

# Soutenance de thèse

présentée le 16/01/2019 par

**Simon Bussy** <sup>(1)</sup>

devant le jury composé de

Pr.	Stéphane Gaïffas <sup>(1,2)</sup>	Examineur
Pr.	Agathe Guilloux <sup>(3)</sup>	Directrice de thèse
Dr.	Anne-Sophie Jannot <sup>(4,5)</sup>	Co-directrice de thèse
Pr.	Gregory Nuel <sup>(1)</sup>	Examineur
Dr.	Franck Picard <sup>(6)</sup>	Examineur
Dr.	Raphaël Porcher <sup>(7)</sup>	Rapporteur
Pr.	Rodolphe Thiebaut <sup>(8,9)</sup>	Examineur
Pr.	Jean-Philippe Vert <sup>(10,11)</sup>	Rapporteur

<sup>(1)</sup>LPSM, UMR 8001, Sorbonne Université, Paris, France. <sup>(2)</sup>CMAP, UMR 7641, École Polytechnique CNRS, Paris, France.

<sup>(3)</sup>LAMME, Université Evry, CNRS, Université Paris-Saclay, Paris, France. <sup>(4)</sup>APHP, Département d'Informatique Biomédicale et de Santé Publique, HEGP, Paris, France. <sup>(5)</sup>INSERM, UMRS 1138, Eq22, Centre de Recherche des Cordeliers, Université Paris Descartes, Paris, France. <sup>(6)</sup>LBBE, UMR CNRS 5558 Université Lyon 1, F-69622 Villeurbanne, France. <sup>(7)</sup>CRESS INSERM, UMR 1153, Université Paris Descartes, Paris, France. <sup>(8)</sup>SISTM, Inria Bordeaux Sud-Ouest, Epidémiologie et Biostatistique, Bordeaux, France. <sup>(9)</sup>INSERM, UMR 897, Université de Bordeaux, Bordeaux, France. <sup>(10)</sup>MINES ParisTech, Université de recherche PSL, CBIO-Centre for Computational Biology, Paris, France. <sup>(11)</sup>Google Brain, Paris, France.

# Introduction

Intitulé de la thèse :

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

## Introduction

### Trajectoires

Contexte  
Méthode  
Résultats

### Réadmissions

Contexte  
Méthode  
Résultats

### C-mix

Modèle  
Applications  
Conclusion

### Binarsity

Méthode  
Applications  
Conclusion

### Binacox

Méthode  
Applications  
Conclusion

### Conclusion

### Références

# Introduction

Intitulé de la thèse :

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

## Cadre

- ▶ Apprentissage supervisé

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Introduction

Intitulé de la thèse :

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

## Cadre

- ▶ Apprentissage supervisé
- ▶ Analyse de survie

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Introduction

Intitulé de la thèse :

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

## Cadre

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Introduction

Intitulé de la thèse :

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

## Cadre

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Introduction

Intitulé de la thèse :

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

## 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

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Introduction

Intitulé de la thèse :

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

## 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

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Introduction

Intitulé de la thèse :

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

### 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

#### Trajectoires

Contexte  
Méthode  
Résultats

#### Réadmissions

Contexte  
Méthode  
Résultats

#### C-mix

Modèle  
Applications  
Conclusion

#### Binarsity

Méthode  
Applications  
Conclusion

#### Binacox

Méthode  
Applications  
Conclusion

#### Conclusion

#### Références

# Introduction

Intitulé de la thèse :

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

### 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

#### Trajectoires

Contexte  
Méthode  
Résultats

#### Réadmissions

Contexte  
Méthode  
Résultats

#### C-mix

Modèle  
Applications  
Conclusion

#### Binarsity

Méthode  
Applications  
Conclusion

#### Binacox

Méthode  
Applications  
Conclusion

#### Conclusion

#### Références

# Introduction

Intitulé de la thèse :

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

### 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

#### Introduction

#### Trajectoires

Contexte  
Méthode  
Résultats

#### Réadmissions

Contexte  
Méthode  
Résultats

#### C-mix

Modèle  
Applications  
Conclusion

#### Binarsity

Méthode  
Applications  
Conclusion

#### Binacox

Méthode  
Applications  
Conclusion

#### Conclusion

#### Références

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

Modèle  
Applications  
Conclusion

Binarsity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

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

## Contexte

- ▶ Drépanocytose : maladie génétique la plus fréquente [17]

Introduction

Trajectoires

**Contexte**

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Contexte

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

Introduction

Trajectoires

**Contexte**

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Contexte

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

Introduction

Trajectoires

**Contexte**

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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

Trajectoires

**Contexte**

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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

Trajectoires

**Contexte**

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Données

- ▶ Patients admis à l'HEGP pour CVO entre 2010 → 2015

Introduction

Trajectoires

Contexte

**Méthode**

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Données

- ▶ Patients admis à l'HEGP pour CVO entre 2010 → 2015
- ▶ 329 séjours, 164 patients

Introduction

Trajectoires

Contexte

**Méthode**

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Méthode

## Données

- ▶ Patients admis à l'HEGP pour CVO entre 2010 → 2015
- ▶ 329 séjours, 164 patients
- ▶ Données longitudinales

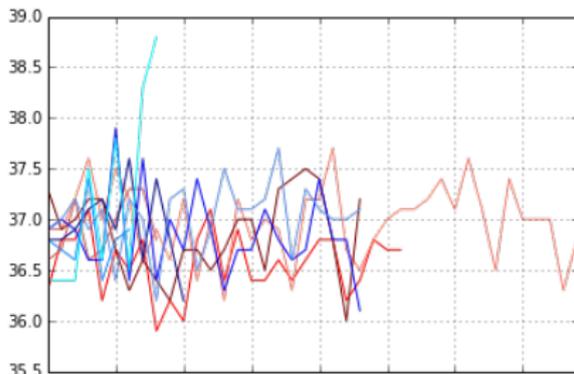


Figure 1: Évolution de la température pour 10 patients

Introduction

Trajectoires

Contexte

**Méthode**

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Données

- ▶ Patients admis à l'HEGP pour CVO entre 2010 → 2015
- ▶ 329 séjours, 164 patients
- ▶ Données longitudinales

## Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour  $i$  :

- ▶ on génère une grille uniforme de temps  $t_k$

Introduction

Trajectoires

Contexte

**Méthode**

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Données

- ▶ Patients admis à l'HEGP pour CVO entre 2010 → 2015
- ▶ 329 séjours, 164 patients
- ▶ Données longitudinales

## Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour  $i$  :

- ▶ on génère une grille uniforme de temps  $t_k$
- ▶ on ajuste un *spline*  $f_i$

Introduction

Trajectoires

Contexte

**Méthode**

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Données

- ▶ Patients admis à l'HEGP pour CVO entre 2010 → 2015
- ▶ 329 séjours, 164 patients
- ▶ Données longitudinales

## Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour  $i$  :

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

Introduction

Trajectoires

Contexte

**Méthode**

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Données

- ▶ Patients admis à l'HEGP pour CVO entre 2010 → 2015
- ▶ 329 séjours, 164 patients
- ▶ Données longitudinales

## Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour  $i$  :

