-occlusive crises

II. Early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework

III. C-mix, a high dimensional mixture model for censored durations

Introduction

Cadre

- Apprentissage supervisé
- Analyse de survie
- Statistique en grande dimension
- Données longitudinales

Principaux projets entrepris

I. Trajectories of biological values and vital parameters, a retrospective cohort study on non-complicated vaso-occlusive crises
II. Early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework
III. C-mix, a high dimensional mixture model for censored durations
IV. Binarsity, a penalization for one-hot encoded features

Introduction

Cadre

- Apprentissage supervisé
- Analyse de survie
- Statistique en grande dimension
- Données longitudinales

Principaux projets entrepris

I. Trajectories of biological values and vital parameters, a retrospective cohort study on non-complicated vaso-occlusive crises
II. Early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework
III. C-mix, a high dimensional mixture model for censored durations
IV. Binarsity, a penalization for one-hot encoded features
V. Binacox, automatic cut-points detection in a high-dimensional Cox model

I. Trajectories of biological values and vital parameters, a retrospective cohort study on non-complicated vaso-occlusive crises

Introduction

Contexte

- Drépanocytose : maladie génétique la plus fréquente [17]

Introduction

Contexte

- Drépanocytose : maladie génétique la plus fréquente [17]
- Crises vaso-occlusives (CVO)

Introduction

Contexte

- Drépanocytose : maladie génétique la plus fréquente [17]
- Crises vaso-occlusives (CVO)
- Pas de biomarqueur pour le suivi

Introduction

Contexte

- Drépanocytose : maladie génétique la plus fréquente [17]
- Crises vaso-occlusives (CVO)
- Pas de biomarqueur pour le suivi
- Étude rétrospective : cohorte de l'HEGP

Introduction

Contexte

- Drépanocytose : maladie génétique la plus fréquente [17]
- Crises vaso-occlusives (CVO)
- Pas de biomarqueur pour le suivi
- Étude rétrospective : cohorte de l'HEGP

Objectifs

- Décrire l'évolution des biomarqueurs et paramètres vitaux lors d'une CVO "non compliquée"

Introduction

Contexte

- Drépanocytose : maladie génétique la plus fréquente [17]
- Crises vaso-occlusives (CVO)
- Pas de biomarqueur pour le suivi
- Étude rétrospective : cohorte de l'HEGP

Objectifs

- Décrire l'évolution des biomarqueurs et paramètres vitaux lors d'une CVO "non compliquée"
- Détecter la présence d'une complication

Introduction

Contexte

- Drépanocytose : maladie génétique la plus fréquente [17]
- Crises vaso-occlusives (CVO)
- Pas de biomarqueur pour le suivi
- Étude rétrospective : cohorte de l'HEGP

Objectifs

- Décrire l'évolution des biomarqueurs et paramètres vitaux lors d'une CVO "non compliquée"
- Détecter la présence d'une complication
- Identifier quel(s) biomarqueur(s) surveiller

Méthode

Données

- Patients admis à l'HEGP pour CVO entre 2010 $\rightarrow$ 2015

Méthode

Données

- Patients admis à l'HEGP pour CVO entre 2010 $\rightarrow$ 2015
- 329 séjours, 164 patients

Méthode

Données

- Patients admis à l'HEGP pour CVO entre 2010 $\rightarrow$ 2015
- 329 séjours, 164 patients
- Données longitudinales

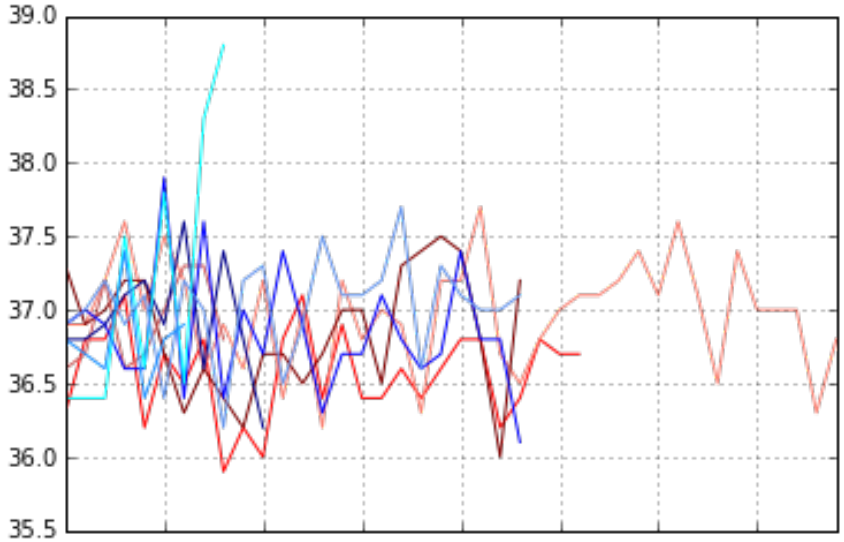


Figure 1: Évolution de la température pour 10 patients

Méthode

Données

- Patients admis à l'HEGP pour CVO entre 2010 $\rightarrow$ 2015
- 329 séjours, 164 patients
- Données longitudinales

Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour i :

- on génére une grille uniforme de temps t_k

Méthode

Données

- Patients admis à l'HEGP pour CVO entre 2010 $\rightarrow$ 2015
- 329 séjours, 164 patients
- Données longitudinales

Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour i :

- on génére une grille uniforme de temps t_k
- on ajuste un *spline* f_i

Méthode

Données

- Patients admis à l'HEGP pour CVO entre 2010 $\rightarrow$ 2015
- 329 séjours, 164 patients
- Données longitudinales

Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour i :

- on génére une grille uniforme de temps t_k
- on ajuste un *spline* f_i
- on calcule $f_i(t_k)$ pour chaque temps t_k

Méthode

Données

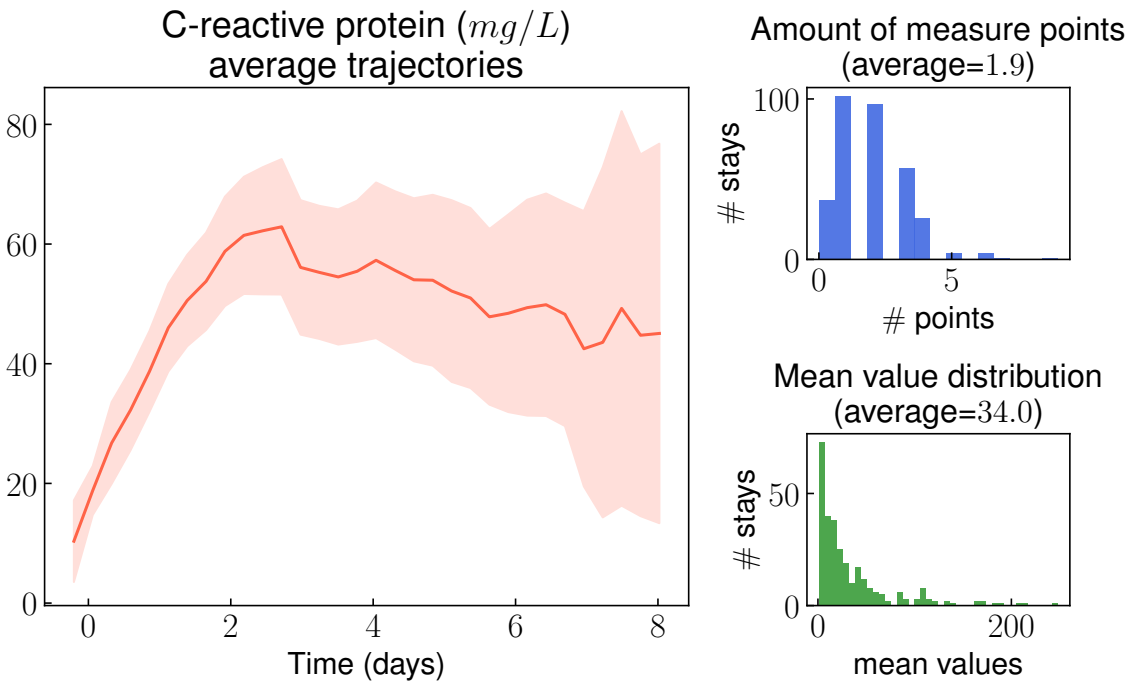
- Patients admis à l'HEGP pour CVO entre 2010 $\rightarrow$ 2015
- 329 séjours, 164 patients
- Données longitudinales

Description des trajectoires moyennes

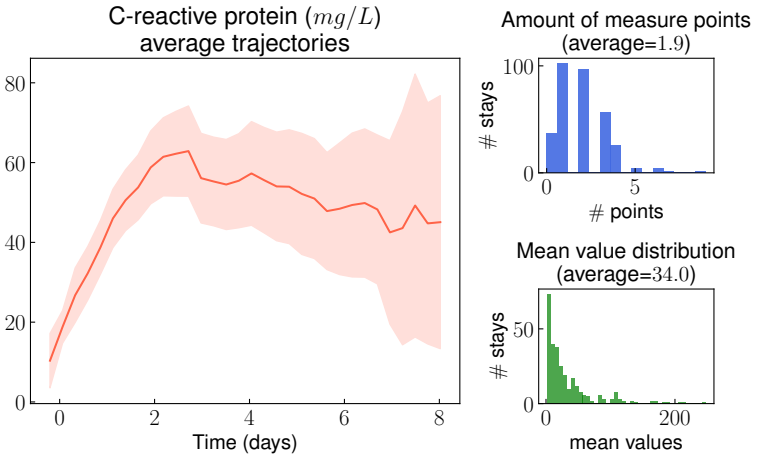
Pour chaque trajectoire et pour chaque séjour i :

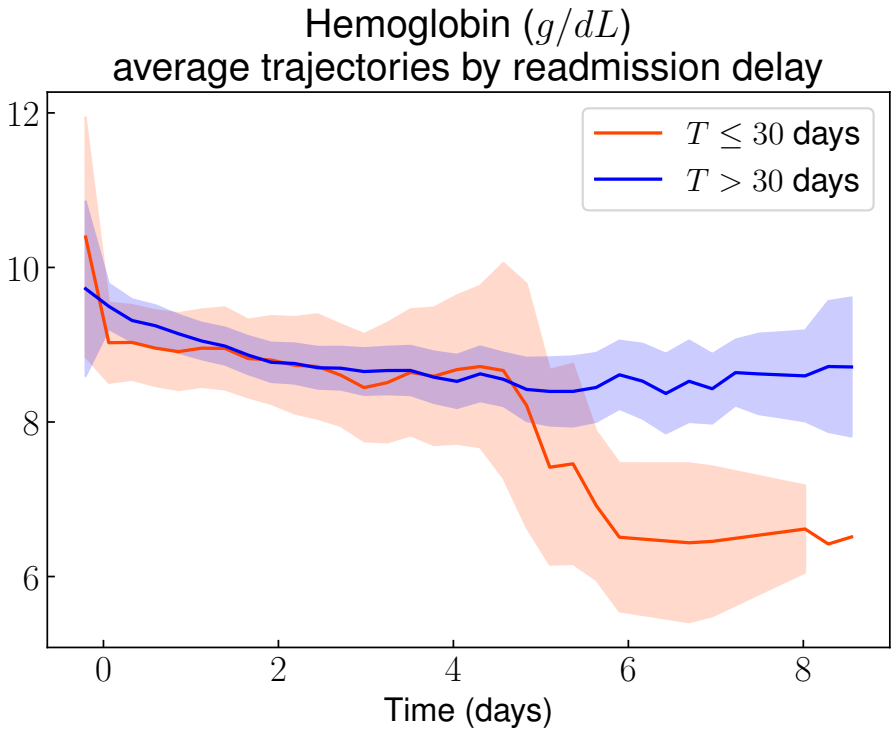
- on génére une grille uniforme de temps t_k
- on ajuste un *spline* f_i
- on calcule $f_i(t_k)$ pour chaque temps t_k
- on déduit une trajectoire moyenne avec IC

Examples de résultats graphiques



Examples de résultats graphiques





Conclusion

- Plusieurs biomarqueurs et paramètres vitaux évoluent

Conclusion

- Plusieurs biomarqueurs et paramètres vitaux évoluent
- Extraction d'information pertinente à partir d'un entrepôt de données cliniques de grande dimension

Conclusion

- Plusieurs biomarqueurs et paramètres vitaux évoluent
- Extraction d'information pertinente à partir d'un entrepôt de données cliniques de grande dimension

Article associé

R. Veil, **S. Bussy**, V. Looten, J.B. Arlet, J. Pouchot Camoz, A.S. Jannot et B. Ranque

Trajectories of biological values and vital parameters : a retrospective cohort study on non-complicated vaso-occlusive crises

Publié dans *Journal of Clinical Medicine*, 2019.

Conclusion

- Plusieurs biomarqueurs et paramètres vitaux évoluent
- Extraction d'information pertinente à partir d'un entrepôt de données cliniques de grande dimension

Article associé

R. Veil, **S. Bussy**, V. Looten, J.B. Arlet, J. Pouchot Camoz, A.S. Jannot et B. Ranque

Trajectories of biological values and vital parameters : a retrospective cohort study on non-complicated vaso-occlusive crises

Publié dans *Journal of Clinical Medicine*, 2019.

