





Prix de thèse Daniel Schwartz

présentée le 02/10/2020 par

Simon Bussy (1)

https://simonbussy.com/

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

sous la direction de

Pr. Stéphane Gaïffas (1,2)

Pr. Agathe Guilloux (3)

Dr. Anne-Sophie Jannot (4,5)

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

Soutenance de thèse 1/27

ntroduction

rajectoire

Méthode

Résultats

éadmissions

ontexte

Résultats

-mix

Modèle

---l...:--

narsity

Aáthoda

Conclusion

Jonetusion

inacox

Applications

Conclusion

Conclusion

2/27

Soutenance

de thèse

Cadre

Apprentissage supervisé

Introduction

rajectoires

Contexte

Dácultati

Réadmission

Contexte

resultat:

C-mi

1odèle

applications

inarsity

Méthode Applications

Applications

Binacox

Méthode

Conclusion

Conclusion

Cadre

- Apprentissage supervisé
- ► Analyse de survie

Soutenance de thèse 2/27

Introduction

Trajectoires

Conte

D.C. I

Résultat

Réadmissions

Contexte

Dácultata

-:

/lodèle

Applications

Application:

Binarsity

Méthode Applications

Applications

inacox

Make

Application Conclusion

Conclusion

Cadre

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

Soutenance de thèse 2/27

Cadre

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

Soutenance de thèse 2/27

Soutenance de thèse 2/27

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

Soutenance de thèse 2/27

Cadre

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

Principaux projets entrepris

- Trajectories of biological values and vital parameters, a retrospective cohort study on non-complicated vaso-occlusive crises
- Early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework

Introduction

Trajec

Méthod

Résultats

Réadmissions

Contexte

Résultats

miv

/lodèle

Applications

onclusion

narsity

Méthode Applications

Conclusion

Binacox

Applicatio

Conclusion

Soutenance de thèse 2/27

Cadre

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

Principaux projets entrepris

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

Haje

Méthode

Résultats

Réadmissions

ontexte

Résultats

miv

mix

/lodéle

onclusion

narsity

Méthode Applications

Conclusion

sinacox

Méthode

Conclusion

Conclusion

Soutenance de thèse 2/27

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
- Binarsity, a penalization for one-hot encoded features

Soutenance de thèse 2/27

Cadre

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

Principaux projets entrepris

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

Traje

Méthod

Résultats

Réadmissions

ontexte

Résultats

miv

mix

viodele Applications

onclusion

narsity

Méthode Applications

Conclusion

Sinacox

Méthode

Application Conclusion

Conclusion

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

Soutenance de thèse 2/27

....

Trajectoires

Méthod

Réadmission

Contexte Méthode

_ .

Modèle

Applications

Binarsity

Applications

Binacox

Méthode Application

Application

Conclusion

Contexte

► Drépanocytose : maladie génétique la plus fréquente [17]

Soutenance de thèse 3/27

Contexte

Contexte

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

Soutenance de thèse 3/27

Introduction

Trajectoires

Contexte

Récultate

Réadmissions

Contexte

Résultats

C-mi

Modèle

pplications

Binarsity

Méthode Applications

Applications Conclusion

inacox

Máthada

Applicatio Conclusion

Conclusion

Contexte

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

Soutenance de thèse 3/27

Contexte

Contexte

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

Soutenance de thèse 3/27

Introduction

Trajectoires

Contexte

Méthode

Réadmission

Contexte

Résultats

-miv

A DI

Applications

Applications

inarsity

Méthode Applications

inacox

inacox

Applicatio Conclusion

Conclusion

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"

Soutenance de thèse 3/27

Contexte

3/27

Soutenance

de thèse

.....

Trajectoire:

Contexte

Methode

Réadmissions

Contexte

Résultat

î-mix

∠-IIIIX

1odèle

Applications

inarsity

Méthode

Applications

Binacox

Méthode

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

Soutenance de thèse 3/27

Introduction

Trajectoires

Contexte

Méthod

Réadmissions

-

Méthode

.-mix

1odèle

Applications

inarsity

sinarsity

Applications

Conclusion

omacox

Application

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

Soutenance de thèse 4/27

Introductio

Trajectoires

Méthode

Résultats

Réadmissions

Contexte

Résultats

-HHX

Applications

Méthode

Applications

Conclusion

Sinacox

Méthode Applications

Conclusion

Références

Données

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

Données

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

Soutenance de thèse 4/27

IIIIIOddctioi

Trajectoires

Méthode

Résultats

Réadmission

Contexte

Résultats

-mix

Modèle Applications

...........