- ▶ on génère 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

Introduction

Trajectoires

Contexte

**Méthode**

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Introduction

## Trajectoires

Contexte

Méthode

**Résultats**

## Réadmissions

Contexte

Méthode

Résultats

## C-mix

Modèle

Applications

Conclusion

## Binarity

Méthode

Applications

Conclusion

## Binacox

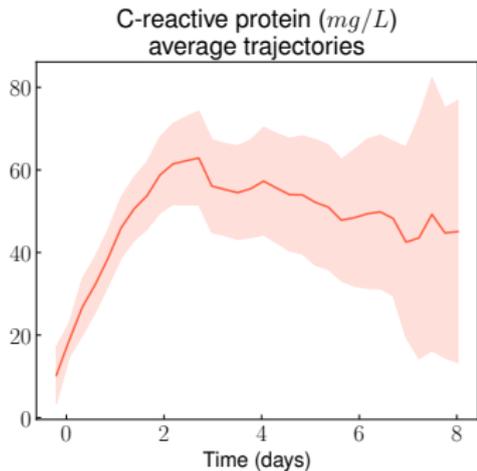
Méthode

Applications

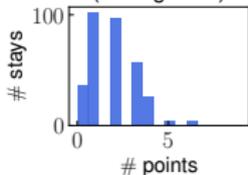
Conclusion

## Conclusion

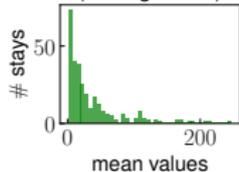
## Références



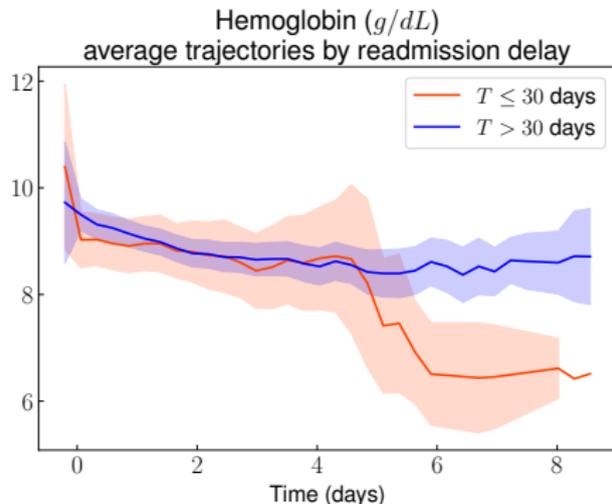
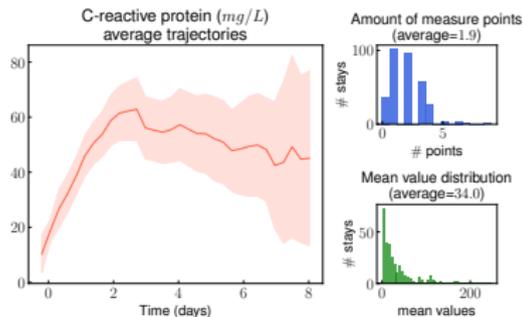
Amount of measure points  
(average=1.9)



Mean value distribution  
(average=34.0)



# Exemples de résultats graphiques



Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ Plusieurs biomarqueurs et paramètres vitaux évoluent

Introduction

Trajectoires

Contexte

Méthode

**Résultats**

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

Introduction

Trajectoires

Contexte

Méthode

**Résultats**

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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**

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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**

## Code Python

- ▶ Disponible à <https://github.com/SimonBussy/redcvo>

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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**

## Code Python

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

- ▶ Facteurs prédictifs de la réhospitalisation pour CVO

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

**Contexte**

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

**Contexte**

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

**Contexte**

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

**Contexte**

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

**Contexte**

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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  $\epsilon$  pré-défini

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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  $\epsilon$  pré-défini
  - ▶ Résultats très dépendants du choix de  $\epsilon$  [4]

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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  $\epsilon$  pré-défini
  - ▶ Résultats très dépendants du choix de  $\epsilon$  [4]
- ▶ *Survival analysis setting* : pas de seuil *a priori*

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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  $\epsilon$  pré-défini
  - ▶ Résultats très dépendants du choix de  $\epsilon$  [4]
- ▶ *Survival analysis setting* : pas de seuil *a priori*

## Objectif

Comparer des méthodes d'apprentissage issues de ces deux cadres

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ Étude rétrospective monocentrique sur la cohorte de I.

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

## Introduction

### Trajectoires

Contexte  
Méthode  
Résultats

### Réadmissions

Contexte  
**Méthode**  
Résultats

### C-mix

Modèle  
Applications  
Conclusion

### Binarsity

Méthode  
Applications  
Conclusion

### Binacox

Méthode  
Applications  
Conclusion

### Conclusion

### Références

# 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)

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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]

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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)

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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)

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Modèles considérés

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Modèles considérés

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Modèles considérés

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Modèles considérés

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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]

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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 neurones à une couche cachée (NN) [23]

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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 neurones à une couche cachée (NN) [23]
- ▶ *Survival analysis setting*, score : C-index [9]

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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 neurones à une couche cachée (NN) [23]
- ▶ *Survival analysis setting*, score : C-index [9]
  - ▶ Cox PH [5]

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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 neurones à une couche cachée (NN) [23]
- ▶ *Survival analysis setting*, score : C-index [9]
  - ▶ Cox PH [5]
  - ▶ CURE [6]

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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 neurones à une couche cachée (NN) [23]
- ▶ *Survival analysis setting*, score : C-index [9]
  - ▶ Cox PH [5]
  - ▶ CURE [6]
  - ▶ C-mix [3]

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

**Méthode**

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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 neurones 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

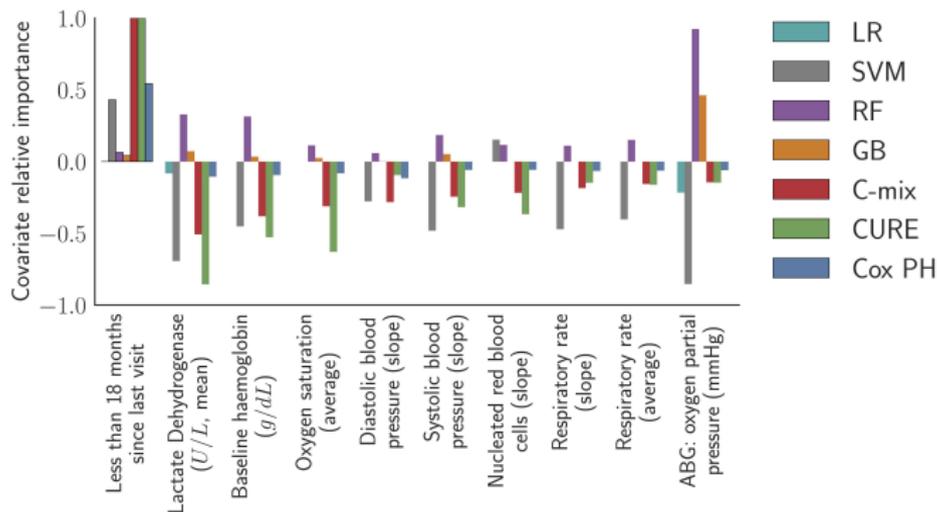
Prédiction de  $T_i > \epsilon$  par  $\hat{S}_i(\epsilon | X_i = x_i)$ , score : AUC

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarsity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

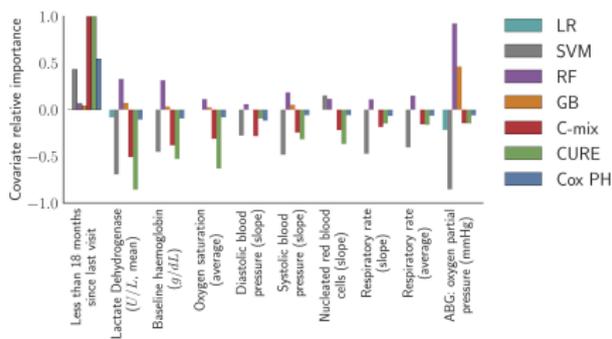
Table 1: Performances en prédiction

Setting	Métrique	Modèle	Score
Survival analysis	C-index	CURE	0.718
		Cox PH	0.725
		C-mix	<b>0.754</b>
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
		$\hat{S}^{\text{C-mix}}(\epsilon)$	<b>0.940</b>

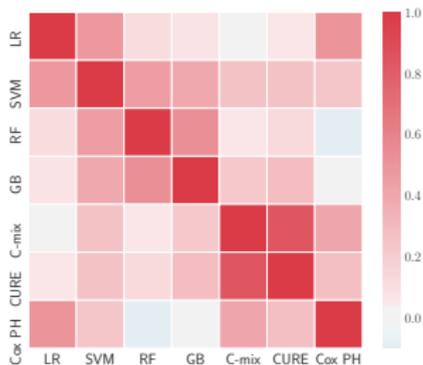


(a) Importance des 10 covariables les plus prédictives

# Résultats en sélection de variables



(a) Importance des 10 covariables les plus prédictives



(b) Corrélation linéaire en sélection de variables

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

**Résultats**

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

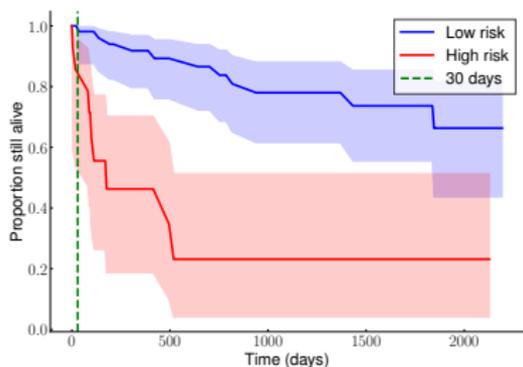
Méthode

Applications

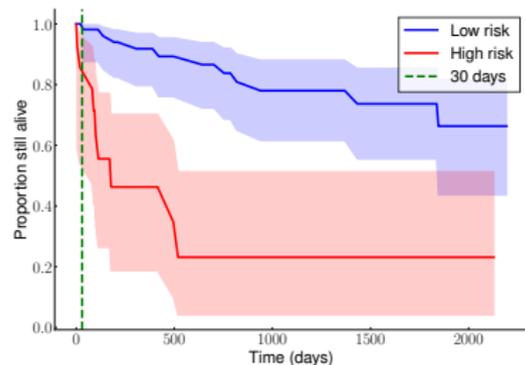
Conclusion

Conclusion

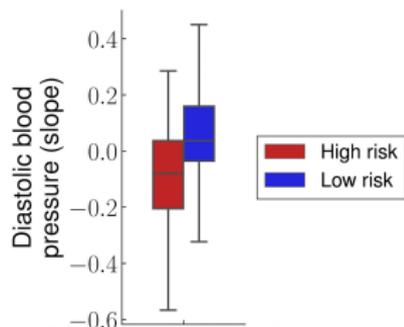
Références



(a) Courbes de survie



(a) Courbes de survie



(b) Boxplot

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ Modèles d'analyse de survie → fonctions de survie estimées → prédictions binaires pour un  $\epsilon$  donné