Code `Python`

- Disponible à `https://github.com/SimonBussy/redcvo`

Conclusion

- Plusieurs biomarqueurs et paramètres vitaux évoluent
- Extraction d'information pertinente à partir d'un entrepôt de données cliniques de grande dimension

Article associé

R. Veil, **S. Bussy**, V. Looten, J.B. Arlet, J. Pouchot Camoz, A.S. Jannot et B. Ranque

Trajectories of biological values and vital parameters : a retrospective cohort study on non-complicated vaso-occlusive crises

Publié dans *Journal of Clinical Medicine*, 2019.

Code `Python`

- Disponible à `https://github.com/SimonBussy/redcvo`
- Figures pour tous les biomarqueurs et paramètres vitaux

II. Early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework

Introduction

- Facteurs prédictifs de la réhospitalisation pour CVO

Introduction

- Facteurs prédictifs de la réhospitalisation pour CVO
- Problèmes récurrents dans les études cliniques :

Introduction

- Facteurs prédictifs de la réhospitalisation pour CVO
- Problèmes récurrents dans les études cliniques :
  - prédiction du risque

Introduction

- Facteurs prédictifs de la réhospitalisation pour CVO
- Problèmes récurrents dans les études cliniques :
  - prédiction du risque
  - identification des covariables impliquées

Introduction

- Facteurs prédictifs de la réhospitalisation pour CVO
- Problèmes récurrents dans les études cliniques :
  - prédiction du risque
  - identification des covariables impliquées

Deux cadres

- *Binary outcome setting*

Introduction

- Facteurs prédictifs de la réhospitalisation pour CVO
- Problèmes récurrents dans les études cliniques :
  - prédiction du risque
  - identification des covariables impliquées

Deux cadres

- *Binary outcome setting*
  - Réhospitalisation "précoce" en se basant sur un seuil ϵ pré-défini

Introduction

- Facteurs prédictifs de la réhospitalisation pour CVO
- Problèmes récurrents dans les études cliniques :
  - prédiction du risque
  - identification des covariables impliquées

Deux cadres

- *Binary outcome setting*
  - Réhospitalisation "précoce" en se basant sur un seuil ϵ pré-défini
  - Résultats très dépendants du choix de ϵ [4]

Introduction

- Facteurs prédictifs de la réhospitalisation pour CVO
- Problèmes récurrents dans les études cliniques :
  - prédiction du risque
  - identification des covariables impliquées

Deux cadres

- *Binary outcome setting*
  - Réhospitalisation "précoce" en se basant sur un seuil ϵ pré-défini
  - Résultats très dépendants du choix de ϵ [4]
- *Survival analysis setting* : pas de seuil *a priori*

Introduction

- Facteurs prédictifs de la réhospitalisation pour CVO
- Problèmes récurrents dans les études cliniques :
  - prédiction du risque
  - identification des covariables impliquées

Deux cadres

- *Binary outcome setting*
  - Réhospitalisation "précoce" en se basant sur un seuil ϵ pré-défini
  - Résultats très dépendants du choix de ϵ [4]
- *Survival analysis setting* : pas de seuil *a priori*

Objectif

Comparer des méthodes d'apprentissage issues de ces deux cadres

Cas d'étude

- Étude rétrospective monocentrique sur la cohorte de I.

Cas d'étude

- Étude rétrospective monocentrique sur la cohorte de I.
- On tire aléatoirement 1 séjour par patient (*i.i.d.*)

Cas d'étude

- Étude rétrospective monocentrique sur la cohorte de I.
- On tire aléatoirement 1 séjour par patient (*i.i.d.*)

Extraction de covariables

À partir des données longitudinales, par exemple :

- Pente d'une régression linéaire (dernières 48h)

Cas d'étude

- Étude rétrospective monocentrique sur la cohorte de I.
- On tire aléatoirement 1 séjour par patient (*i.i.d.*)

Extraction de covariables

À partir des données longitudinales, par exemple :

- Pente d'une régression linéaire (dernières 48h)
- Hyper-paramètres des noyaux de PG [18]

Cas d'étude

- Étude rétrospective monocentrique sur la cohorte de I.
- On tire aléatoirement 1 séjour par patient (*i.i.d.*)

Extraction de covariables

À partir des données longitudinales, par exemple :

- Pente d'une régression linéaire (dernières 48h)
- Hyper-paramètres des noyaux de PG [18]

174 covariables

- démographiques

Cas d'étude

- Étude rétrospective monocentrique sur la cohorte de I.
- On tire aléatoirement 1 séjour par patient (*i.i.d.*)

Extraction de covariables

À partir des données longitudinales, par exemple :

- Pente d'une régression linéaire (dernières 48h)
- Hyper-paramètres des noyaux de PG [18]

174 covariables

- démographiques
- qualitatives (ex: type d'opioïde administré)

Cas d'étude

- Étude rétrospective monocentrique sur la cohorte de I.
- On tire aléatoirement 1 séjour par patient (*i.i.d.*)

Extraction de covariables

À partir des données longitudinales, par exemple :

- Pente d'une régression linéaire (dernières 48h)
- Hyper-paramètres des noyaux de PG [18]

174 covariables

- démographiques
- qualitatives (ex: type d'opioïde administré)
- quantitative (ex: paramètres biologiques/vitaux/doses d'opioïde)

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]
  - Régression Logistique (LR) [12]

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]
  - Régression Logistique (LR) [12]
  - SVM avec noyau linéaire [19]

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]
  - Régression Logistique (LR) [12]
  - SVM avec noyau linéaire [19]
  - Forêts aléatoires (RF) [2]

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]
  - Régression Logistique (LR) [12]
  - SVM avec noyau linéaire [19]
  - Forêts aléatoires (RF) [2]
  - Gradient Boosting (GB) [7]

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]
  - Régression Logistique (LR) [12]
  - SVM avec noyau linéaire [19]
  - Forêts aléatoires (RF) [2]
  - Gradient Boosting (GB) [7]
  - Réseaux de neuronnes à une couche cachée (NN) [23]

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]
  - Régression Logistique (LR) [12]
  - SVM avec noyau linéaire [19]
  - Forêts aléatoires (RF) [2]
  - Gradient Boosting (GB) [7]
  - Réseaux de neuronnes à une couche cachée (NN) [23]
- *Survival analysis setting*, score : C-index [9]

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]
  - Régression Logistique (LR) [12]
  - SVM avec noyau linéaire [19]
  - Forêts aléatoires (RF) [2]
  - Gradient Boosting (GB) [7]
  - Réseaux de neuronnes à une couche cachée (NN) [23]
- *Survival analysis setting*, score : C-index [9]
  - Cox PH [5]

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]
  - Régression Logistique (LR) [12]
  - SVM avec noyau linéaire [19]
  - Forêts aléatoires (RF) [2]
  - Gradient Boosting (GB) [7]
  - Réseaux de neuronnes à une couche cachée (NN) [23]
- *Survival analysis setting*, score : C-index [9]
  - Cox PH [5]
  - CURE [6]

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]
  - Régression Logistique (LR) [12]
  - SVM avec noyau linéaire [19]
  - Forêts aléatoires (RF) [2]
  - Gradient Boosting (GB) [7]
  - Réseaux de neuronnes à une couche cachée (NN) [23]
- *Survival analysis setting*, score : C-index [9]
  - Cox PH [5]
  - CURE [6]
  - C-mix [3]

Méthode

Modèles considérés

- *Binary outcome setting* ($\epsilon = 30$ jours), score : AUC [1]
  - Régression Logistique (LR) [12]
  - SVM avec noyau linéaire [19]
  - Forêts aléatoires (RF) [2]
  - Gradient Boosting (GB) [7]
  - Réseaux de neuronnes à une couche cachée (NN) [23]
- *Survival analysis setting*, score : C-index [9]
  - Cox PH [5]
  - CURE [6]
  - C-mix [3]

Comparaison des deux cadres

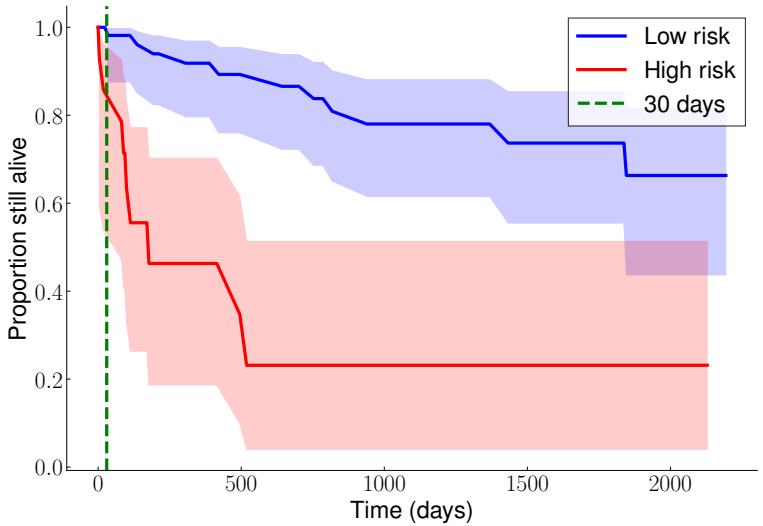
Prédiction de $T_i > \epsilon$ par $\hat{S}_i(\epsilon|X_i = x_i)$, score : AUC

Résultats en prédiction

Table 1: Performances en prédiction

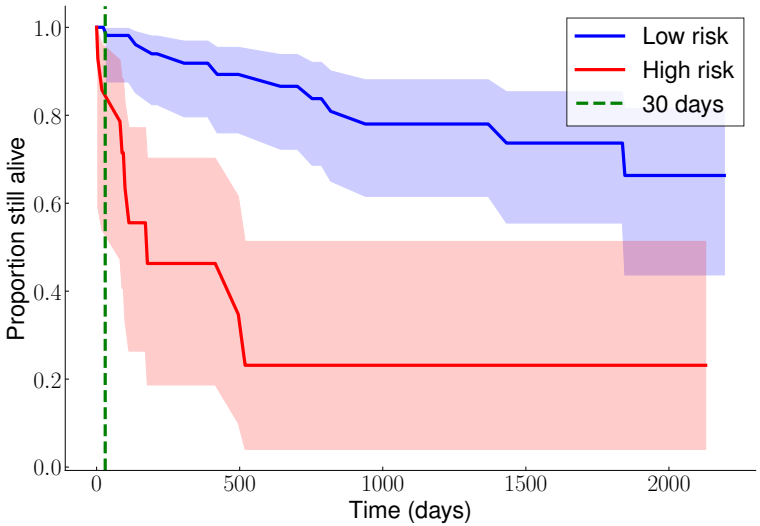
Setting	Métrique	Modèle	Score
Survival analysis	C-index	CURE	0.718
		Cox PH	0.725
		C-mix	**0.754**
Binary outcome ($\epsilon = 30$)	AUC	SVM	0.524
		GB	0.561
		LR	0.616
		NN	0.707
		RF	0.738
		$\hat{S}^{\text{CURE}}(\epsilon)$	0.831
		$\hat{S}^{\text{Cox PH}}(\epsilon)$	0.855
		$\hat{S}^{\text{C-mix}}(\epsilon)$	**0.940**

Résultats : le C-mix

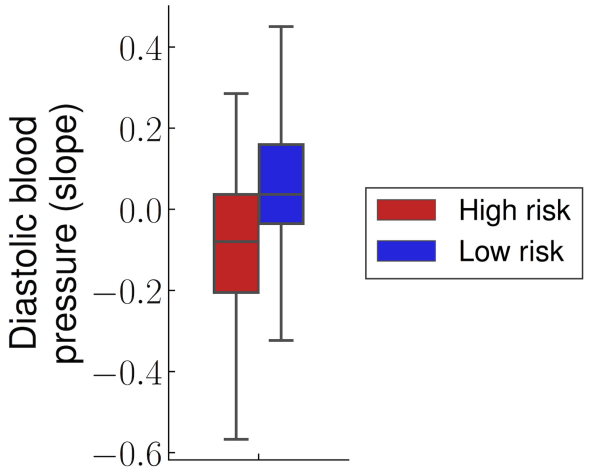


(a) Courbes de survie

Résultats : le C-mix



(a) Courbes de survie



(b) Boxplot

Conclusion

- Modèles d'analyse de survie $\rightarrow$ fonctions de survie estimées $\rightarrow$ prédictions binaires pour un ϵ donné

Conclusion

- Modèles d'analyse de survie $\rightarrow$ fonctions de survie estimées $\rightarrow$ prédictions binaires pour un ϵ donné
- Méthodologie pour la création de features pertinentes

Conclusion

- Modèles d'analyse de survie $\rightarrow$ fonctions de survie estimées $\rightarrow$ prédictions binaires pour un ϵ donné
- Méthodologie pour la création de features pertinentes
- Bonnes performances du C-mix + aspects d'interprétation intéressants

Conclusion

- Modèles d'analyse de survie $\rightarrow$ fonctions de survie estimées $\rightarrow$ prédictions binaires pour un ϵ donné
- Méthodologie pour la création de features pertinentes
- Bonnes performances du C-mix + aspects d'interprétation intéressants

Article associé

S. Bussy, R. Veil, V. Looten, A. Burgun, S. Gaïffas, A. Guilloux, B. Ranque et A.S. Jannot

Comparison of methods for early-readmission prediction in a high dimensional heterogeneous covariates and time-to-event outcome framework

Publié dans *BMC Medical Research Methodology*, 2019.