Méthode Applications

inacox

Méthode Applications

Conclusion

Données

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

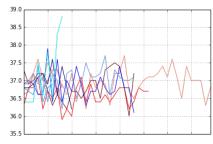


Figure 1: Évolution de la température pour 10 patients

Soutenance de thèse 4/27

Introduction

Trajectoires

Méthode

B/ 1 1 1

Réadmissions

ontexte

-

Modèle

Applications

inarsity

Méthode

Applications Conclusion

Binacox

Méthode

Applicatio

Conclusion

Soutenance de thèse 4/27

Máthoda

Données

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

Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour i :

 \triangleright on génére une grille uniforme de temps t_k

Soutenance de thèse 4/27

Máthoda

Données

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

Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour i :

- ightharpoonup on génére une grille uniforme de temps t_k
- on ajuste un spline fi

Soutenance de thèse 4/27

Máthoda

Données

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

Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour i :

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

Soutenance de thèse 4/27

Máthoda

Données

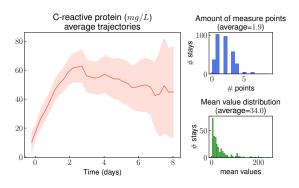
- \blacktriangleright Patients admis à l'HEGP pour CVO entre 2010 \rightarrow 2015
- 329 séjours, 164 patients
- Données longitudinales

Description des trajectoires moyennes

Pour chaque trajectoire et pour chaque séjour i :

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

Examples de résultats graphiques



Soutenance de thèse 5/27

Introductio

Trajectoire

Méthod

Résultats

Réadmission

Contexte

Résultats

l-mix

Modèle Application

Conclusion

inarsity

Méthode Applications

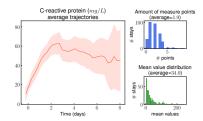
Conclusion

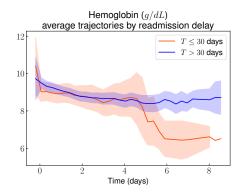
Binacox

Application Conclusion

Conclusion

Examples de résultats graphiques





Soutenance de thèse 5/27

Introductio

Trajectoire

Méthod

Résultats

Réadmission

ontexte

Résult

-mix

Modèle

pplications

narsity

Méthode

Applications Conclusion

Binacox

Méthode Application

Conclusion

Conclusion

▶ Plusieurs biomarqueurs et paramètres vitaux évoluent

Soutenance de thèse 6/27

IIIIIoductio

rajectoires

Context

Résultats

Réadmissions

Contexte

Résultats

mix

lodèle

Applications

Binarsity

Méthode

Applications

Binacox

JIIIaCOX

Applications Conclusion

Conclusion

Conclusion

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

Soutenance de thèse 6/27

IIIIIoductio

Trajectoires

Métho

Résultats

Réadmission

Contexte

Résultats

-mix

Applications

onclusion

inarsity

Vléthode

Applications

inacox

Méthode

Applicatio Conclusion

Conclusion

Conclusion

Soutenance de thèse 6/27

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

Article associé

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

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

Publié dans Journal of Clinical Medicine, 2019.

Introduction

Frajectoires

Méthode

Résultats

Réadmissions

Contexte Méthode

Modèle

Applications

Binarsity

Applications

Binacox

Méthode Applicatio

a malicata m

Contexte

 Extraction d'information pertinente à partir d'un entrepôt de données cliniques de grande dimension

Résultats

Article associé

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

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

Publié dans Journal of Clinical Medicine, 2019.

Code Python

▶ Disponible à https://github.com/SimonBussy/redcvo

.......