## Introduction

### Trajectoires

Contexte  
Méthode  
Résultats

### Réadmissions

Contexte  
Méthode  
**Résultats**

### C-mix

Modèle  
Applications  
Conclusion

### Binarsity

Méthode  
Applications  
Conclusion

### Binacox

Méthode  
Applications  
Conclusion

### Conclusion

### Références

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

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
**Résultats**

C-mix

Modèle  
Applications  
Conclusion

Binarsity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

# Conclusion

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

**Résultats**

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Conclusion

- ▶ Modèles d'analyse de survie → fonctions de survie estimées → prédictions binaires pour un  $\epsilon$  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**

Accepté à *BMC Medical Research Methodology*, 2018.

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

**Résultats**

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Conclusion

- ▶ Modèles d'analyse de survie → fonctions de survie estimées → prédictions binaires pour un  $\epsilon$  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**

Accepté à *BMC Medical Research Methodology*, 2018.

## Code Python

Disponible à

<https://github.com/SimonBussy/early-readmission-prediction>

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

- ▶ Contexte de l'analyse de survie

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

**Modèle**

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ Contexte de l'analyse de survie

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

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

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

**Modèle**  
Applications  
Conclusion

Binarsity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

- ▶ 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  $f(t|X = x) = \sum_{k=0}^{K-1} \pi_{\beta_k}(x) f_k(t; \alpha_k)$

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

**Modèle**  
Applications  
Conclusion

Binarity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

- ▶ 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)

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

**Modèle**  
Applications  
Conclusion

Binarity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

- ▶ 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)

- ▶  $Z|X = x \sim \mathcal{M}(\pi_{\beta_0}(x), \dots, \pi_{\beta_{K-1}}(x))$

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

**Modèle**  
Applications  
Conclusion

Binarity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

- ▶ 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)
- ▶  $Z|X = x \sim \mathcal{M}(\pi_{\beta_0}(x), \dots, \pi_{\beta_{K-1}}(x))$
- ▶ Hypothèses : (i)  $T|Z, X \perp\!\!\!\perp C|Z, X$  [13] (ii)  $C \perp\!\!\!\perp Z$  [14]

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

Modèle  
Applications  
Conclusion

Binarity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

- ▶ 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)
- ▶  $Z|X = x \sim \mathcal{M}(\pi_{\beta_0}(x), \dots, \pi_{\beta_{K-1}}(x))$
- ▶ Hypothèses : (i)  $T|Z, X \perp\!\!\!\perp C|Z, X$  [13] (ii)  $C \perp\!\!\!\perp Z$  [14]
- ▶ Échantillon *i.i.d.*  $(x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n) \in \mathbb{R}^d \times \mathbb{R}_+ \times \{0, 1\}$

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

Modèle  
Applications  
Conclusion

Binarity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

- ▶ 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)

- ▶  $Z|X=x \sim \mathcal{M}(\pi_{\beta_0}(x), \dots, \pi_{\beta_{K-1}}(x))$

- ▶ Hypothèses : (i)  $T|Z, X \perp\!\!\!\perp C|Z, X$  [13] (ii)  $C \perp\!\!\!\perp Z$  [14]

- ▶ Échantillon *i.i.d.*  $(x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n) \in \mathbb{R}^d \times \mathbb{R}_+ \times \{0, 1\}$

- ▶  $\theta = (\alpha_0, \dots, \alpha_K, \beta_0, \dots, \beta_K)^\top$ , log-vraisemblance du C-mix

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

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

Modèle  
Applications  
Conclusion

Binarity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

► Objectif pénalisé

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

**Modèle**

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ 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-vraisemblance complétée (négative)

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

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

- ▶ 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-vraisemblance complétée (négative)

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

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

$$\begin{aligned} q_{i,k}^{(l)} &= \mathbb{E}_{\theta^{(l)}}[\mathbb{1}_{\{z_i=k\}} | y_i, \delta_i] = \mathbb{P}_{\theta^{(l)}}[z_i = k | y_i, \delta_i] = \frac{\Lambda_{k,i}^{(l)}}{\sum_{r=0}^{K-1} \Lambda_{r,i}^{(l)}} \\ \Lambda_{k,i}^{(l)} &= [f_k(y_i; \alpha_k^{(l)}) \bar{G}(y_i^-)]^{\delta_i} [g(y_i) \bar{F}_k(y_i^-; \alpha_k^{(l)})]^{1-\delta_i} \pi_{\beta_k^{(l)}}(x_i) \end{aligned}$$

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

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

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

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

**Modèle**

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

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

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

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

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

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

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

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

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

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

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

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

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

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

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

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

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

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

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

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

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

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

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

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

- ▶ Convergence vers un minimum local prouvée

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

# Évaluation des performances

- ▶ Paramétrisation:  $K = 2$ ,  $Z = 1$  pour un risque de décès élevé

## Introduction

### Trajectoires

Contexte  
Méthode  
Résultats

### Réadmissions

Contexte  
Méthode  
Résultats

### C-mix

Modèle  
**Applications**  
Conclusion

### Binarsity

Méthode  
Applications  
Conclusion

### Binacox

Méthode  
Applications  
Conclusion

### Conclusion

### Références

- ▶ Paramétrisation:  $K = 2, Z = 1$  pour un risque de décès élevé
- ▶ Lois

$$\text{Weibull} \quad f_k(t; \alpha_k) = (1 - \phi_k)^{t^{\mu_k}} - (1 - \phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géométrique} \quad f_k(t; \alpha_k) = \alpha_k(1 - \alpha_k)^{t-1}$$

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

**Applications**

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ Paramétrisation:  $K = 2$ ,  $Z = 1$  pour un risque de décès élevé
- ▶ Lois

$$\text{Weibull } f_k(t; \alpha_k) = (1 - \phi_k)^{t^{\mu_k}} - (1 - \phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géométrique } f_k(t; \alpha_k) = \alpha_k(1 - \alpha_k)^{t-1}$$

- ▶ Modèles concurrents considérés
  - ▶ Cox PH pénalisé par l'Elastic-Net

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

**Applications**

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ Paramétrisation:  $K = 2$ ,  $Z = 1$  pour un risque de décès élevé
- ▶ Lois

$$\text{Weibull } f_k(t; \alpha_k) = (1 - \phi_k)^{t^{\mu_k}} - (1 - \phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géométrique } f_k(t; \alpha_k) = \alpha_k(1 - \alpha_k)^{t-1}$$

- ▶ Modèles concurrents considérés
  - ▶ Cox PH pénalisé par l'Elastic-Net
  - ▶ CURE pénalisé par l'Elastic-Net

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

**Applications**

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Évaluation des performances

- ▶ Paramétrisation:  $K = 2$ ,  $Z = 1$  pour un risque de décès élevé
- ▶ Lois

$$\text{Weibull } f_k(t; \alpha_k) = (1 - \phi_k)^{t^{\mu_k}} - (1 - \phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géométrique } f_k(t; \alpha_k) = \alpha_k(1 - \alpha_k)^{t-1}$$

- ▶ Modèles concurrents considérés
  - ▶ Cox PH pénalisé par l'Elastic-Net
  - ▶ CURE pénalisé par l'Elastic-Net
- ▶ Marqueur  $M = \pi_{\hat{\beta}}(X)$  ou  $\exp(X^\top \hat{\beta}^{\text{cox}})$

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarsity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