Conclusion

- Modèles d'analyse de survie $\rightarrow$ fonctions de survie estimées $\rightarrow$ prédictions binaires pour un ϵ donné
- Méthodologie pour la création de features pertinentes
- Bonnes performances du C-mix + aspects d'interprétation intéressants

Article associé

S. Bussy, R. Veil, V. Looten, A. Burgun, S. Gaïffas, A. Guilloux, B. Ranque et A.S. Jannot

Comparison of methods for early-readmission prediction in a high dimensional heterogeneous covariates and time-to-event outcome framework

Publié dans *BMC Medical Research Methodology*, 2019.

Code `Python`

Disponible à

`https://github.com/SimonBussy/early-readmission-prediction`

III. C-mix, a high dimensional mixture model for censored durations

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \ldots, K-1\}$

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \ldots, K-1\}$
- Modèle de mélange $f(t|X = x) = \sum_{k=0}^{K-1} \pi_{\beta_k}(x) f_k(t; \alpha_k)$

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \ldots, K-1\}$
- Modèle de mélange $f(t|X = x) = \sum_{k=0}^{K-1} \pi_{\beta_k}(x) f_k(t; \alpha_k)$
- $\pi_{\beta_k}(x) = \mathbb{P}[Z = k|X = x] = \frac{e^{x^\top \beta_k}}{\sum_{k=0}^{K-1} e^{x^\top \beta_k}}$ (softmax)

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \dots, K-1\}$
- Modèle de mélange $f(t|X = x) = \sum_{k=0}^{K-1} \pi_{\beta_k}(x) f_k(t; \alpha_k)$
- $\pi_{\beta_k}(x) = \mathbb{P}[Z = k|X = x] = \frac{e^{x^\top \beta_k}}{\sum_{k=0}^{K-1} e^{x^\top \beta_k}}$ (softmax)
- Échantillon *i.i.d.* $(x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n) \in \mathbb{R}^d \times \mathbb{R}_+ \times \{0, 1\}$

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \ldots, K-1\}$
- Modèle de mélange $f(t|X = x) = \sum_{k=0}^{K-1} \pi_{\beta_k}(x) f_k(t; \alpha_k)$
- $\pi_{\beta_k}(x) = \mathbb{P}[Z = k|X = x] = \frac{e^{x^\top \beta_k}}{\sum_{k=0}^{K-1} e^{x^\top \beta_k}}$ (softmax)
- Échantillon *i.i.d.* $(x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n) \in \mathbb{R}^d \times \mathbb{R}_+ \times \{0, 1\}$
- $\theta = (\alpha_0, \ldots, \alpha_K, \beta_0, \ldots, \beta_K)^\top$, log-vraissemblance du C-mix

$$\ell_n(\theta) = n^{-1} \sum_{i=1}^{n} \left\{ \delta_i \log \left[\bar{G}(y_i^-) \sum_{k=0}^{K-1} \pi_{\beta_k}(x_i) f_k(y_i; \alpha_k) \right] \right.$$
$$\left. + (1 - \delta_i) \log \left[g(y_i) \sum_{k=0}^{K-1} \pi_{\beta_k}(x_i) \bar{F}_k(y_i^-; \alpha_k) \right] \right\}$$

L'algorithme QNEM

- Objectif pénalisé

$$\ell_n^{\text{pen}}(\theta) = -\ell_n(\theta) + \sum_{k=0}^{K-1} \gamma_k \left((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2 \right) \qquad (1)$$

L'algorithme QNEM

- Objectif pénalisé

$$\ell_n^{\text{pen}}(\theta) = -\ell_n(\theta) + \sum_{k=0}^{K-1} \gamma_k \left((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2 \right) \quad (1)$$

- $\ell_n^{\text{comp}}(\theta)$ log-vraissemblance complétée (négative)

$$-n^{-1} \sum_{i=1}^{n} \left\{ \delta_i \left[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_i=k\}} \left(\log \pi_{\beta_k}(x_i) + \log f_k(y_i; \alpha_k) \right) + \log \bar{G}(y_i^-) \right] \right.$$
$$\left. + (1-\delta_i) \left[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_i=k\}} \left(\log \pi_{\beta_k}(x_i) + \log \bar{F}_k(y_i^-; \alpha_k) \right) + \log g(y_i) \right] \right\}$$

L'algorithme QNEM

- Objectif pénalisé

$$\ell_n^{\text{pen}}(\theta) = -\ell_n(\theta) + \sum_{k=0}^{K-1} \gamma_k \left((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2 \right) \quad (1)$$

- $\ell_n^{\text{comp}}(\theta)$ log-vraissemblance complétée (négative)

$$-n^{-1} \sum_{i=1}^{n} \left\{ \delta_i \left[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_i=k\}} \left(\log \pi_{\beta_k}(x_i) + \log f_k(y_i; \alpha_k) \right) + \log \bar{G}(y_i^-) \right] \right.$$
$$\left. + (1-\delta_i) \left[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_i=k\}} \left(\log \pi_{\beta_k}(x_i) + \log \bar{F}_k(y_i^-; \alpha_k) \right) + \log g(y_i) \right] \right\}$$

- Étape E: $Q_n(\theta, \theta^{(l)}) = \mathbb{E}_{\theta^{(l)}}[\ell_n^{\text{comp}}(\theta) | \boldsymbol{y}, \boldsymbol{\delta}]$

$$q_{i,k}^{(l)} = \mathbb{E}_{\theta^{(l)}}[\mathbb{1}_{\{z_i=k\}} | y_i, \delta_i] = \mathbb{P}_{\theta^{(l)}}[z_i = k | y_i, \delta_i] = \frac{\Lambda_{k,i}^{(l)}}{\sum_{r=0}^{K-1} \Lambda_{r,i}^{(l)}}$$

$$\Lambda_{k,i}^{(l)} = \left[f_k(y_i; \alpha_k^{(l)}) \bar{G}(y_i^-) \right]^{\delta_i} \left[g(y_i) \bar{F}_k(y_i^-; \alpha_k^{(l)}) \right]^{1-\delta_i} \pi_{\beta_k^{(l)}}(x_i)$$

L'algorithme QNEM

- Étape M: problème d'optimisation convexe...

$$\text{minimiser} \quad R_{n,k}^{(l)}(\beta_k) + \gamma_k\big((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2\big), \qquad (2)$$

avec $R_{n,k}^{(l)}(\beta_k) = -n^{-1}\sum_{i=1}^{n} q_{i,k}^{(l)} \log \pi_{\beta_k}(x_i)$

L'algorithme QNEM

- Étape M: problème d'optimisation convexe...

$$\text{minimiser} \quad R^{(l)}_{n,k}(\beta_k) + \gamma_k\big((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2\big), \qquad (2)$$

avec $R^{(l)}_{n,k}(\beta_k) = -n^{-1}\sum_{i=1}^{n} q^{(l)}_{i,k} \log \pi_{\beta_k}(x_i)$

- ...mais non différentiable! On réécrit alors (2) comme :

$$\text{minimiser } R^{(l)}_{n,k}(\beta_k^+ - \beta_k^-) + \gamma_k(1-\eta)\sum_{j=1}^{d}(\beta_{k,j}^+ + \beta_{k,j}^-) + \gamma_k\frac{\eta}{2}\|\beta_k^+ - \beta_k^-\|_2^2$$

tel que $\beta_{k,j}^+ \geq 0$ et $\beta_{k,j}^- \geq 0$ pour tout $j \in \{1,\ldots,d\}$

L'algorithme QNEM

- Étape M: problème d'optimisation convexe...

$$\text{minimiser} \quad R_{n,k}^{(l)}(\beta_k) + \gamma_k\big((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2\big), \qquad (2)$$

avec $R_{n,k}^{(l)}(\beta_k) = -n^{-1}\sum_{i=1}^{n} q_{i,k}^{(l)} \log \pi_{\beta_k}(x_i)$

- ...mais non différentiable! On réécrit alors (2) comme :

$$\text{minimiser } R_{n,k}^{(l)}(\beta_k^+ - \beta_k^-) + \gamma_k(1-\eta)\sum_{j=1}^{d}(\beta_{k,j}^+ + \beta_{k,j}^-) + \gamma_k\frac{\eta}{2}\|\beta_k^+ - \beta_k^-\|_2^2$$

tel que $\beta_{k,j}^+ \geq 0$ et $\beta_{k,j}^- \geq 0$ pour tout $j \in \{1,\ldots,d\}$

- Solveur L-BFGS-B, requiert le gradient qui s'écrit

$$\frac{\partial R_{n,k}^{(l)}(\beta_k)}{\partial \beta_k} = -n^{-1}\sum_{i=1}^{n} q_{i,k}^{(l)}\big(1 - \pi_{\beta_k}(x_i)\big)x_i$$

L'algorithme QNEM

- Étape M: problème d'optimisation convexe...

$$\text{minimiser} \quad R_{n,k}^{(l)}(\beta_k) + \gamma_k\big((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2\big), \qquad (2)$$

avec $R_{n,k}^{(l)}(\beta_k) = -n^{-1}\sum_{i=1}^{n} q_{i,k}^{(l)} \log \pi_{\beta_k}(x_i)$

- ...mais non différentiable! On réécrit alors (2) comme :

$$\text{minimiser } R_{n,k}^{(l)}(\beta_k^+ - \beta_k^-) + \gamma_k(1-\eta)\sum_{j=1}^{d}(\beta_{k,j}^+ + \beta_{k,j}^-) + \gamma_k\frac{\eta}{2}\|\beta_k^+ - \beta_k^-\|_2^2$$

tel que $\beta_{k,j}^+ \geq 0$ et $\beta_{k,j}^- \geq 0$ pour tout $j \in \{1,\ldots,d\}$

- Solveur L-BFGS-B, requiert le gradient qui s'écrit

$$\frac{\partial R_{n,k}^{(l)}(\beta_k)}{\partial \beta_k} = -n^{-1}\sum_{i=1}^{n} q_{i,k}^{(l)}\big(1 - \pi_{\beta_k}(x_i)\big)x_i$$

- Convergence vers un minimum local prouvée

Application sur données génétiques

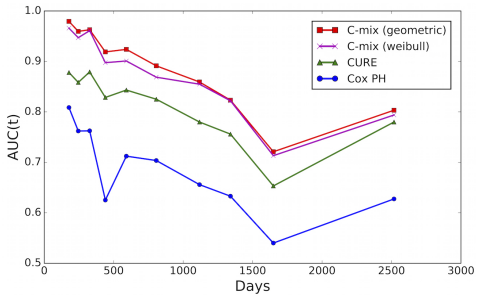
Table 2: Comparison du C-index sur les données TCGA ($d = 20531$)

Cancer		BRCA ($n = 1211$)			GBM ($n = 168$)			KIRC ($n = 605$)		
Modèle		C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
d	100	**0.792**	0.764	0.705	**0.826**	0.695	0.571	**0.768**	0.732	0.716
	300	**0.782**	0.753	0.723	**0.849**	0.697	0.571	**0.755**	0.691	0.698
	1000	**0.817**	0.613	0.577	**0.775**	0.699	0.592	**0.743**	0.690	0.685

Application sur données génétiques

Table 2: Comparison du C-index sur les données TCGA ($d = 20531$)

Cancer		BRCA ($n = 1211$)			GBM ($n = 168$)			KIRC ($n = 605$)		
Modèle		C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
	100	**0.792**	0.764	0.705	**0.826**	0.695	0.571	**0.768**	0.732	0.716
d	300	**0.782**	0.753	0.723	**0.849**	0.697	0.571	**0.755**	0.691	0.698
	1000	**0.817**	0.613	0.577	**0.775**	0.699	0.592	**0.743**	0.690	0.685



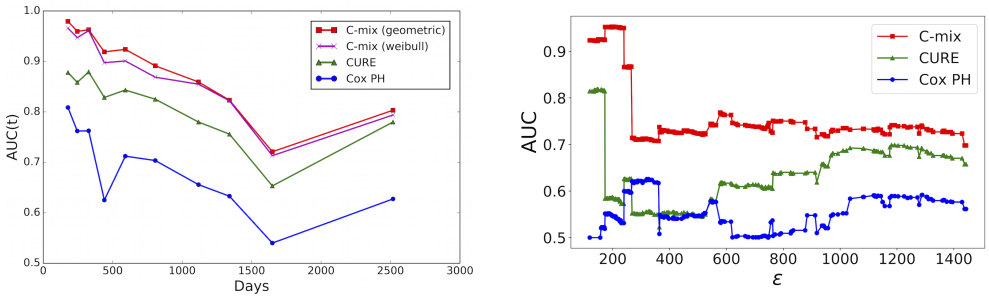
(a) AUC(t)

Figure 3: Résultats sur le cancer BRCA

Application sur données génétiques

Table 2: Comparison du C-index sur les données TCGA ($d = 20531$)

Cancer		BRCA ($n = 1211$)			GBM ($n = 168$)			KIRC ($n = 605$)		
Modèle		C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
d	100	**0.792**	0.764	0.705	**0.826**	0.695	0.571	**0.768**	0.732	0.716
	300	**0.782**	0.753	0.723	**0.849**	0.697	0.571	**0.755**	0.691	0.698
	1000	**0.817**	0.613	0.577	**0.775**	0.699	0.592	**0.743**	0.690	0.685