Applications

Conclusion

Binacox

Applicatio

Conclusion

L-IIIIX Madàla

Applications

Conclusion

Binarsity

Méthode

Applications Conclusion

Binacox

Máthada

Applicatio

Conclusion

Références

▶ Plusieurs biomarqueurs et paramètres vitaux évoluent

Extraction d'information pertinente à partir d'un

entrepôt de données cliniques de grande dimension

Article associé

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

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

Publié dans Journal of Clinical Medicine, 2019.

Code Python

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

Soutenance de thèse 6/27

Réadmissions

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

Facteurs prédictifs de la réhospitalisation pour CVO

Soutenance de thèse 7/27

Contexte

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

Soutenance de thèse 7/27

Introduction

Trajectoire

Méthod

Résultats

Réadmissions

Contexte

Récultate

. .

1odèle

Conclusion

Inarsity

Méthode

Applications

......

Sinacox

Applications

Conclusion

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

Soutenance de thèse 7/27

Contexte

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

Soutenance de thèse 7/27

Contexte

- 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

Soutenance de thèse 7/27

Introduction

Frajectoire

Méthod

Résultats

Réadmissions

Contexte

Récultate

Modèle

Applications

Conclusion

inarsity

Applications

Binacox

Applications

Conclusion

- 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
 - Péhospitalisation "précoce" en se basant sur un seuil ϵ pré-défini

Soutenance de thèse 7/27

Introduction

rajectoires

Méthod

Résultats

Réadmissions

Contexte

Résultats

...

lodèle

applications

inarsity

Méthode Applications

inacox

Applications

Conclusion

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

Deux cadres

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

Soutenance de thèse 7/27

Introduction

Trajectoires

Méthod

Résultats

Réadmissions

Contexte

Résultats

...

lodèle

Applications

inarsity

Méthode Applications

Applications Conclusion

inacox

Applications
Conclusion

Conclusion

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

Soutenance de thèse 7/27

Introduction

Trajectoires

Méthod

B/ 1 1

Réadmissions

Contexte

Résultats

HIIX

odèle

Applications

inarsity

Applications

Binacox

Application

Conclusion

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

Objectif

Comparer des méthodes d'apprentissage issues de ces deux cadres

Introduction

Frajectoires

Méthode

Réadmission

Contexte

Résultats

-mix

nodele

onclusion

inarsity

Applications

.

Méthode

Application Conclusion

Conclusion

► Étude rétrospective monocentrique sur la cohorte de l.

Soutenance de thèse 8/27

ntroduction

Trajectoires

Méthode

Résultats

Réadmissions

Contexte Méthode

Récultate

lodèle

Applications

Rinarsity

Méthode Applications

Binacox

Méthode Applications

Conclusion

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

Soutenance de thèse 8/27

Introduction

Trajectoire

Méthod

Réadmissions

ontexte

Méthode

...

/lodèle

Modele Applications

Binarsity

Méthode Applications

Binacox

Méthode Applications

Conclusion

- Étude rétrospective monocentrique sur la cohorte de l.
- 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)

Soutenance de thèse 8/27

IIIIIOductioi

Trajectoire

Méthodo

Réadmissions

Contexte

Résultat

Modèle

Applications

inarsity

Applications

inacox

Méthode Applications

Conclusion

- ▶ Étude rétrospective monocentrique sur la cohorte de l.
- 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]

Soutenance de thèse 8/27

IIIIIOductioi

Trajectoires

Méthod

Réadmissions

Contexte

Méthode

Résultat

L-MIX

Applications

Conclusion

inarsity

Applications

Conclusion

Binacox

Applicatio

Conclusion

- Étude rétrospective monocentrique sur la cohorte de l.
- 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

Soutenance de thèse 8/27

Introduction

Trajectoires

Méthod

Réadmissions

ontexte

Méthode Résultats

-

Modèle

Applications

inarsity

Máthoda

Applications Conclusion

Binacox

Méthode

Conclusion

Conclusion

- ▶ Étude rétrospective monocentrique sur la cohorte de l.
- 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)

Soutenance de thèse 8/27

Introduction

Trajectoires

Méthod

Réadmissions

ontexte

Méthode Résultats

-

Modèle

Applications

inarsity

Máthoda

Applications

D:

Binacox

Application Conclusion

Conclusion

Réadmissions

Contexte Méthode

Récultate

Resultat

M- 41

Applications

LOTICIDSION

Binarsity

Applications

Rinacov

Sinacox

Méthode Applications

Conclusion

Conclusion

Références

▶ Étude rétrospective monocentrique sur la cohorte de l.