- ▶ Paramétrisation:  $K = 2$ ,  $Z = 1$  pour un risque de décès élevé
- ▶ Lois

$$\text{Weibull } f_k(t; \alpha_k) = (1 - \phi_k)^{t^{\mu_k}} - (1 - \phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géométrique } f_k(t; \alpha_k) = \alpha_k(1 - \alpha_k)^{t-1}$$

- ▶ Modèles concurrents considérés
  - ▶ Cox PH pénalisé par l'Elastic-Net
  - ▶ CURE pénalisé par l'Elastic-Net
- ▶ Marqueur  $M = \pi_{\hat{\beta}}(X)$  ou  $\exp(X^\top \hat{\beta}^{\text{cox}})$
- ▶ Métriques :

- ▶  $\text{AUC}^{\text{C},\text{D}}(t) = \int_{-\infty}^{\infty} \text{TPR}^{\text{C}}(\xi, t) \left| \frac{\partial \text{FPR}^{\text{D}}(\xi, t)}{\partial \xi} \right| d\xi$  [11]

- ▶ Paramétrisation:  $K = 2$ ,  $Z = 1$  pour un risque de décès élevé
- ▶ Lois

$$\text{Weibull } f_k(t; \alpha_k) = (1 - \phi_k)^{t^{\mu_k}} - (1 - \phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géométrique } f_k(t; \alpha_k) = \alpha_k(1 - \alpha_k)^{t-1}$$

- ▶ Modèles concurrents considérés
  - ▶ Cox PH pénalisé par l'Elastic-Net
  - ▶ CURE pénalisé par l'Elastic-Net
- ▶ Marqueur  $M = \pi_{\hat{\beta}}(X)$  ou  $\exp(X^\top \hat{\beta}^{\text{cox}})$
- ▶ Métriques :

- ▶  $\text{AUC}^{\text{C}, \text{D}}(t) = \int_{-\infty}^{\infty} \text{TPR}^{\text{C}}(\xi, t) \left| \frac{\partial \text{FPR}^{\text{D}}(\xi, t)}{\partial \xi} \right| d\xi$  [11]

- ▶ C-index  $\mathcal{C}_\tau = \mathbb{P}[M_i > M_j | Y_i < Y_j, Y_i < \tau]$

- ▶ Paramétrisation:  $K = 2$ ,  $Z = 1$  pour un risque de décès élevé
- ▶ Lois

$$\text{Weibull } f_k(t; \alpha_k) = (1 - \phi_k)^{t^{\mu_k}} - (1 - \phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géométrique } f_k(t; \alpha_k) = \alpha_k(1 - \alpha_k)^{t-1}$$

- ▶ Modèles concurrents considérés
  - ▶ Cox PH pénalisé par l'Elastic-Net
  - ▶ CURE pénalisé par l'Elastic-Net
- ▶ Marqueur  $M = \pi_{\hat{\beta}}(X)$  ou  $\exp(X^T \hat{\beta}^{\text{cox}})$
- ▶ Métriques :

- ▶  $\text{AUC}^{\text{C}, \text{D}}(t) = \int_{-\infty}^{\infty} \text{TPR}^{\text{C}}(\xi, t) \left| \frac{\partial \text{FPR}^{\text{D}}(\xi, t)}{\partial \xi} \right| d\xi$  [11]

- ▶ C-index  $\mathcal{C}_\tau = \mathbb{P}[M_i > M_j | Y_i < Y_j, Y_i < \tau]$

- ▶ Prédiction de  $T_i > \epsilon$  avec  $\hat{S}_i(\epsilon | X_i = x_i)$  pour différents  $\epsilon$ , où

$$\hat{S}_i(t | X_i = x_i) = \pi_{\hat{\beta}}(x_i) \hat{S}_1(t) + (1 - \pi_{\hat{\beta}}(x_i)) \hat{S}_0(t) \quad (\text{C-mix \& CURE})$$

$$\hat{S}_i(t | X_i = x_i) = [\hat{S}_0^{\text{cox}}(t)]^{\exp(x_i^T \hat{\beta}^{\text{cox}})} \quad (\text{Cox PH})$$

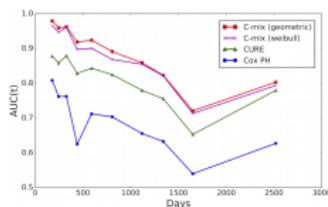
[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

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

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

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

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



(a) AUC( $t$ )

Figure 4: Résultats sur le cancer BRCA

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

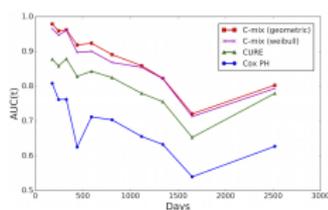
Conclusion

Conclusion

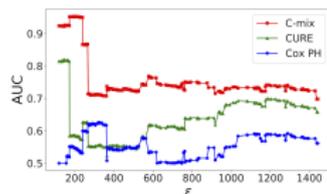
Références

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

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



(a) AUC( $t$ )

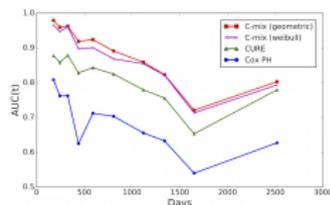


(b) AUC

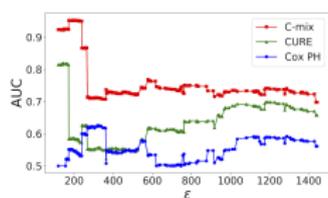
Figure 4: Résultats sur le cancer BRCA

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

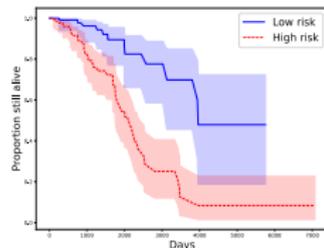
Cancer	Modèle	BRCA ( $n = 1211$ )			GBM ( $n = 168$ )			KIRC ( $n = 605$ )		
		C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
$d$	100	<b>0.792</b>	0.764	0.705	<b>0.826</b>	0.695	0.571	<b>0.768</b>	0.732	0.716
	300	<b>0.782</b>	0.753	0.723	<b>0.849</b>	0.697	0.571	<b>0.755</b>	0.691	0.698
	1000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



(a) AUC( $t$ )



(b) AUC



(c) Courbes de survie

Figure 4: Résultats sur le cancer BRCA

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

**Conclusion**

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ 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

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

Modèle  
Applications  
**Conclusion**

Binarsity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

# 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*, 2017.

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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*, 2017.

## Code Python

- ▶ Disponible à <https://github.com/SimonBussy/C-mix>

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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*, 2017.

## Code Python

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Conclusion

## Code Python

- ▶ Programmes annotés, notebooks et tutoriels

```
1 from QNEM.inference import QNEM
2 from QNEM.simulation import CensoredGeomMixtureRegression
3
4 # Generate data
5 simu = CensoredGeomMixtureRegression(n_samples=1000, n_features=100,
6                                     n_active_features=30)
7 X, Y, delta = simu.simulate()
8
9 # Choose between C-mix or CURE model
10 model = 'C-mix'
11
12 # Fit the model with a penalty strength equal to 10
13 learner = QNEM(model=model, l_elastic_net=10, eta=.1, max_iter=100,
14                tol=1e-6, warm_start=True, fit_intercept=True)
15 learner.fit(X, Y, delta)
16
17 # Obtain the estimated marker
18 coeffs = learner.coefs
19 marker = QNEM.predict_proba(X, fit_intercept, coeffs)
```

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

# 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*, 2017.

## Code Python

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

**Binarsity**

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

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

## Introduction

### Trajectoires

Contexte  
Méthode  
Résultats

### Réadmissions

Contexte  
Méthode  
Résultats

### C-mix

Modèle  
Applications  
Conclusion

### Binarsity

**Méthode**  
Applications  
Conclusion

### Binacox

Méthode  
Applications  
Conclusion

### Conclusion

### Références

► 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

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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

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

$X_{\bullet,1}$

$\vdots$

12,5

9,2

3,1

8,7

$\vdots$

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

**Méthode**

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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

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

$$\begin{array}{cccc}
 & \overset{d_1 = 4}{\curvearrowright} & & & \\
 X_{\bullet,1} & & q_1(0.25) = 7,9 & q_1(0.5) = 9,6 & q_1(0.75) = 11,7 \\
 \vdots & & & & \\
 12,5 & & & & \\
 9,2 & & & & \\
 3,1 & & & & \\
 8,7 & & & & \\
 \vdots & & & & 
 \end{array}$$

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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