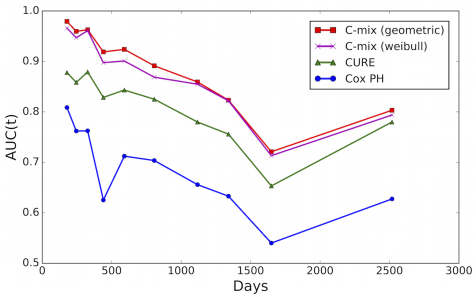
(a) AUC(t) (b) AUC

Figure 3: Résultats sur le cancer BRCA

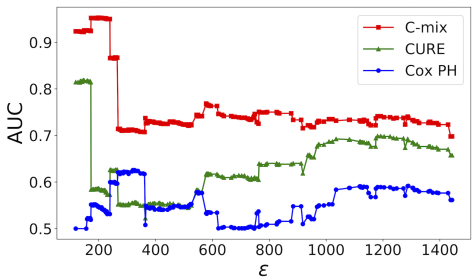
Application sur données génétiques

Table 2: Comparison du C-index sur les données TCGA ($d = 20531$)

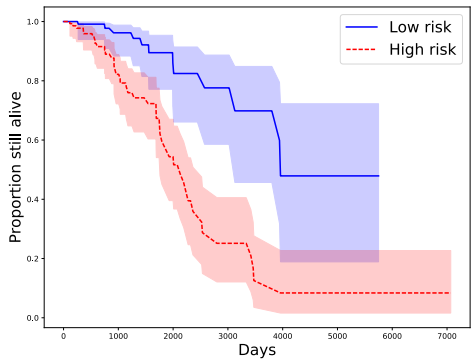
Cancer		BRCA ($n = 1211$)			GBM ($n = 168$)			KIRC ($n = 605$)		
Modèle		C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
	100	**0.792**	0.764	0.705	**0.826**	0.695	0.571	**0.768**	0.732	0.716
d	300	**0.782**	0.753	0.723	**0.849**	0.697	0.571	**0.755**	0.691	0.698
	1000	**0.817**	0.613	0.577	**0.775**	0.699	0.592	**0.743**	0.690	0.685



(a) AUC(t)



(b) AUC



(c) Courbes de survie

Figure 3: Résultats sur le cancer BRCA

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité (gènes connus)

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité (gènes connus)
- Détection de sous-groupes de patients relativement à leurs risques

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité (gènes connus)
- Détection de sous-groupes de patients relativement à leurs risques

Article associé

S. Bussy, A. Guilloux, S. Gaïffas, A.S. Jannot
C-mix: a high dimensional mixture model for censored durations, with applications to genetic data
Publié dans *Statistical Methods in Medical Research*, 2018.

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité (gènes connus)
- Détection de sous-groupes de patients relativement à leurs risques

Article associé

S. Bussy, A. Guilloux, S. Gaïffas, A.S. Jannot
C-mix: a high dimensional mixture model for censored durations, with applications to genetic data
Publié dans *Statistical Methods in Medical Research*, 2018.

Code `Python`

- Disponible à `https://github.com/SimonBussy/C-mix`

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité (gènes connus)
- Détection de sous-groupes de patients relativement à leurs risques

Article associé

S. Bussy, A. Guilloux, S. Gaïffas, A.S. Jannot
C-mix: a high dimensional mixture model for censored durations, with applications to genetic data
Publié dans *Statistical Methods in Medical Research*, 2018.

Code `Python`

- Disponible à `https://github.com/SimonBussy/C-mix`
- Programmes annotés, notebooks et tutoriels

IV. Binarsity, a penalization for one-hot encoded features

Le contexte

- Apprentissage supervisé $(x_i, y_i)_{i=1,\dots,n}$ avec x_i continues

Le contexte

- Apprentissage supervisé $(x_i, y_i)_{i=1,\dots,n}$ avec x_i continues
- Encodage one-hot [22] : $x_i = (x_{i,1}, \dots, x_{i,p})^\top \in \mathbb{R}^p$ transformé en
$x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_1}^B, x_{i,2,1}^B, \dots, x_{i,2,d_2}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_p}^B)^\top \in \mathbb{R}^d$,
t.q. pour $i = 1, \dots, n$ et $k = 1, \dots, d_j$, on a

$$x_{i,j,k}^B = \begin{cases} 1 & \text{si } x_{i,j} \in I_{j,k} = \left(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right], \text{ (interquantiles)} \\ 0 & \text{sinon} \end{cases}$$

Le contexte

- Apprentissage supervisé $(x_i, y_i)_{i=1,\dots,n}$ avec x_i continues
- Encodage one-hot [22] : $x_i = (x_{i,1}, \dots, x_{i,p})^\top \in \mathbb{R}^p$ transformé en
  $x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_1}^B, x_{i,2,1}^B, \dots, x_{i,2,d_2}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_p}^B)^\top \in \mathbb{R}^d$,
  t.q. pour $i = 1, \dots, n$ et $k = 1, \dots, d_j$, on a

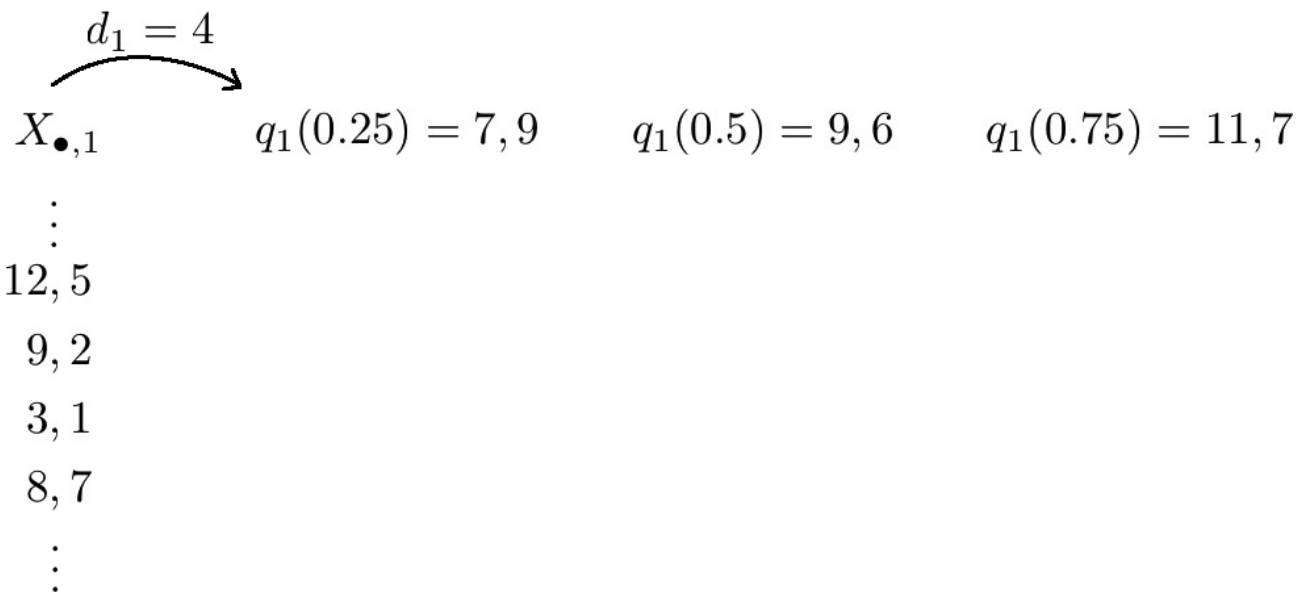
$$x_{i,j,k}^B = \begin{cases} 1 & \text{si } x_{i,j} \in I_{j,k} = \left(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right], \text{ (interquantiles)} \\ 0 & \text{sinon} \end{cases}$$

$$\begin{matrix} X_{\bullet,1} \\ \vdots \\ 12,5 \\ 9,2 \\ 3,1 \\ 8,7 \\ \vdots \end{matrix}$$

Le contexte

- Apprentissage supervisé $(x_i, y_i)_{i=1,\dots,n}$ avec x_i continues
- Encodage one-hot [22] : $x_i = (x_{i,1}, \dots, x_{i,p})^\top \in \mathbb{R}^p$ transformé en
$x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_1}^B, x_{i,2,1}^B, \dots, x_{i,2,d_2}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_p}^B)^\top \in \mathbb{R}^d$,
t.q. pour $i = 1, \dots, n$ et $k = 1, \dots, d_j$, on a

$$x_{i,j,k}^B = \begin{cases} 1 & \text{si } x_{i,j} \in I_{j,k} = \left(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right], \text{ (interquantiles)} \\ 0 & \text{sinon} \end{cases}$$



Le contexte

- Apprentissage supervisé $(x_i, y_i)_{i=1,\dots,n}$ avec x_i continues
- Encodage one-hot [22] : $x_i = (x_{i,1}, \dots, x_{i,p})^\top \in \mathbb{R}^p$ transformé en
$x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_1}^B, x_{i,2,1}^B, \dots, x_{i,2,d_2}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_p}^B)^\top \in \mathbb{R}^d$,
t.q. pour $i = 1, \dots, n$ et $k = 1, \dots, d_j$, on a

$$x_{i,j,k}^B = \begin{cases} 1 & \text{si } x_{i,j} \in I_{j,k} = \left(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right], \text{ (interquantiles)} \\ 0 & \text{sinon} \end{cases}$$

$d_1 = 4$

$X_{\bullet,1}$	$q_1(0.25) = 7,9$	$q_1(0.5) = 9,6$	$q_1(0.75) = 11,7$
$\vdots$			
$i \rightarrow 12,5$			
$9,2$			
$3,1$			
$8,7$			
$\vdots$			

Le contexte

- Apprentissage supervisé $(x_i, y_i)_{i=1,\dots,n}$ avec x_i continues
- Encodage one-hot [22] : $x_i = (x_{i,1}, \dots, x_{i,p})^\top \in \mathbb{R}^p$ transformé en
$x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_1}^B, x_{i,2,1}^B, \dots, x_{i,2,d_2}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_p}^B)^\top \in \mathbb{R}^d$,
t.q. pour $i = 1, \dots, n$ et $k = 1, \dots, d_j$, on a

$$x_{i,j,k}^B = \begin{cases} 1 & \text{si } x_{i,j} \in I_{j,k} = \left(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right], \text{ (interquantiles)} \\ 0 & \text{sinon} \end{cases}$$

$d_1 = 4$

$X_{\bullet,1}$		$q_1(0.25) = 7,9$	$q_1(0.5) = 9,6$	$q_1(0.75) = 11,7$	
$\vdots$					
$i \to 12,5$	$\to$	0	0	0	1
$9,2$					
$3,1$					
$8,7$					
$\vdots$					

Le contexte

- Apprentissage supervisé $(x_i, y_i)_{i=1,\dots,n}$ avec x_i continues
- Encodage one-hot [22] : $x_i = (x_{i,1}, \dots, x_{i,p})^\top \in \mathbb{R}^p$ transformé en
$x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_1}^B, x_{i,2,1}^B, \dots, x_{i,2,d_2}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_p}^B)^\top \in \mathbb{R}^d$,
t.q. pour $i = 1, \dots, n$ et $k = 1, \dots, d_j$, on a

$$x_{i,j,k}^B = \begin{cases} 1 & \text{si } x_{i,j} \in I_{j,k} = \left(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right], \text{ (interquantiles)} \\ 0 & \text{sinon} \end{cases}$$

$d_1 = 4$

$X_{\bullet,1}$		$q_1(0.25) = 7,9$	$q_1(0.5) = 9,6$	$q_1(0.75) = 11,7$	
$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i \to 12,5$	$\to$	0	0	0	1
$9,2$	$\to$	0	1	0	0
$3,1$	$\to$	1	0	0	0
$8,7$	$\to$	0	1	0	0
$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$

Le contexte

- Apprentissage supervisé $(x_i, y_i)_{i=1,\dots,n}$ avec x_i continues
- Encodage one-hot [22] : $x_i = (x_{i,1}, \dots, x_{i,p})^\top \in \mathbb{R}^p$ transformé en $x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_1}^B, x_{i,2,1}^B, \dots, x_{i,2,d_2}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_p}^B)^\top \in \mathbb{R}^d$, t.q. pour $i = 1, \dots, n$ et $k = 1, \dots, d_j$, on a

$$x_{i,j,k}^B = \begin{cases} 1 & \text{si } x_{i,j} \in I_{j,k} = \left(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right], \text{ (interquantiles)} \\ 0 & \text{sinon} \end{cases}$$

- Risque empirique $R_n(\beta) = n^{-1} \sum_{i=1}^n \ell(y_i, m_\beta(x_i))$ où $m_\beta(x_i) = \beta^\top x_i^B$ et $\beta \in \mathbb{R}^d$, avec $d = \sum_{j=1}^p d_j$