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)

Modèles considérés

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

Soutenance de thèse 9/27

minoduction

Trajectoires

Méthod

Résultats

Réadmissions

Contexte Méthode

Dácultate

_mix

Modèle

Applications

inarsity

Méthode Applications

Binacox

Méthode Applications

. . .

Lonciusion

Modèles considérés

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

Soutenance de thèse 9/27

....

Trajectoires

Méthod

Résultats

Réadmissions

Contexte Méthode

Récultate

. .

Modèle

Applications

Binarsity

Méthode Applications

Binacox

Méthode Applications

Camalinatan

Modèles considérés

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

Soutenance de thèse 9/27

....

Trajectoires

Méthod

Résultats

Réadmissions

Contexte Méthode

Récultate

Modèle

Applications

inarsity

Applications

Binacox

Applications

Conclusion

Conclusion

Modèles considérés

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

Soutenance de thèse 9/27

Trajectoires

Méthode

Réadmissions

Contexte

Méthode

Resultats

C-mi

Modèle Applications

onclusion

inarsity

Applications

inacox

Méthode

Application Conclusion

Conclusion

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]

Soutenance de thèse 9/27

....

Trajectoires

Méthode

Réadmission

- Caulilissio

Méthode

Résultats

C-mix

Applications

.

Méthode

Applications Conclusion

inacox

Méthode Applications

Conclusion

Modèles considérés

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

Soutenance de thèse 9/27

.....

Trajectoires

Méthode

Réadmission

. . .

Méthode

Résultats

C-mix

Modèle

Applications Conclusion

inarsity

Applications

Conclusion

Binacox

Application:

Conclusion

Modèles considérés

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

Soutenance de thèse 9/27

Introduction

Trajectoires

Méthode

Résultats

Réadmissions

Contexte Méthode

Résultats

Cmiv

Modèle

Applications

Olicidsion

Binarsity

Méthode

Applications Conclusion

inacox

Applicatio

Modèles considérés

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

Soutenance de thèse 9/27

Introduction

Trajectoires

Méthode

Résultats

Réadmissions

Contexte Méthode

Résultats

C-mix

Modèle

Applications

tananatan.

Binarsity

Applications

inacox

Applicatio

Conclusion

Modèles considérés

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

Soutenance de thèse 9/27

Introduction

Trajectoires

Méthode

Résultats

Réadmissions

Contexte Méthode

Résultats

C-mix

Modèle

Applications

inarsity

Applications

Binacox

Application

Conclusion

Modèles considérés

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

Soutenance de thèse 9/27

Méthode

Máthoda

Modèles considérés

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

Comparaison des deux cadres

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

Résultats en prédiction

Table 1: Performances en prédiction

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

Soutenance de thèse 10/27

Introduction

Trajectoire

Méthode

Réadmissions

.ontexte //éthode

Résultats

L-mix

Applications

inarsity

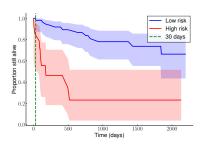
Méthode Applications

inacox

Méthode Applications

Conclusion

Résultats : le C-mix

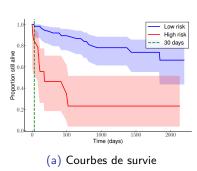


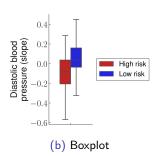
(a) Courbes de survie

Soutenance de thèse 11/27

Résultats

Résultats : le C-mix





Soutenance de thèse 11/27

Introduction

Trajectoires

Méthode

Réadmissions

Méthode

Résultats

C-mix

Applications

inarsity

Méthode Applications

Binacox

Applications Conclusion

Conclusion

Conclusion

Modèles d'analyse de survie \to fonctions de survie estimées \to prédictions binaires pour un ϵ donné

Soutenance de thèse 12/27

.....

rajectoire

-

Résultat

Réadmission:

Lontexte

Résultats

C-mi

Modèle

Applications

Rinarsity

Méthode Applications

Applications

inacox

Méthode

Conclusion

Conclusion

Conclusion

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

Soutenance de thèse 12/27

Résultats

Conclusion

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

Soutenance de thèse 12/27

meroductio

Fraiectoire

Méthode

Résultats

Réadmission

Méthode

Résultats

C-mi

Modèle

Applications

Binarsity

Méthode

Applications

inacox

Méthode

Conclusion

Conclusion

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. Rangue et A.S. Jannot

Publié dans BMC Medical Research Methodology, 2019.