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

	$d_1 = 4$			
$X_{\bullet,1}$	$q_1(0.25) = 7,9$	$q_1(0.5) = 9,6$	$q_1(0.75) = 11,7$	
$\vdots$				
$i \rightarrow 12,5$				
9,2				
3,1				
8,7				
$\vdots$				

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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

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

	$d_1 = 4$				
	↘				
$X_{\bullet,1}$	$q_1(0.25) = 7,9$	$q_1(0.5) = 9,6$	$q_1(0.75) = 11,7$		
⋮					
$i \rightarrow 12,5$	→ 0	0	0	1	
9,2					
3,1					
8,7					
⋮					

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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

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

		$\overset{d_1 = 4}{\curvearrowright}$			
$X_{\bullet,1}$		$q_1(0.25) = 7,9$	$q_1(0.5) = 9,6$	$q_1(0.75) = 11,7$	
⋮	⋮	⋮	⋮	⋮	⋮
$i \rightarrow 12,5$	→ 0	0	1	0	1
9,2	→ 0	0	1	0	0
3,1	→ 1	1	0	0	0
8,7	→ 0	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ 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

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

- ▶ 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  $d = \sum_{j=1}^p d_j$

- ▶ 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

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

- ▶ 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

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

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

## Introduction

### Trajectoires

Contexte  
Méthode  
Résultats

### Réadmissions

Contexte  
Méthode  
Résultats

### C-mix

Modèle  
Applications  
Conclusion

### Binarity

**Méthode**  
Applications  
Conclusion

### Binacox

Méthode  
Applications  
Conclusion

### Conclusion

### Références

## Problèmes

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

**Méthode**

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Problèmes

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

## Introduction

### Trajectoires

Contexte  
Méthode  
Résultats

### Réadmissions

Contexte  
Méthode  
Résultats

### C-mix

Modèle  
Applications  
Conclusion

### Binarity

**Méthode**  
Applications  
Conclusion

### Binacox

Méthode  
Applications  
Conclusion

### Conclusion

### Références

## Problèmes

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

## Problèmes

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

## Réponses

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

## Problèmes

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

## Réponses

- ▶ (P1)  $\rightarrow$  imposer  $\sum_{k=1}^{d_j} \beta_{j,k} = 0$
- ▶ (P2)  $\rightarrow$  pénalisation TV par bloc  $\sum_{j=1}^p \|\beta_{j,\bullet}\|_{TV, \hat{w}_{j,\bullet}}$   
où  $\|\beta_{j,\bullet}\|_{TV, \hat{w}_{j,\bullet}} = \sum_{k=2}^{d_j} \hat{w}_{j,k} |\beta_{j,k} - \beta_{j,k-1}|$

## Problèmes

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

## Réponses

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

# La méthode

## Problèmes

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

## Réponses

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

## Pénalité binarsity

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

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

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarsity](#)**Méthode**[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

# La méthode

## Problèmes

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

## Réponses

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

## Pénalité binarsity

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

$$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \{ R_n(\beta) + \text{bina}(\beta) \}$$

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarsity](#)**Méthode**[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

# Résultat théorique

- Fonction de risque associée [21]

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

**Méthode**

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Résultat théorique

- ▶ Fonction de risque associée [21]

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

- ▶  $J(\beta) = [J_1(\beta), \dots, J_p(\beta)]$ , avec

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

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarsity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

# Résultat théorique

- ▶ Fonction de risque associée [21]

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

- ▶  $J(\beta) = [J_1(\beta), \dots, J_p(\beta)]$ , avec

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

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

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

# Résultat théorique

- ▶ Fonction de risque associée [21]

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

- ▶  $J(\beta) = [J_1(\beta), \dots, J_p(\beta)]$ , avec

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

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

## Inégalité oracle non-asymptotique à vitesse rapide

Avec grande probabilité, on a

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

avec  $c_1, c_2 > 0$

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

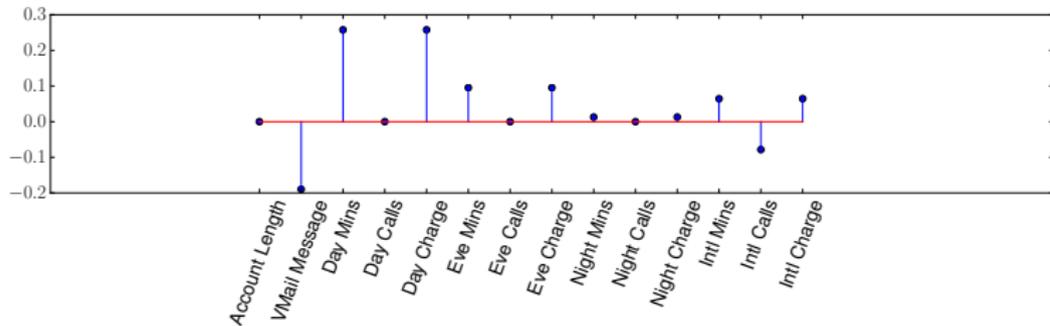


Figure 5: LR sur "Churn" ( $n = 3333, p = 14$ )

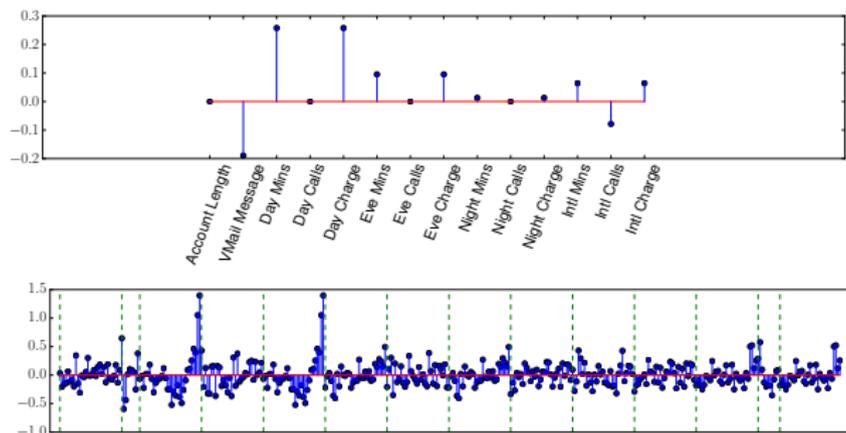


Figure 5: LR sur "Churn" ( $n = 3333$ ,  $p = 14$ )

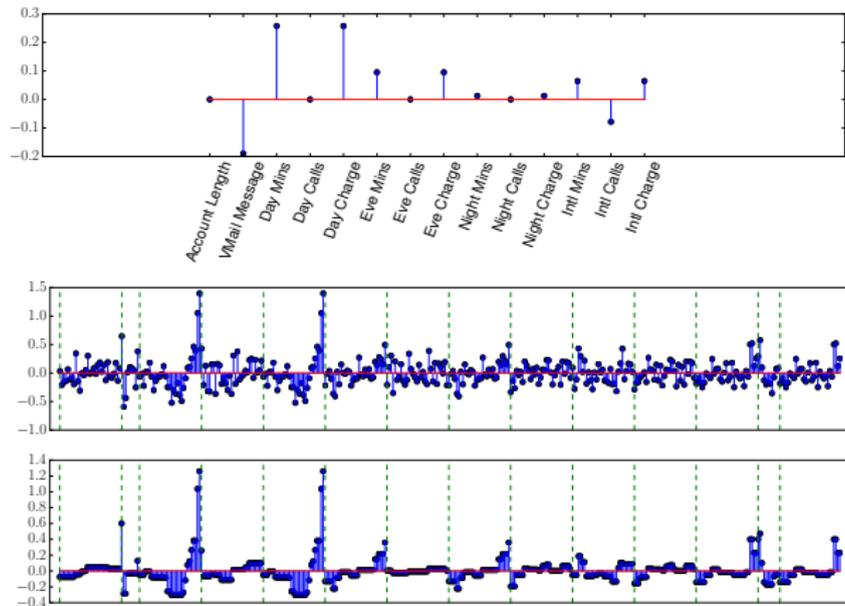


Figure 5: LR sur "Churn" ( $n = 3333, p = 14$ )