Le contexte

- Apprentissage supervisé $(x_i, y_i)_{i=1,\dots,n}$ avec x_i continues
- Encodage one-hot [22] : $x_i = (x_{i,1}, \dots, x_{i,p})^\top \in \mathbb{R}^p$ transformé en
  $x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_1}^B, x_{i,2,1}^B, \dots, x_{i,2,d_2}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_p}^B)^\top \in \mathbb{R}^d$,
  t.q. pour $i = 1, \dots, n$ et $k = 1, \dots, d_j$, on a

$$x_{i,j,k}^B = \begin{cases} 1 & \text{si } x_{i,j} \in I_{j,k} = \left(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right], \text{ (interquantiles)} \\ 0 & \text{sinon} \end{cases}$$

- Risque empirique $R_n(\beta) = n^{-1} \sum_{i=1}^n \ell(y_i, m_\beta(x_i))$ où $m_\beta(x_i) = \beta^\top x_i^B$ et $\beta \in \mathbb{R}^d$, avec $d = \sum_{j=1}^p d_j$
- GLM [8] : $Y|X = x \sim$ famille exp à 1 paramètre

$$y|x \mapsto \exp\left(\frac{y m^0(x) - b\big(m^0(x)\big)}{\phi} + c(y, \phi)\right),$$

et $\ell\big(y_1, y_2\big) = -y_1 y_2 + b(y_2)$

La méthode

Problèmes

- (P1) $\forall j = 1, \ldots, p, \sum_{k=1}^{d_j} x_{i,j,k}^B = 1 \Rightarrow X^B$ pas de rang plein

La méthode

Problèmes

- (P1) $\forall j = 1, \ldots, p, \sum_{k=1}^{d_j} x^B_{i,j,k} = 1 \Rightarrow X^B$ pas de rang plein
- (P2) Choix des d_j? Valeurs élevées $\Rightarrow$ sur-apprentissage

La méthode

Problèmes

- (P1) $\forall j = 1, \ldots, p, \sum_{k=1}^{d_j} x^B_{i,j,k} = 1 \Rightarrow X^B$ pas de rang plein
- (P2) Choix des d_j? Valeurs élevées $\Rightarrow$ sur-apprentissage
- (P3) Sélection de variable? (sparsité par bloc dans β)

La méthode

Problèmes

- (P1) $\forall j = 1, \ldots, p, \sum_{k=1}^{d_j} x^B_{i,j,k} = 1 \Rightarrow X^B$ pas de rang plein
- (P2) Choix des d_j? Valeurs élevées $\Rightarrow$ sur-apprentissage
- (P3) Sélection de variable? (sparsité par bloc dans β)

Réponses

- (P1) $\rightarrow$ imposer $\sum_{k=1}^{d_j} \beta_{j,k} = 0$

La méthode

Problèmes

- (P1) $\forall j = 1, \ldots, p, \sum_{k=1}^{d_j} x^B_{i,j,k} = 1 \Rightarrow X^B$ pas de rang plein
- (P2) Choix des d_j? Valeurs élevées $\Rightarrow$ sur-apprentissage
- (P3) Sélection de variable? (sparsité par bloc dans β)

Réponses

- (P1) $\rightarrow$ imposer $\sum_{k=1}^{d_j} \beta_{j,k} = 0$
- (P2) $\rightarrow$ pénalisation TV par bloc $\sum_{j=1}^{p} \|\beta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}$,

$$\text{où } \|\beta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} = \sum_{k=2}^{d_j} \hat{w}_{j,k} |\beta_{j,k} - \beta_{j,k-1}|$$

La méthode

Problèmes

- (P1) $\forall j = 1, \ldots, p, \sum_{k=1}^{d_j} x^B_{i,j,k} = 1 \Rightarrow X^B$ pas de rang plein
- (P2) Choix des d_j? Valeurs élevées $\Rightarrow$ sur-apprentissage
- (P3) Sélection de variable? (sparsité par bloc dans β)

Réponses

- (P1) $\rightarrow$ imposer $\sum_{k=1}^{d_j} \beta_{j,k} = 0$
- (P2) $\rightarrow$ pénalisation TV par bloc $\sum_{j=1}^{p} \|\beta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}$
- (P3) $\rightarrow$ réponses à (P1) et (P2)!

La méthode

Problèmes

- (P1) $\forall j = 1, \ldots, p, \sum_{k=1}^{d_j} x^B_{i,j,k} = 1 \Rightarrow X^B$ pas de rang plein
- (P2) Choix des d_j? Valeurs élevées $\Rightarrow$ sur-apprentissage
- (P3) Sélection de variable? (sparsité par bloc dans β)

Réponses

- (P1) $\rightarrow$ imposer $\sum_{k=1}^{d_j} \beta_{j,k} = 0$
- (P2) $\rightarrow$ pénalisation TV par bloc $\sum_{j=1}^{p} \|\beta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}$
- (P3) $\rightarrow$ réponses à (P1) et (P2)!

Pénalité binarsity

- $\mathrm{bina}(\beta) = \sum_{j=1}^{p} \left(\sum_{k=2}^{d_j} \hat{w}_{j,k} |\beta_{j,k} - \beta_{j,k-1}| + \delta_1(\beta_{j,\bullet}) \right)$,

$$\text{où } \delta_1(u) = \begin{cases} 0 & \text{si} \quad \mathbf{1}^\top u = 0, \\ \infty & \text{sinon} \end{cases}$$

La méthode

Problèmes

- (P1) $\forall j = 1, \dots, p, \sum_{k=1}^{d_j} x^B_{i,j,k} = 1 \Rightarrow X^B$ pas de rang plein
- (P2) Choix des d_j? Valeurs élevées $\Rightarrow$ sur-apprentissage
- (P3) Sélection de variable? (sparsité par bloc dans β)

Réponses

- (P1) $\rightarrow$ imposer $\sum_{k=1}^{d_j} \beta_{j,k} = 0$
- (P2) $\rightarrow$ pénalisation TV par bloc $\sum_{j=1}^{p} \|\beta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}$
- (P3) $\rightarrow$ réponses à (P1) et (P2)!

Pénalité binarsity

- $\mathrm{bina}(\beta) = \sum_{j=1}^{p} \left(\sum_{k=2}^{d_j} \hat{w}_{j,k} |\beta_{j,k} - \beta_{j,k-1}| + \delta_1(\beta_{j,\bullet}) \right)$
- Problème d'optimisation

$$\hat{\beta} \in \mathrm{argmin}_{\beta \in \mathbb{R}^d} \left\{ R_n(\beta) + \mathrm{bina}(\beta) \right\}$$

Résultat théorique

- Fonction de risque associée [21]

$$R(m_\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{ - b'\left(m^0(X_i)\right) m_\beta(X_i) + b\left(m_\beta(X_i)\right)\right\}$$

Résultat théorique

- Fonction de risque associée [21]

$$R(m_\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{ - b'\left(m^0(X_i)\right) m_\beta(X_i) + b\left(m_\beta(X_i)\right)\right\}$$

- $J(\beta) = \left[J_1(\beta), \ldots, J_p(\beta)\right]$, avec

$$J_j(\beta) = \left\{k : \beta_{j,k} \neq \beta_{j,k-1}, \text{ for } k = 2, \ldots, d_j\right\}$$

Résultat théorique

- Fonction de risque associée [21]

$$R(m_\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{ - b'\left(m^0(X_i)\right) m_\beta(X_i) + b\left(m_\beta(X_i)\right)\right\}$$

- $J(\beta) = \left[J_1(\beta), \ldots, J_p(\beta)\right]$, avec

$$J_j(\beta) = \left\{k : \beta_{j,k} \neq \beta_{j,k-1}, \text{ for } k = 2, \ldots, d_j\right\}$$

- $\mathscr{B}_d(R) = \{\beta \in \mathbb{R}^d : \|\beta\|_2 \leq R\}$, avec $R > 0$

Résultat théorique

- Fonction de risque associée [21]

$$R(m_\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{ - b'\left(m^0(X_i)\right) m_\beta(X_i) + b\left(m_\beta(X_i)\right)\right\}$$

- $J(\beta) = \left[J_1(\beta), \ldots, J_p(\beta)\right]$, avec

$$J_j(\beta) = \left\{k : \beta_{j,k} \neq \beta_{j,k-1}, \text{ for } k = 2, \ldots, d_j\right\}$$

- $\mathscr{B}_d(R) = \{\beta \in \mathbb{R}^d : \|\beta\|_2 \leq R\}$, avec $R > 0$

Inégalité oracle non-asymptotique à vitesse rapide

Avec grande probabilité, on a

$$R(m_{\hat{\theta}}) - R(m^0) \leq (1+c_1) \inf_{\substack{\theta \in \mathscr{B}_d(R) \\ \forall j,\ \mathbf{1}^\top \theta_{j,\bullet} = 0 \\ |J(\theta)| \leq J^*}} \left\{ R(m_\theta) - R(m^0) + c_2 \frac{|J(\theta)| \log d}{n} \right\}$$

avec $c_1, c_2 > 0$

Illustration de binarsity

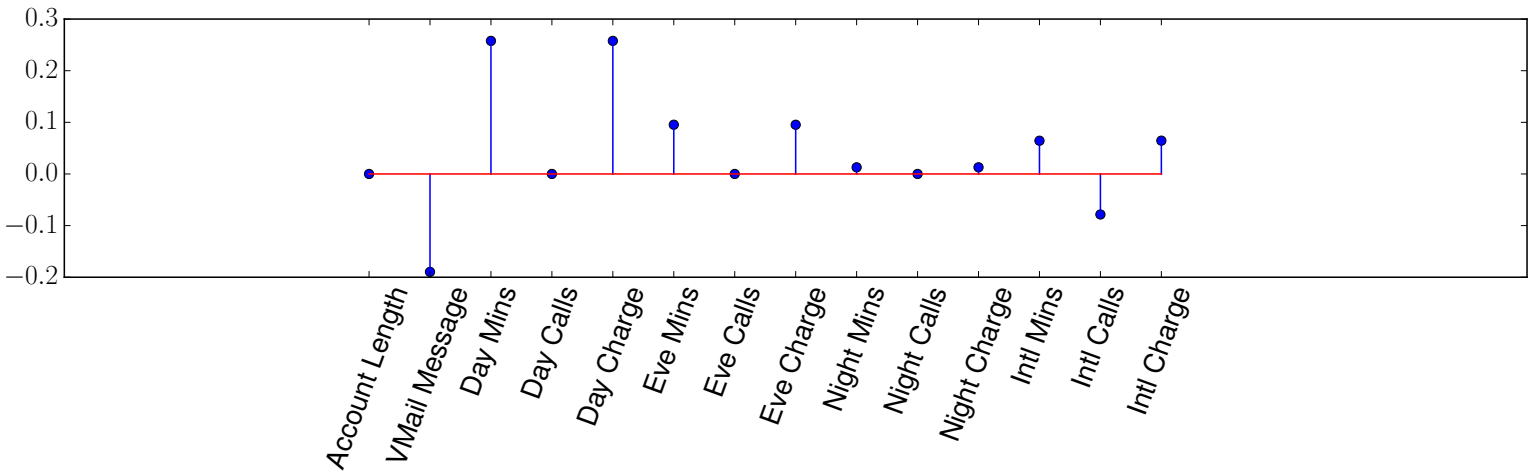


Figure 4: LR sur "Churn" ($n = 3333, p = 14$)

Illustration de binarsity

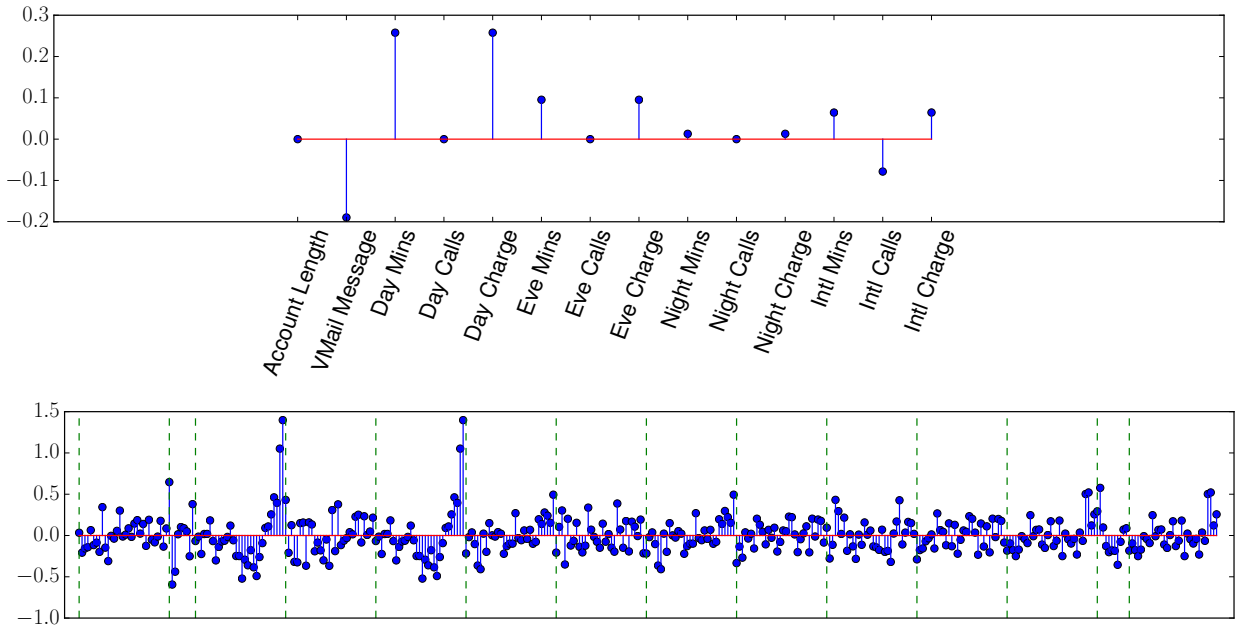


Figure 4: LR sur "Churn" ($n = 3333, p = 14$)