Récultate

Comparison of methods for early-readmission prediction in a high dimensional heterogeneous covariates and

time-to-event outcome framework

Méthodologie pour la création de features pertinentes

▶ Bonnes performances du C-mix + aspects d'interprétation intéressants

Récultate

Article associé

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

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

Publié dans BMC Medical Research Methodology, 2019.

Code Python

Disponible à

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

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

Soutenance de thèse 12/27

Introduction

rajectoires

Méthod

Résultat

Réadmissions

Contexte

Résulta

C-mix

Modà

Application:

Conclusion

inarsity

Méthode

Applications Conclusion

Binacox

Méthode

Applicatio Conclusion

Conclusion

Le modèle

Contexte de l'analyse de survie

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

Soutenance de thèse 13/27

Modèle

Le modèle

Contexte de l'analyse de survie

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

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

Soutenance de thèse 13/27

minoduction

Frajectoires

Méthode

Résultats

Réadmissions

Contexte

Résultats

C-mi>

Modèle

pplications

narsity

Méthode Applications

onclusion

Sinacox

Applications Conclusion

Conclusion

Le modèle

Contexte de l'analyse de survie

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

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

Soutenance de thèse 13/27

.....

rajectoires

Méthode

Résultats

Réadmissions

Contexte

Résultats

Modèle

l**odele** pplications

......

inarsity

Applications

inacox

Méthode Applications

Conclusion

Le modèle

Contexte de l'analyse de survie

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

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

Soutenance de thèse 13/27

milioduction

rajectoires

Méthode

Résultats

Réadmissions

Zontexte Véthodo

Resultats

-mix

Modèle

pplications

inarsity

Applications

.

inacox

Applications Conclusion

Conclusion

Le modèle

Contexte de l'analyse de survie

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

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

Soutenance de thèse 13/27

Introduction

ajectoir

Méthode

Dánduninai

Réadmissions

ontexte láthoda

Modèle

Modele Application

Application

inarsity

Méthode Applications

inacox

Méthode Applications Conclusion

Conclusion

Le modèle

Contexte de l'analyse de survie

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

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

$$\qquad \qquad \pi_{\beta_k}(x) = \mathbb{P}[Z = k | X = x] = \frac{e^{x^\top \beta_k}}{\sum_{k=0}^{K-1} e^{x^\top \beta_k}} \quad \text{(softmax)}$$

- Échantillon *i.i.d.* $(x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n) \in \mathbb{R}^d \times \mathbb{R}_+ \times \{0, 1\}$
- $\theta = (\alpha_0, \dots, \alpha_K, \beta_0, \dots, \beta_K)^{\top}$, log-vraissemblance du C-mix

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

Soutenance de thèse 13/27

Introduction

aiectoire

Méthode

57 1 . . .

Réadmissions

léthode

-mix

Modèle

Application

inarcity

inarsity

Applications Conclusion

Binacox

Application

Conclusion

L'algorithme QNEM

Objectif pénalisé

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

Soutenance de thèse 14/27

Trajectoires

Méthode

_

Réadmissions

Méthode

Résultats

C-mix

Modèle

pplications

inarsity

Applications

Rinacox

Méthode Applications

Conclusion

L'algorithme QNEM

Objectif pénalisé

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

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

$$-n^{-1} \sum_{i=1}^{n} \left\{ \delta_{i} \left[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_{i}=k\}} \left(\log \pi_{\beta_{k}}(x_{i}) + \log f_{k}(y_{i}; \alpha_{k}) \right) + \log \bar{G}(y_{i}^{-}) \right] \right.$$

$$\left. + (1 - \delta_{i}) \left[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_{i}=k\}} \left(\log \pi_{\beta_{k}}(x_{i}) + \log \bar{F}_{k}(y_{i}^{-}; \alpha_{k}) \right) + \log g(y_{i}) \right] \right\}$$