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

**Applications**

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Illustration de binarsity

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

**Applications**

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

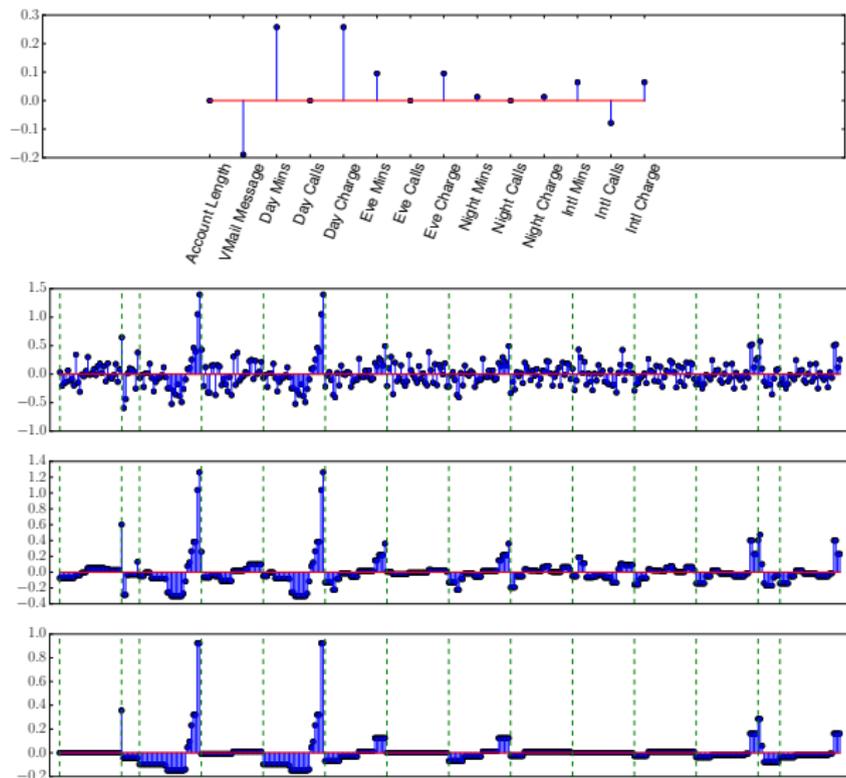


Figure 5: LR sur "Churn" ( $n = 3333$ ,  $p = 14$ )

# Illustration de binarsity

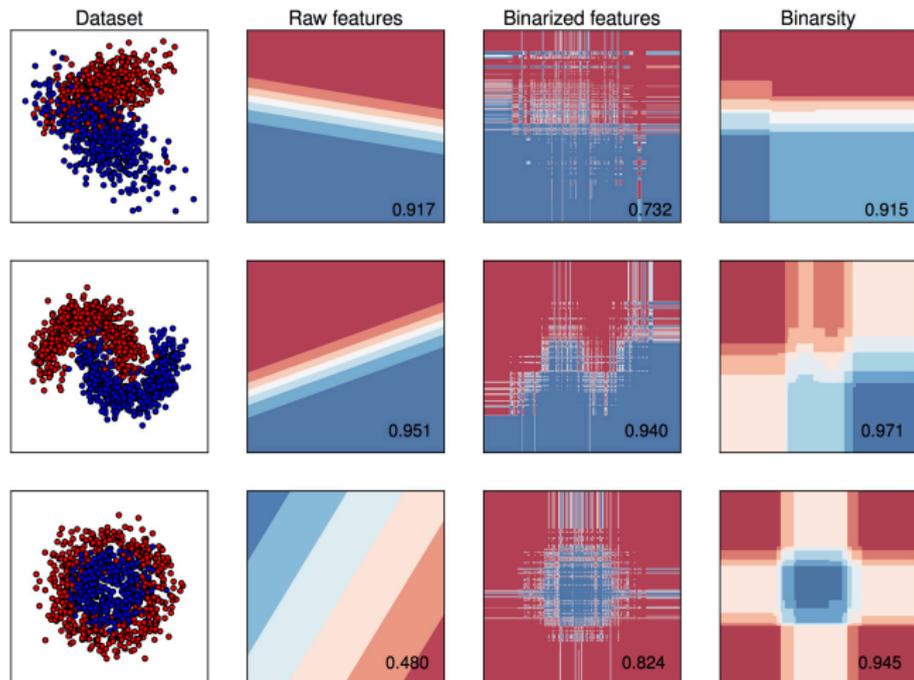


Figure 6: Illustration sur 3 jeux de données simulés avec  $n = 1000$ ,  $p = 2$  et  $d_1 = d_2 = 100$

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

Jeux de données	$n$	$p$
Ionosphere	351	34
Churn	3333	21
Default of credit card	30000	24
Adult	32561	14
Bank marketing	45211	17
Covertypes	550088	10
SUSY	5000000	18
HEPMASS	10500000	28
HIGGS	11000000	24

Table 3: Les 9 jeux de données considérés

Jeux de données	$n$	$p$
Ionosphere	351	34
Churn	3333	21
Default of credit card	30000	24
Adult	32561	14
Bank marketing	45211	17
Covertypes	550088	10
SUSY	5000000	18
HEPMASS	10500000	28
HIGGS	11000000	24

Table 3: Les 9 jeux de données considérés

Nom	Description
Lasso [20]	Régression Logistique (LR) avec pénalité $\ell_1$
Group L1 [16]	LR avec pénalité group $\ell_1$
Group TV	LR avec pénalité group total-variation
SVM	Support vector machine avec noyau gaussien
GAM [10]	Generalized additive model
RF	Random forest
GB	Gradient boosting

Table 4: Modèles concurrents

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

**Applications**

Conclusion

Binacox

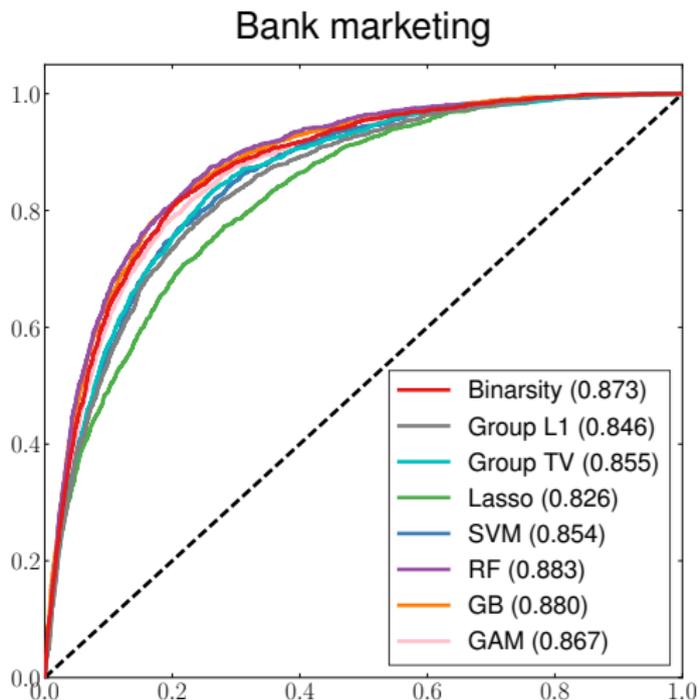
Méthode

Applications

Conclusion

Conclusion

Références



(a) Courbes ROC

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

**Applications**

Conclusion

Binacox

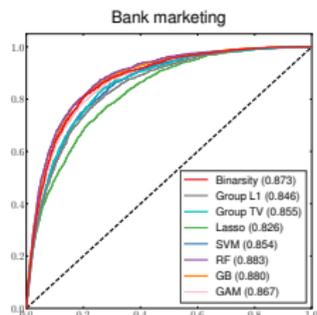
Méthode

Applications

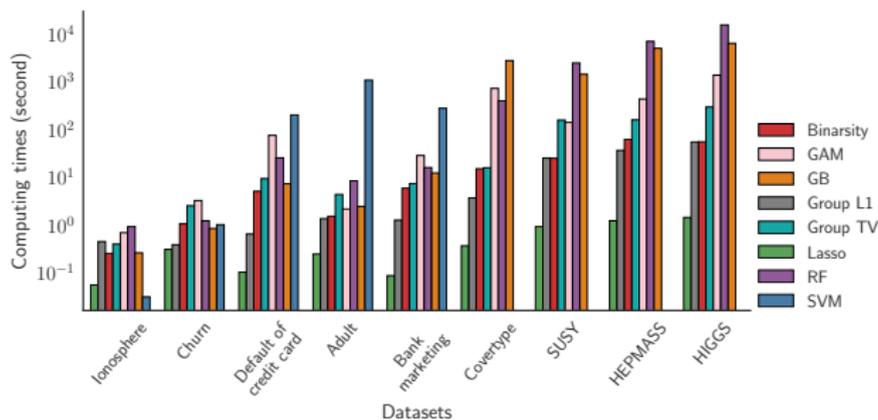
Conclusion

Conclusion

Références



(a) Courbes ROC



(b) Temps de calcul

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

**Conclusion**

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

**Conclusion**

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

**Conclusion**

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

**Conclusion**

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Conclusion

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

## Article associé

M.Z. Alaya, **S. Bussy**, S. Gaïffas, A. Guilloux

### **Binarsity: a penalization for one-hot encoded features**

En révision mineure dans *Journal of Machine Learning Research*, 2018.