Illustration de binarsity

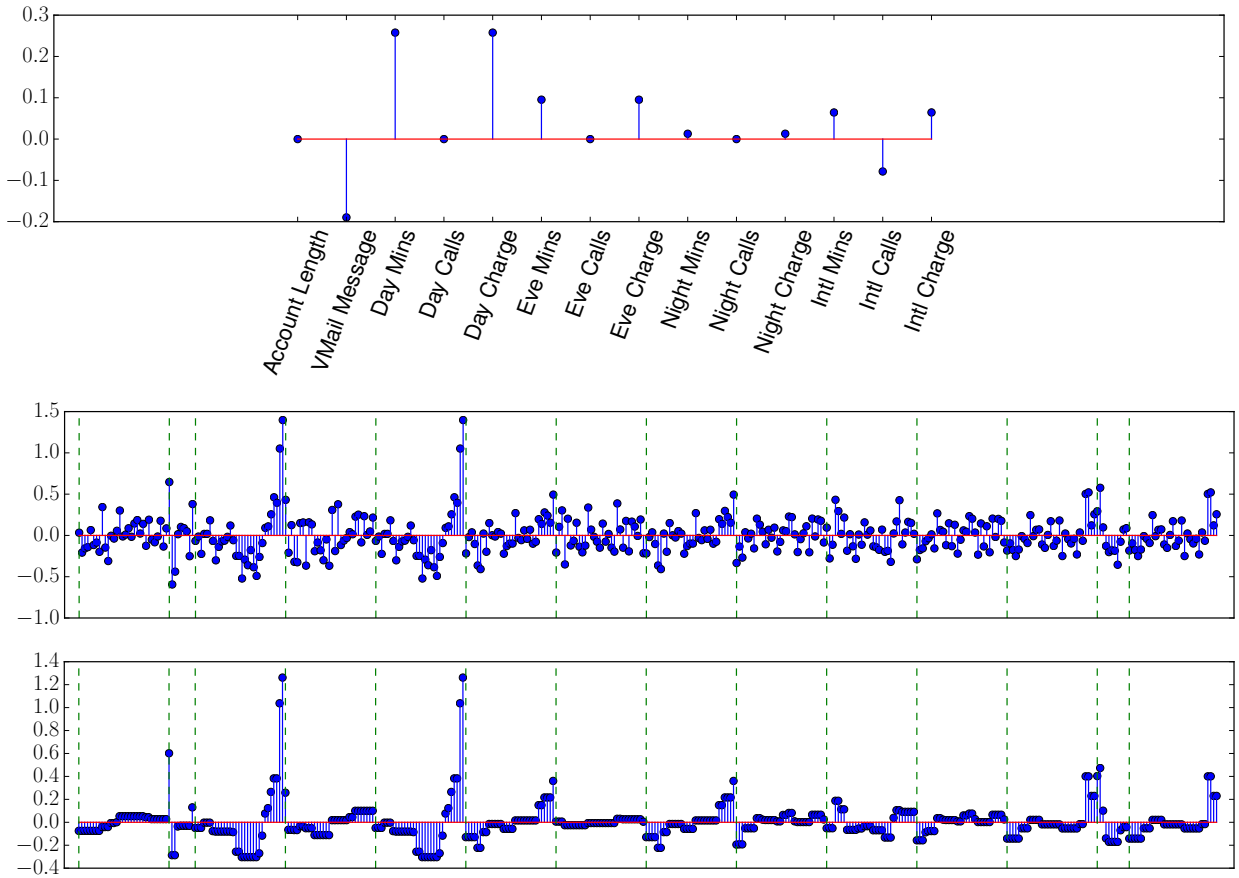


Figure 4: LR sur "Churn" ($n = 3333, p = 14$)

Illustration de binarsity

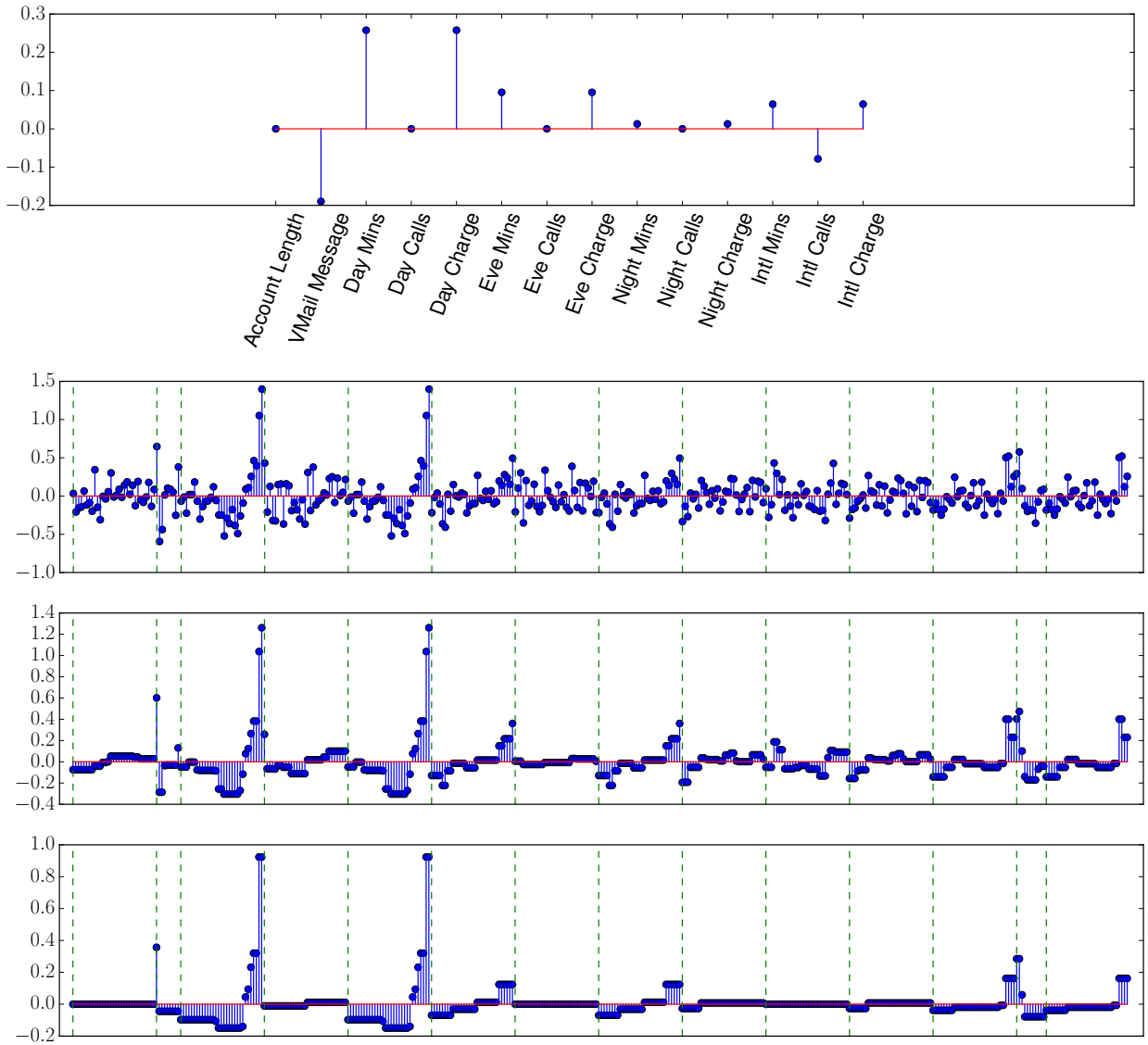
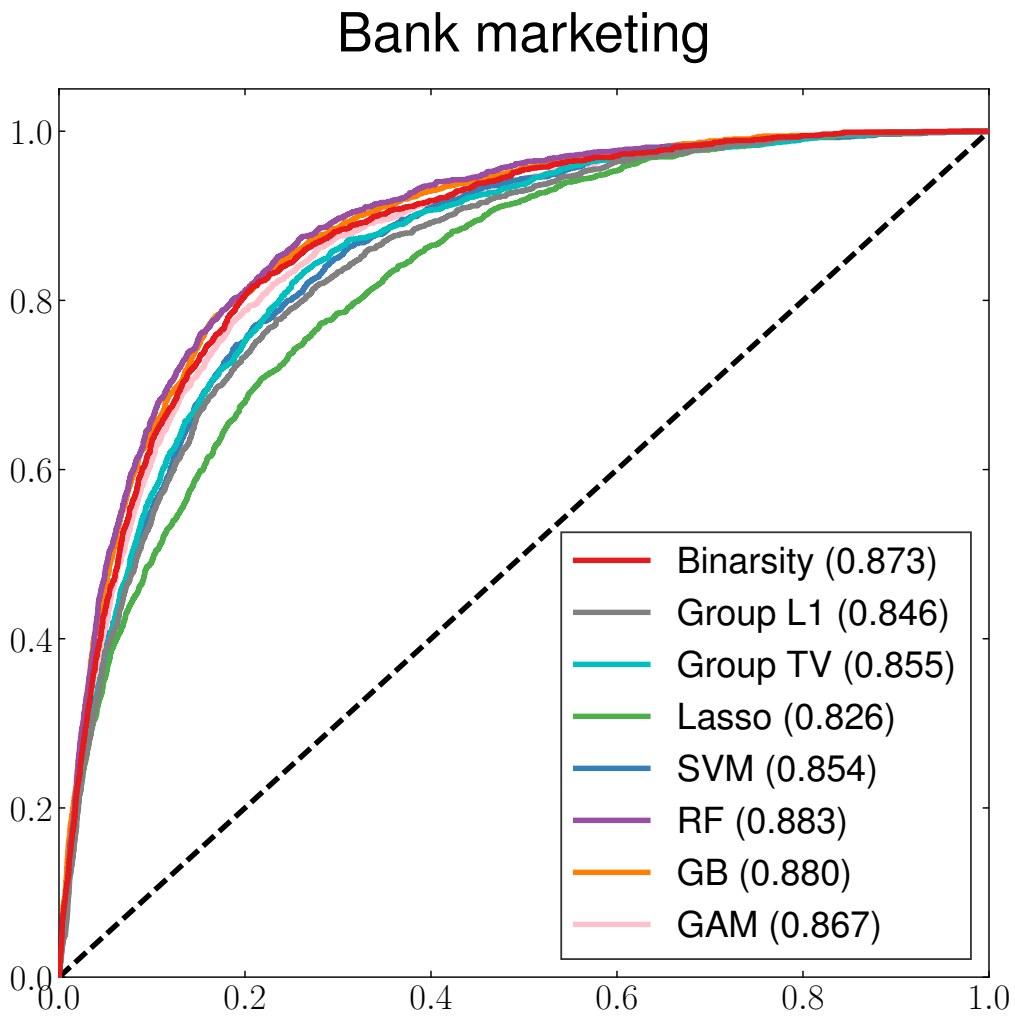


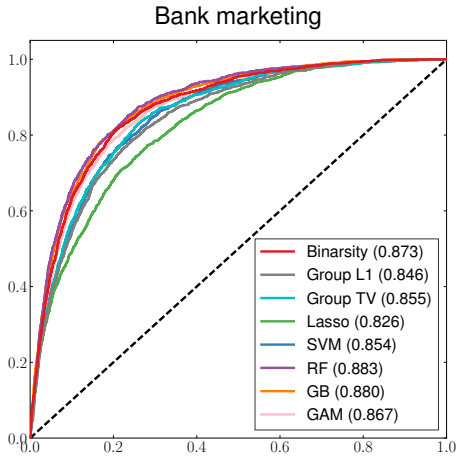
Figure 4: LR sur "Churn" ($n = 3333, p = 14$)

Applications

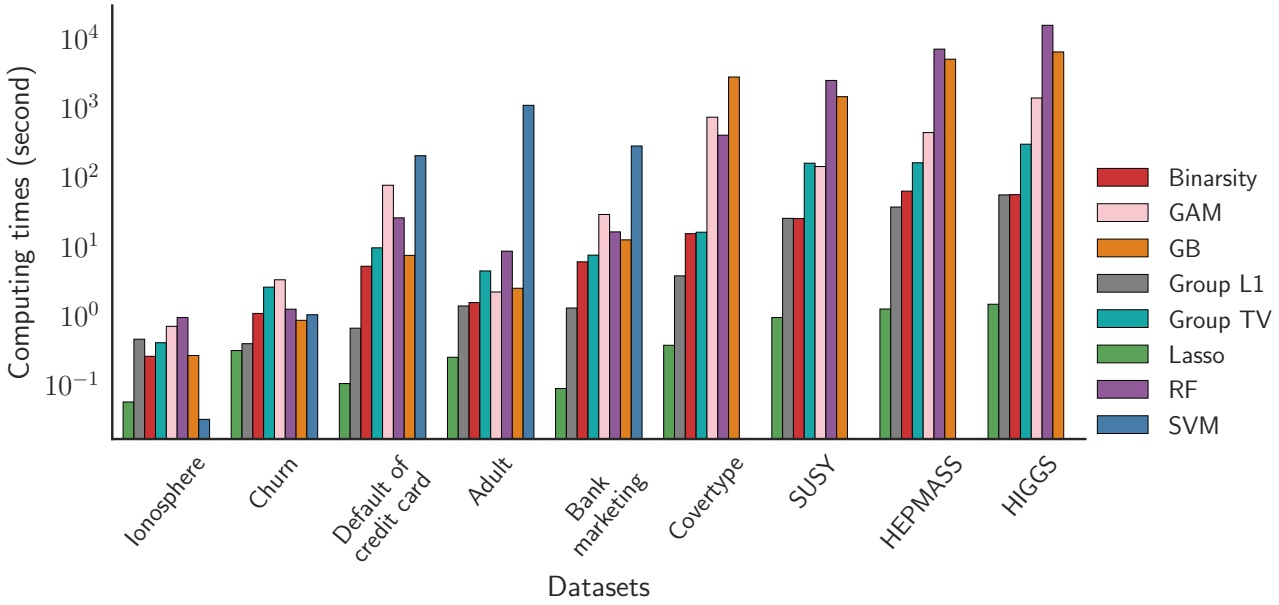


(a) Courbes ROC

Applications



(a) Courbes ROC



(b) Temps de calcul

Conclusion

- Nouvelle pénalisation pour l'encodage "one-hot" de covariables continues

Conclusion

- Nouvelle pénalisation pour l'encodage "one-hot" de covariables continues
- Bonnes propriétés théoriques

Conclusion

- Nouvelle pénalisation pour l'encodage "one-hot" de covariables continues
- Bonnes propriétés théoriques
- Bonnes performances sur données réelles et simulées

Conclusion

- Nouvelle pénalisation pour l'encodage "one-hot" de covariables continues
- Bonnes propriétés théoriques
- Bonnes performances sur données réelles et simulées
- Et surtout $\rightarrow$ interprétabilité!

Conclusion

- Nouvelle pénalisation pour l'encodage "one-hot" de covariables continues
- Bonnes propriétés théoriques
- Bonnes performances sur données réelles et simulées
- Et surtout $\rightarrow$ interprétabilité!

Article associé

M.Z. Alaya, **S. Bussy**, S. Gaïffas, A. Guilloux
Binarsity: a penalization for one-hot encoded features
Publié dans *Journal of Machine Learning Research*, 2019.

Conclusion

- Nouvelle pénalisation pour l'encodage "one-hot" de covariables continues
- Bonnes propriétés théoriques
- Bonnes performances sur données réelles et simulées
- Et surtout $\rightarrow$ interprétabilité!

Article associé

M.Z. Alaya, **S. Bussy**, S. Gaïffas, A. Guilloux
Binarsity: a penalization for one-hot encoded features
Publié dans *Journal of Machine Learning Research*, 2019.

Code `Python/C++`

- Disponible à `https://github.com/SimonBussy/binarsity`

Conclusion

- Nouvelle pénalisation pour l'encodage "one-hot" de covariables continues
- Bonnes propriétés théoriques
- Bonnes performances sur données réelles et simulées
- Et surtout $\rightarrow$ interprétabilité!

Article associé

M.Z. Alaya, **S. Bussy**, S. Gaïffas, A. Guilloux
Binarsity: a penalization for one-hot encoded features
Publié dans *Journal of Machine Learning Research*, 2019.

Code `Python/C++`

- Disponible à `https://github.com/SimonBussy/binarsity`
- Library `tick` : `https://github.com/X-DataInitiative/tick`

V. Binacox, automatic cut-points detection in a high-dimensional Cox model

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]
- Binacox : méthode pronostique pour

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]
- Binacox : méthode pronostique pour
  - détecter de **multiples** seuils par covariable continue,

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]
- Binacox : méthode pronostique pour
  - détecter de **multiples** seuils par covariable continue,
  - de façon **multivariée**,

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]
- Binacox : méthode pronostique pour
  - détecter de **multiples** seuils par covariable continue,
  - de façon **multivariée**,
  - dans un contexte de **grande dimension**

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]
- Binacox : méthode pronostique pour
  - détecter de **multiples** seuils par covariable continue,
  - de façon **multivariée**,
  - dans un contexte de **grande dimension**
- $Z = T \wedge C, \quad \Delta = \mathbb{1}(T \leq C)$ et $X \in \mathbb{R}^p$

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]
- Binacox : méthode pronostique pour
  - détecter de **multiples** seuils par covariable continue,
  - de façon **multivariée**,
  - dans un contexte de **grande dimension**
- $Z = T \wedge C, \quad \Delta = \mathbb{1}(T \leq C) \quad$ et $\quad X \in \mathbb{R}^p$
- Risque instantané pour un patient i donné par

$$\lambda^\star(t|X_i = x_i) = \lambda_0^\star(t)\exp\underbrace{\left\{ \sum_{j=1}^{p} \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \mathbb{1}(x_{i,j} \in I_{j,k}^\star) \right\}}_{f^\star(x_i)}$$

où $I_{j,k}^\star = (\mu_{j,k-1}^\star, \mu_{j,k}^\star]$ pour $k \in \{1, \ldots, K_j^\star + 1\}$

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]
- Binacox : méthode pronostique pour
  - détecter de **multiples** seuils par covariable continue,
  - de façon **multivariée**,
  - dans un contexte de **grande dimension**
- $Z = T \wedge C$, $\Delta = \mathbb{1}(T \leq C)$ et $X \in \mathbb{R}^p$
- Risque instantané pour un patient i donné par

$$\lambda^\star(t|X_i = x_i) = \lambda_0^\star(t)\exp\underbrace{\left\{ \sum_{j=1}^{p} \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \mathbb{1}(x_{i,j} \in I_{j,k}^\star) \right\}}_{f^\star(x_i)}$$

  où $I_{j,k}^\star = (\mu_{j,k-1}^\star, \mu_{j,k}^\star]$ pour $k \in \{1, \ldots, K_j^\star + 1\}$
- But : estimer simultanément

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]
- Binacox : méthode pronostique pour
  - détecter de **multiples** seuils par covariable continue,
  - de façon **multivariée**,
  - dans un contexte de **grande dimension**
- $Z = T \wedge C, \quad \Delta = \mathbb{1}(T \leq C)$ et $X \in \mathbb{R}^p$
- Risque instantané pour un patient i donné par

$$\lambda^\star(t|X_i = x_i) = \lambda_0^\star(t)\exp\Big\{\underbrace{\sum_{j=1}^{p}\sum_{k=1}^{K_j^\star+1}\beta_{j,k}^\star \mathbb{1}(x_{i,j} \in I_{j,k}^\star)}_{f^\star(x_i)}\Big\}$$

  où $I_{j,k}^\star = (\mu_{j,k-1}^\star, \mu_{j,k}^\star]$ pour $k \in \{1, \ldots, K_j^\star + 1\}$
- But : estimer simultanément
  - $\mu^\star = (\mu_{1,1}^\star, \ldots, \mu_{1,K_1^\star}^\star, \ldots, \mu_{p,1}^\star, \ldots, \mu_{p,K_p^\star}^\star)^\top \in \mathbb{R}^{K^\star}$

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]
- Binacox : méthode pronostique pour
  - détecter de **multiples** seuils par covariable continue,
  - de façon **multivariée**,
  - dans un contexte de **grande dimension**
- $Z = T \wedge C, \quad \Delta = \mathbb{1}(T \leq C) \quad$ et $\quad X \in \mathbb{R}^p$
- Risque instantané pour un patient i donné par

$$\lambda^\star(t|X_i = x_i) = \lambda_0^\star(t)\exp\Big\{\underbrace{\sum_{j=1}^{p}\sum_{k=1}^{K_j^\star+1}\beta_{j,k}^\star \mathbb{1}(x_{i,j} \in I_{j,k}^\star)}_{f^\star(x_i)}\Big\}$$

  où $I_{j,k}^\star = (\mu_{j,k-1}^\star, \mu_{j,k}^\star]$ pour $k \in \{1, \ldots, K_j^\star + 1\}$
- But : estimer simultanément
  - $\mu^\star = (\mu_{1,1}^\star, \ldots, \mu_{1,K_1^\star}^\star, \ldots, \mu_{p,1}^\star, \ldots, \mu_{p,K_p^\star}^\star)^\top \in \mathbb{R}^{K^\star}$
  - $\beta^\star = (\beta_{1,1}^\star, \ldots, \beta_{1,K_1^\star+1}^\star, \ldots, \beta_{p,1}^\star, \ldots, \beta_{p,K_p^\star+1}^\star)^\top \in \mathbb{R}^{K^\star+p}$

Le contexte

- Biomarqueurs $\rightarrow$ décision clinique : choix de seuils
- Méthodes actuelles basées sur les tests multiples (MT) [15]
- Binacox : méthode pronostique pour
  - détecter de **multiples** seuils par covariable continue,
  - de façon **multivariée**,
  - dans un contexte de **grande dimension**
- $Z = T \wedge C, \quad \Delta = \mathbb{1}(T \leq C) \quad$ et $\quad X \in \mathbb{R}^p$
- Risque instantané pour un patient i donné par

$$\lambda^\star(t|X_i = x_i) = \lambda_0^\star(t)\exp\Big\{\underbrace{\sum_{j=1}^{p}\sum_{k=1}^{K_j^\star+1}\beta_{j,k}^\star \mathbb{1}(x_{i,j} \in I_{j,k}^\star)}_{f^\star(x_i)}\Big\}$$

  où $I_{j,k}^\star = (\mu_{j,k-1}^\star, \mu_{j,k}^\star]$ pour $k \in \{1, \ldots, K_j^\star + 1\}$
- But : estimer simultanément
  - $\mu^\star = (\mu_{1,1}^\star, \ldots, \mu_{1,K_1^\star}^\star, \ldots, \mu_{p,1}^\star, \ldots, \mu_{p,K_p^\star}^\star)^\top \in \mathbb{R}^{K^\star}$
  - $\beta^\star = (\beta_{1,1}^\star, \ldots, \beta_{1,K_1^\star+1}^\star, \ldots, \beta_{p,1}^\star, \ldots, \beta_{p,K_p^\star+1}^\star)^\top \in \mathbb{R}^{K^\star+p}$
  - $K^\star = \sum_{j=1}^p K_j^\star$

La méthode

- Binarisation par encodage "one-hot" $x_i^B \in \mathbb{R}^{p+d}$

$$\text{t.q. } x_i^B = (x_{i,1,1}^B, \ldots, x_{i,1,d_1+1}^B, \ldots, x_{i,p,1}^B, \ldots, x_{i,p,d_p+1}^B)^\top,$$

où

$$x_{i,j,l}^B = \begin{cases} 1 & \text{si } x_{i,j} \in I_{j,l}, \\ 0 & \text{sinon,} \end{cases}$$

et $I_{j,l} = (\mu_{j,l-1}, \mu_{j,l}]$ interquantiles avec $\mu_{j,l} = q_j(l/d_j + 1)$

La méthode

- Binarisation par encodage "one-hot" $x_i^B \in \mathbb{R}^{p+d}$
- $f_\beta : x_i \mapsto \beta^\top x_i^B = \sum_{j=1}^{p} f_{\beta_j,\bullet}(x_i)$ pour estimer $f^\star$

$$\text{où } f_{\beta_j,\bullet}(x_i) = \sum_{l=1}^{d_j+1} \beta_{j,l} \mathbb{1}(x_{i,j} \in I_{j,l})$$

La méthode

- Binarisation par encodage "one-hot" $x_i^B \in \mathbb{R}^{p+d}$
- $f_\beta : x_i \mapsto \beta^\top x_i^B = \sum_{j=1}^{p} f_{\beta_j,\bullet}(x_i)$ pour estimer $f^\star$
- log-vraissemblance négative partielle $\ell_n(f_\beta)$

$$\ell_n(f_\beta) = -\frac{1}{n}\sum_{i=1}^{n} \delta_i \left\{ f_\beta(x_i) - \log \sum_{i' : z_{i'} \geq z_i} e^{f_\beta(x_{i'})} \right\}$$

La méthode

- Binarisation par encodage "one-hot" $x_i^B \in \mathbb{R}^{p+d}$
- $f_\beta : x_i \mapsto \beta^\top x_i^B = \sum_{j=1}^{p} f_{\beta_j,\bullet}(x_i)$ pour estimer $f^\star$
- log-vraissemblance négative partielle $\ell_n(f_\beta)$
- $\hat{\beta} \in \text{argmin}_{\beta \in \mathscr{B}_{p+d}(R)} \left\{ \ell_n(f_\beta) + \text{bina}(\beta) \right\}$

La méthode

- Binarisation par encodage "one-hot" $x_i^B \in \mathbb{R}^{p+d}$
- $f_\beta : x_i \mapsto \beta^\top x_i^B = \sum_{j=1}^{p} f_{\beta_j,\bullet}(x_i)$ pour estimer $f^\star$
- log-vraissemblance négative partielle $\ell_n(f_\beta)$
- $\hat{\beta} \in \text{argmin}_{\beta \in \mathscr{B}_{p+d}(R)} \left\{ \ell_n(f_\beta) + \text{bina}(\beta) \right\}$
- $J_j(\hat{\beta}) = \left\{ l : \beta_{j,l} \neq \beta_{j,l-1}, \text{pour } l = 2, \ldots, d_j + 1 \right\} = \{\hat{l}_{j,1}, \ldots, \hat{l}_{j,s_j}\}$