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Conclusion

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

## Article associé

M.Z. Alaya, **S. Bussy**, S. Gaïffas, A. Guilloux

### **Binarsity: a penalization for one-hot encoded features**

En révision mineure dans *Journal of Machine Learning Research*, 2018.

## Code Python/C++

- ▶ Disponible à <https://github.com/SimonBussy/binarsity>

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Conclusion

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

## Article associé

M.Z. Alaya, **S. Bussy**, S. Gaïffas, A. Guilloux

### **Binarsity: a penalization for one-hot encoded features**

En révision mineure dans *Journal of Machine Learning Research*, 2018.

## Code Python/C++

- ▶ Disponible à <https://github.com/SimonBussy/binarsity>
- ▶ Library tick : <https://github.com/X-DataInitiative/tick>

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

Modèle  
Applications  
Conclusion

Binarsity

Méthode  
Applications  
Conclusion

**Binacox**

Méthode  
Applications  
Conclusion

Conclusion

Références

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

- ▶ Biomarqueurs → décision clinique : choix de seuils

## Introduction

### Trajectoires

Contexte  
Méthode  
Résultats

### Réadmissions

Contexte  
Méthode  
Résultats

### C-mix

Modèle  
Applications  
Conclusion

### Binarsity

Méthode  
Applications  
Conclusion

### Binacox

**Méthode**  
Applications  
Conclusion

## Conclusion

## Références

# Le contexte

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

## Introduction

### Trajectoires

Contexte  
Méthode  
Résultats

### Réadmissions

Contexte  
Méthode  
Résultats

### C-mix

Modèle  
Applications  
Conclusion

### Binarsity

Méthode  
Applications  
Conclusion

### Binacox

**Méthode**  
Applications  
Conclusion

### Conclusion

### Références

# Le contexte

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

**Méthode**

Applications

Conclusion

Conclusion

Références

# Le contexte

- ▶ Biomarqueurs → 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,

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

**Méthode**

Applications

Conclusion

Conclusion

Références

# Le contexte

- ▶ Biomarqueurs → 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**,

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

**Méthode**

Applications

Conclusion

Conclusion

Références

# Le contexte

- ▶ Biomarqueurs → 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**

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

**Méthode**

Applications

Conclusion

Conclusion

Références

# Le contexte

- ▶ Biomarqueurs → 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  $X \in \mathbb{R}^p$

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarsity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)

# Le contexte

- ▶ Biomarqueurs → 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

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

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Le contexte

- ▶ Biomarqueurs → 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

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

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

- ▶ But : estimer simultanément

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Le contexte

- ▶ Biomarqueurs → 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

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

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

- ▶ But : estimer simultanément
  - ▶  $\mu^* = (\mu_{1,1}^*, \dots, \mu_{1,K_1^*}^*, \dots, \mu_{p,1}^*, \dots, \mu_{p,K_p^*}^*)^\top \in \mathbb{R}^{K^*}$

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# Le contexte

- ▶ Biomarqueurs → 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

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

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

- ▶ But : estimer simultanément
  - ▶  $\mu^* = (\mu_{1,1}^*, \dots, \mu_{1,K_1^*}^*, \dots, \mu_{p,1}^*, \dots, \mu_{p,K_p^*}^*)^\top \in \mathbb{R}^{K^*}$
  - ▶  $\beta^* = (\beta_{1,1}^*, \dots, \beta_{1,K_1^*+1}^*, \dots, \beta_{p,1}^*, \dots, \beta_{p,K_p^*+1}^*)^\top \in \mathbb{R}^{K^*+p}$

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ Biomarqueurs → 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

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

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

- ▶ But : estimer simultanément
  - ▶  $\mu^* = (\mu_{1,1}^*, \dots, \mu_{1,K_1^*}^*, \dots, \mu_{p,1}^*, \dots, \mu_{p,K_p^*}^*)^\top \in \mathbb{R}^{K^*}$
  - ▶  $\beta^* = (\beta_{1,1}^*, \dots, \beta_{1,K_1^*+1}^*, \dots, \beta_{p,1}^*, \dots, \beta_{p,K_p^*+1}^*)^\top \in \mathbb{R}^{K^*+p}$
  - ▶  $K^* = \sum_{j=1}^p K_j^*$

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

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

où

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

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

**Méthode**

Applications

Conclusion

Conclusion

Références

- ▶ 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  $f^*$

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ 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  $f^*$
- ▶ log-vraisemblance négative partielle  $\ell_n(f_\beta)$

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

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

Modèle  
Applications  
Conclusion

Binarity

Méthode  
Applications  
Conclusion

Binacox

**Méthode**  
Applications  
Conclusion

Conclusion

Références

## Introduction

## Trajectoires

Contexte  
Méthode  
Résultats

## Réadmissions

Contexte  
Méthode  
Résultats

## C-mix

Modèle  
Applications  
Conclusion

## Binarity

Méthode  
Applications  
Conclusion

## Binacox

**Méthode**  
Applications  
Conclusion

## Conclusion

## Références

- ▶ 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-vraisemblance négative partielle  $\ell_n(f_\beta)$
- ▶  $\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathcal{B}_{p+d}(\mathbb{R})} \{ \ell_n(f_\beta) + \operatorname{bina}(\beta) \}$

- ▶ 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-vraisemblance négative partielle  $\ell_n(f_\beta)$
- ▶  $\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathcal{B}_{p+d}(\mathbb{R})} \{ \ell_n(f_\beta) + \operatorname{bina}(\beta) \}$
- ▶  $J_j(\hat{\beta}) = \{ l : \beta_{j,l} \neq \beta_{j,l-1}, \text{ pour } l = 2, \dots, d_j + 1 \} = \{ \hat{l}_{j,1}, \dots, \hat{l}_{j,s_j} \}$

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## Introduction

## Trajectoires

Contexte  
Méthode  
Résultats

## Réadmissions

Contexte  
Méthode  
Résultats

## C-mix

Modèle  
Applications  
Conclusion

## Binarity

Méthode  
Applications  
Conclusion

## Binacox

Méthode  
Applications  
Conclusion

## Conclusion

## Références

- ▶ 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-vraisemblance négative partielle  $\ell_n(f_\beta)$
- ▶  $\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathcal{B}_{p+d}(\mathbb{R})} \{ \ell_n(f_\beta) + \operatorname{bina}(\beta) \}$
- ▶  $J_j(\hat{\beta}) = \{ l : \beta_{j,l} \neq \beta_{j,l-1}, \text{ pour } l = 2, \dots, d_j + 1 \} = \{ \hat{l}_{j,1}, \dots, \hat{l}_{j,s_j} \}$
- ▶  $\hat{\mu}_{j,\bullet} = (\mu_{j,\hat{l}_{j,1}}, \dots, \mu_{j,\hat{l}_{j,s_j}})^\top$  avec  $s_j = |J_j(\hat{\beta})| = \hat{K}_j$

- ▶ 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  $f^*$
- ▶ log-vraisemblance négative partielle  $\ell_n(f_\beta)$
- ▶  $\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathcal{B}_{p+d}(\mathbb{R})} \{ \ell_n(f_\beta) + \operatorname{bina}(\beta) \}$
- ▶  $J_j(\hat{\beta}) = \{ l : \beta_{j,l} \neq \beta_{j,l-1}, \text{ pour } l = 2, \dots, d_j + 1 \} = \{ \hat{l}_{j,1}, \dots, \hat{l}_{j,s_j} \}$
- ▶  $\hat{\mu}_{j,\bullet} = (\mu_{j,\hat{l}_{j,1}}, \dots, \mu_{j,\hat{l}_{j,s_j}})^\top$  avec  $s_j = |J_j(\hat{\beta})| = \hat{K}_j$

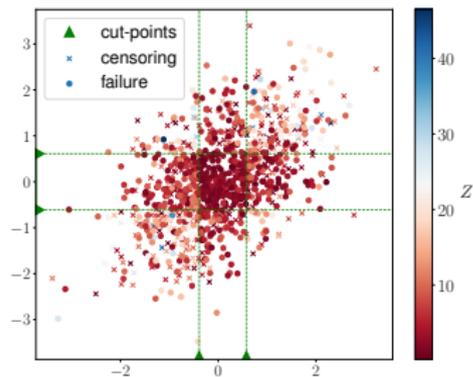
## Inégalité oracle non-asymptotique à vitesse rapide