La méthode

- Binarisation par encodage "one-hot" $x_i^B \in \mathbb{R}^{p+d}$
- $f_\beta : x_i \mapsto \beta^\top x_i^B = \sum_{j=1}^p f_{\beta_j,\bullet}(x_i)$ pour estimer $f^\star$
- log-vraissemblance négative partielle $\ell_n(f_\beta)$
- $\hat{\beta} \in \text{argmin}_{\beta \in \mathscr{B}_{p+d}(R)} \left\{ \ell_n(f_\beta) + \text{bina}(\beta) \right\}$
- $J_j(\hat{\beta}) = \left\{ l : \beta_{j,l} \neq \beta_{j,l-1}, \text{pour } l = 2, \ldots, d_j + 1 \right\} = \{\hat{l}_{j,1}, \ldots, \hat{l}_{j,s_j}\}$
- $\widehat{\mu}_{j,\bullet} = (\mu_{j,\hat{l}_{j,1}}, \ldots, \mu_{j,\hat{l}_{j,s_j}})^\top$ avec $s_j = |J_j(\hat{\beta})| = \widehat{K}_j$

La méthode

- Binarisation par encodage "one-hot" $x_i^B \in \mathbb{R}^{p+d}$
- $f_\beta : x_i \mapsto \beta^\top x_i^B = \sum_{j=1}^p f_{\beta_{j,\bullet}}(x_i)$ pour estimer $f^\star$
- log-vraissemblance négative partielle $\ell_n(f_\beta)$
- $\hat{\beta} \in \text{argmin}_{\beta \in \mathscr{B}_{p+d}(R)} \left\{ \ell_n(f_\beta) + \text{bina}(\beta) \right\}$
- $J_j(\hat{\beta}) = \left\{ l : \beta_{j,l} \neq \beta_{j,l-1}, \text{pour } l = 2, \ldots, d_j + 1 \right\} = \{\hat{l}_{j,1}, \ldots, \hat{l}_{j,s_j}\}$
- $\widehat{\mu}_{j,\bullet} = (\mu_{j,\hat{l}_{j,1}}, \ldots, \mu_{j,\hat{l}_{j,s_j}})^\top$ avec $s_j = |J_j(\hat{\beta})| = \widehat{K}_j$

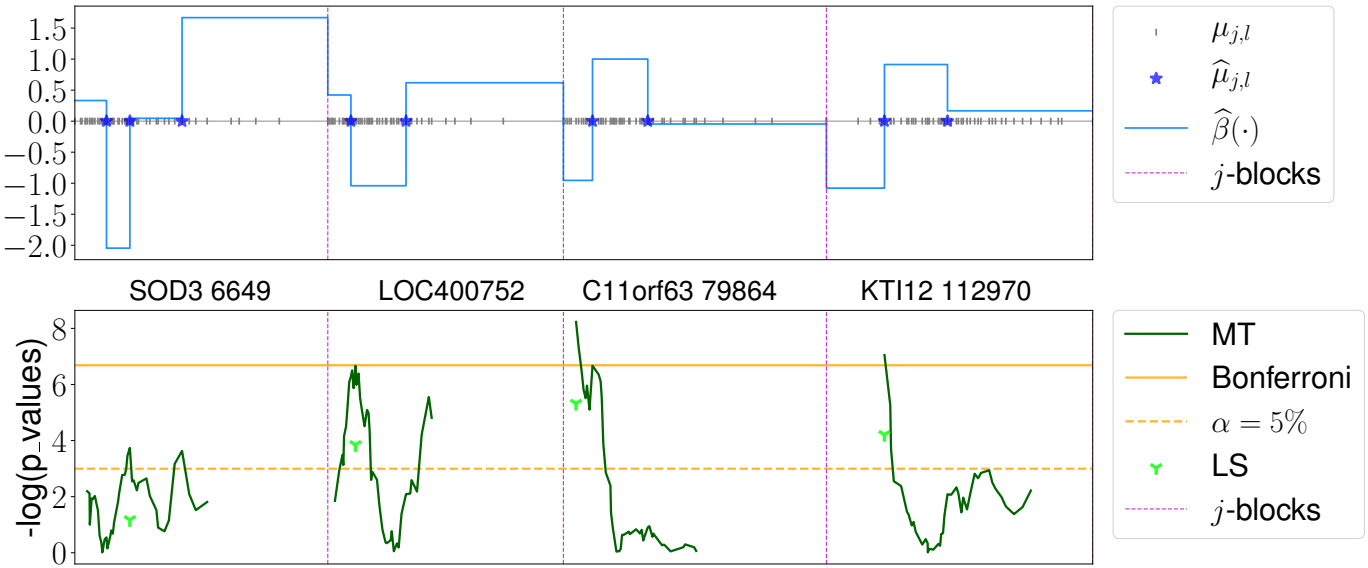
Inégalité oracle non-asymptotique à vitesse rapide

Avec grande probabilité, on a

$$KL_n(f^\star, f_{\hat{\beta}}) \leq (1 + c_1) \inf_{\substack{\beta \in \mathscr{B}_{p+d}(R) \\ |J(\beta)| \leq K^\star \\ \forall j, \mathbf{1}^\top \beta_{j,\bullet} = 0}} \left\{ KL_n(f^\star, f_\beta) + c_2 |J(\beta)| \frac{\log(p+d)}{n} \right\}$$

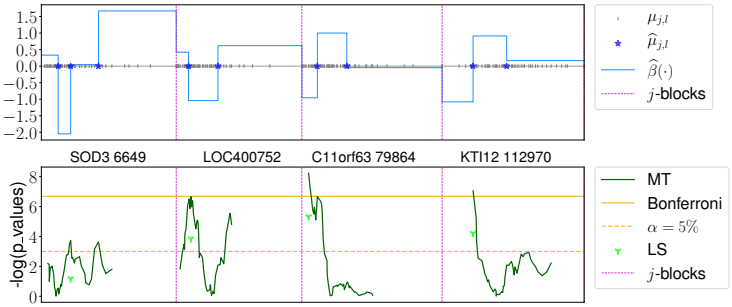
avec $c_1, c_2 > 0$

Applications sur données réelles (TCGA)

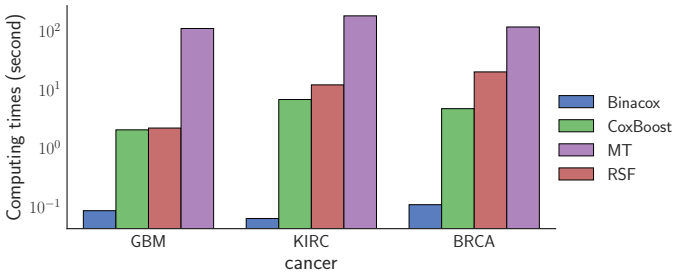


(a) Résultats sur le cancer GBM ($n = 168, p = 20531$)

Applications sur données réelles (TCGA)

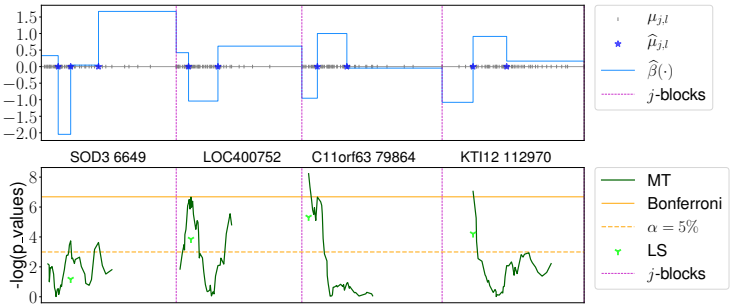


(a) Résultats sur le cancer GBM

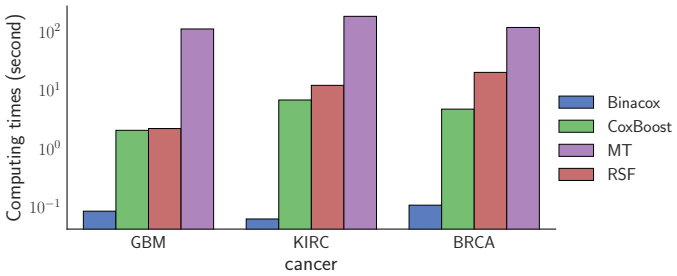


(b) Temps de calcul

Applications sur données réelles (TCGA)



(a) Résultats sur le cancer GBM



(b) Temps de calcul

Table 3: Comparison des C-index

Cancer	Continuous	Binacox	MT-B	MT-LS	CoxBoost	RSF
GBM ($n = 168$)	0.660	**0.806**	0.753	0.768	0.684	0.691
KIRC ($n = 605$)	0.682	**0.727**	0.663	0.663	0.679	0.686
BRCA ($n = 1211$)	0.713	**0.849**	0.741	0.738	0.723	0.746

Conclusion

- Détection multivariée de plusieurs cut-points par covariable continue

Conclusion

- Détection multivariée de plusieurs cut-points par covariable continue
- Implémentation rapide et bons résultats obtenus en estimation sur données simulées et réelles

Conclusion

- Détection multivariée de plusieurs cut-points par covariable continue
- Implémentation rapide et bons résultats obtenus en estimation sur données simulées et réelles

Article associé

S. Bussy, M.Z. Alaya, A. Guilloux et A.S. Jannot
Binacox: automatic cut-points detection in high-dimensional Cox model, with applications to genetic data
En revision dans *Biometrics*, 2020.

Conclusion

- Détection multivariée de plusieurs cut-points par covariable continue
- Implémentation rapide et bons résultats obtenus en estimation sur données simulées et réelles

Article associé

S. Bussy, M.Z. Alaya, A. Guilloux et A.S. Jannot
Binacox: automatic cut-points detection in high-dimensional Cox model, with applications to genetic data
En revision dans *Biometrics*, 2020.

Code `Python/C++`

Disponible à `https://github.com/SimonBussy/binacox`

VI. Conclusion générale

Conclusion générale

Nouvelles méthodes

- Différentes méthodes de machine learning proposées

Conclusion générale

Nouvelles méthodes

- Différentes méthodes de machine learning proposées
- Contexte de grande dimension

Conclusion générale

Nouvelles méthodes

- Différentes méthodes de machine learning proposées
- Contexte de grande dimension
- Interprétabilité des méthodes (sélection de variable, identification de sous-groupes, de seuils *etc.*)

Conclusion générale

Nouvelles méthodes

- Différentes méthodes de machine learning proposées
- Contexte de grande dimension
- Interprétabilité des méthodes (sélection de variable, identification de sous-groupes, de seuils *etc.*)
- Études théoriques et pratiques

Conclusion générale

Nouvelles méthodes

- Différentes méthodes de machine learning proposées
- Contexte de grande dimension
- Interprétabilité des méthodes (sélection de variable, identification de sous-groupes, de seuils *etc.*)
- Études théoriques et pratiques

Implémentation

Tous les codes ayant généré les résultats/figures de la thèse sont disponibles à https://github.com/SimonBussy

Références I

[1] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Simon Bussy, Agathe Guilloux, Stéphane Gaïffas, and Anne-Sophie Jannot. C-mix: A high-dimensional mixture model for censored durations, with applications to genetic data. *Statistical Methods in Medical Research*, 0(0):0962280218766389, 2018.

[4] Hung-Chia Chen, Ralph L Kodell, Kuang Fu Cheng, and James J Chen. Assessment of performance of survival prediction models for cancer prognosis. *BMC medical research methodology*, 12(1):102, 2012.

[5] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[6] Vern T Farewell. The use of mixture models for the analysis of sureval data with long-term survivors. *Biometrics*, 38(4):1041–1046, 1982.

[7] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38 (4):367–378, 2002.

[8] P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall, London, 1994.

[9] F. E. Harrell, K. L. Lee, and D. B. Mark. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387, 1996.

[10] T. Hastie and R. Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

[11] Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

[12] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[13] J. P. Klein and M. L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

Références II

[14] Lynn Kuo and Fengchun Peng. A mixture-model approach to the analysis of survival data. *Biostatistics-Basel-*, 5:255–272, 2000.

[15] B. Lausen and M. Schumacher. Maximally selected rank statistics. *Biometrics*, pages 73–85, 1992.

[16] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[17] Frédéric B Piel, Anand P Patil, Rosalind E Howes, Oscar A Nyangiri, Peter W Gething, Mewahyu Dewi, William H Temperley, Thomas N Williams, David J Weatherall, and Simon I Hay. Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. *The Lancet*, 381(9861):142–151, 2013.

[18] Marco Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. Modelling patient time-series data from electronic health records using gaussian processes. In *Advances in Neural Information Processing Systems: Workshop on Machine Learning for Clinical Data Analysis*, pages 1–4, 2013.

[19] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[20] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[21] S. van de Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, 36(2): 614–645, 2008.

[22] J. Wu and S. Coggeshall. *Foundations of Predictive Analytics (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 1st edition, 2012.

[23] B Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.