Avec grande probabilité, on a

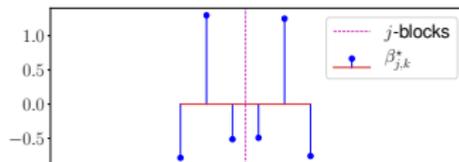
$$KL_n(f^*, \hat{f}_\beta) \leq (1 + c_1) \inf_{\substack{\beta \in \mathcal{B}_{p+d}(\mathbb{R}) \\ |J(\beta)| \leq K^* \\ \forall j, \mathbf{1}^\top \beta_{j,\bullet} = 0}} \left\{ KL_n(f^*, f_\beta) + c_2 |J(\beta)| \frac{\log(p+d)}{n} \right\}$$

avec  $c_1, c_2 > 0$

[Introduction](#)[Trajectoires](#)[Contexte](#)[Méthode](#)[Résultats](#)[Réadmissions](#)[Contexte](#)[Méthode](#)[Résultats](#)[C-mix](#)[Modèle](#)[Applications](#)[Conclusion](#)[Binarity](#)[Méthode](#)[Applications](#)[Conclusion](#)[Binacox](#)[Méthode](#)[Applications](#)[Conclusion](#)[Conclusion](#)[Références](#)



(a) Données générées



(b)  $\beta^*$  correspondant

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

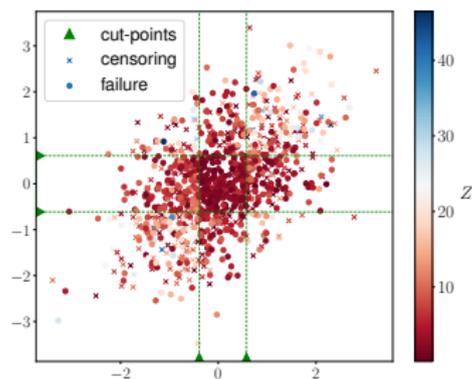
Méthode

**Applications**

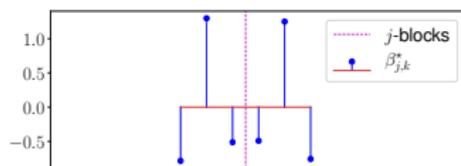
Conclusion

Conclusion

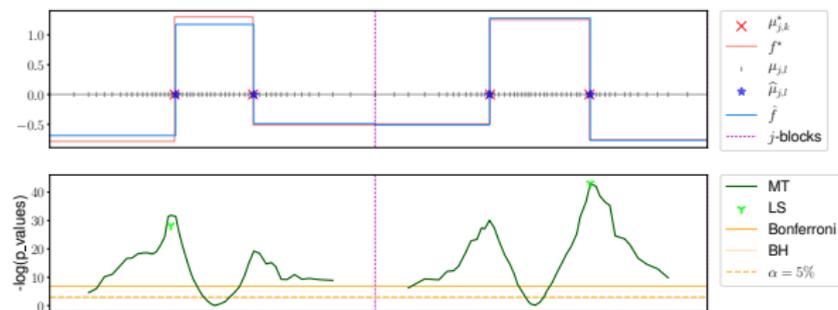
Références



(a) Données générées



(b)  $\beta^*$  correspondant



(c) Résultats des différentes méthodes

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

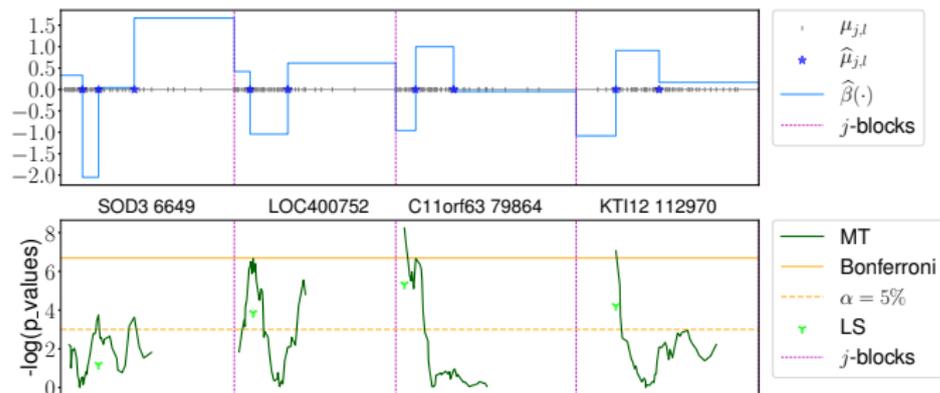
Méthode

Applications

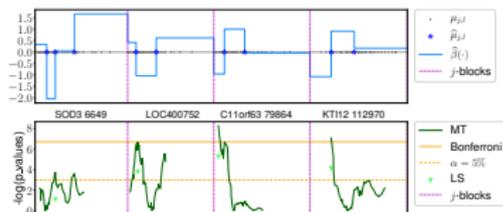
Conclusion

Conclusion

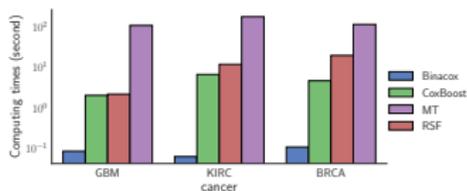
Références



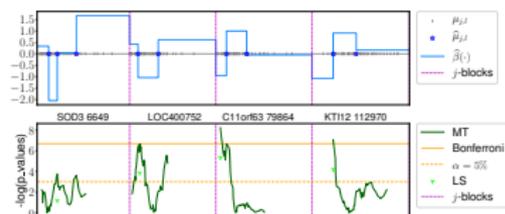
(a) Résultats sur le cancer GBM ( $n = 168, p = 20531$ )



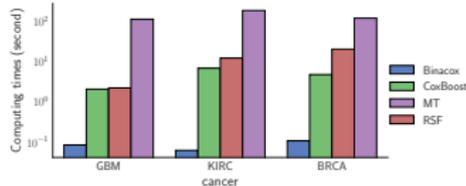
(a) Résultats sur le cancer GBM



(b) Temps de calcul



(a) Résultats sur le cancer GBM



(b) Temps de calcul

Table 5: Comparison des C-index

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- ▶ Détection multivariée de plusieurs cut-points par covariable continue

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

**Conclusion**

Conclusion

Références

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

**Conclusion**

Conclusion

Références

- ▶ Détection multivariée de plusieurs cut-points par covariable continue
- ▶ Implémentation rapide et bons résultats obtenus en estimation sur données simulées et réelles
- ▶ Preuve de la consistance en détection en cours

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

**Conclusion**

Conclusion

Références

# Conclusion

- ▶ Détection multivariée de plusieurs cut-points par covariable continue
- ▶ Implémentation rapide et bons résultats obtenus en estimation sur données simulées et réelles
- ▶ Preuve de la consistance en détection en cours

## 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**

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

**Conclusion**

Conclusion

Références

# Conclusion

- ▶ Détection multivariée de plusieurs cut-points par covariable continue
- ▶ Implémentation rapide et bons résultats obtenus en estimation sur données simulées et réelles
- ▶ Preuve de la consistance en détection en cours

## 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**

## Code Python/C++

Disponible à <https://github.com/SimonBussy/binacox>

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

**Conclusion**

Références

## VI. Conclusion générale

## Nouvelles méthodes

- ▶ Différentes méthodes de machine learning proposées

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

**Conclusion**

Références

# Conclusion générale

## Nouvelles méthodes

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

**Conclusion**

Références

## 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.*)

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

## 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
- ▶ Plusieurs collaborations enrichissantes

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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
- ▶ Plusieurs collaborations enrichissantes

## Implémentation

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

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

# 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
- ▶ Plusieurs collaborations enrichissantes

## Implémentation

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

## Perspectives de recherche

Différents nouveaux projets en cours / extensions des modèles proposés

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références

- [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.

Introduction

Trajectoires

Contexte  
Méthode  
Résultats

Réadmissions

Contexte  
Méthode  
Résultats

C-mix

Modèle  
Applications  
Conclusion

Binarity

Méthode  
Applications  
Conclusion

Binacox

Méthode  
Applications  
Conclusion

Conclusion

Références

- [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.

Introduction

Trajectoires

Contexte

Méthode

Résultats

Réadmissions

Contexte

Méthode

Résultats

C-mix

Modèle

Applications

Conclusion

Binarity

Méthode

Applications

Conclusion

Binacox

Méthode

Applications

Conclusion

Conclusion

Références