Soutenance de thèse 14/27

Introduction

Trajectoires

Contexte

resures

Réadmissions

Contexte

Résultats

-mix

Modèle

pplications onclusion

inarsity

Méthode Applications Conclusion

Sinacox

Méthode Applications Conclusion

Conclusion

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

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

$$-n^{-1} \sum_{i=1}^{n} \left\{ \delta_{i} \left[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_{i}=k\}} \left(\log \pi_{\beta_{k}}(x_{i}) + \log f_{k}(y_{i}; \alpha_{k}) \right) + \log \bar{G}(y_{i}^{-}) \right] + (1 - \delta_{i}) \left[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_{i}=k\}} \left(\log \pi_{\beta_{k}}(x_{i}) + \log \bar{F}_{k}(y_{i}^{-}; \alpha_{k}) \right) + \log g(y_{i}) \right] \right\}$$

$$\begin{split} & \text{ \'etape E: } Q_n(\theta, \theta^{(I)}) = \mathbb{E}_{\theta^{(I)}}[\ell_n^{\mathsf{comp}}(\theta)|\mathbf{y}, \boldsymbol{\delta}] \\ & q_{i,k}^{(I)} = \mathbb{E}_{\theta^{(I)}}[\mathbb{1}_{\{z_i = k\}}|y_i, \delta_i] = \mathbb{P}_{\theta^{(I)}}[z_i = k|y_i, \delta_i] = \frac{\Lambda_{k,i}^{(I)}}{\sum_{r=0}^{K-1} \Lambda_{r,i}^{(I)}} \\ & \Lambda_{k,i}^{(I)} = \left[f_k(y_i; \alpha_k^{(I)}) \bar{G}(y_i^-)\right]^{\delta_i} \left[g(y_i) \bar{F}_k(y_i^-; \alpha_k^{(I)})\right]^{1-\delta_i} \pi_{\beta_k^{(I)}}(x_i) \end{split}$$

Soutenance de thèse 14/27

Introduction

Méthode

Résultats

Réadmissions

ontexte

Résultats

mix

Modèle Applications

onclusion

inarsity

Méthode Applications Conclusion

Binacox

Applications Conclusion

Conclusio

2/6/

L'algorithme QNEM

Étape M: problème d'optimisation convexe...

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

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

Soutenance de thèse 15/27

Frajectoires

Méthode

Résultats

Réadmissions

ontexte

Kesultats

-mix

Modèle

pplications

narsity

Applications

inacov

Méthode

Application

Conclusion

L'algorithme QNEM

Étape M: problème d'optimisation convexe...

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

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

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

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

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

Soutenance de thèse 15/27

·___

Méthode Résultats

Réadmissions

ontexte

Résultats

Modèle

pplications

narsity

éthode

Applications Conclusion

Binacox

Méthode Applications

Conclusion

Soutenance de thèse 15/27

Étape M: problème d'optimisation convexe...

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

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

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

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

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

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

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

Introduction

rajectoires

Contexte

Résultats

éadmissions

ontexte

resultats

Modèle

pplications

narsity

itarSity éthodo

Applications

Binacox

Méthode Applications

Conclusion

Soutenance de thèse 15/27

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

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

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

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

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

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

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

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

Convergence vers un minimum local prouvée

IIILIOUUCLIOII

raioctairec

Contexte Méthode

Résultats

éadmissions

ontexte

Résultats

mix

Modèle Applications

arsity

narsity

Applications

inacox

Méthode Applications

Conclusion

Application sur données génétiques

Soutenance de thèse 16/27

ntroduction

Trajectoires

Méthode

éadmissions

Contexte

Résultats

Modèle

Applications

Conclusion

inarsity

Applications

inacox

illacox

Applicatio

onclusion

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

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

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

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

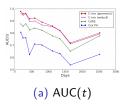


Figure 3: Résultats sur le cancer BRCA

ntroduction

....

Méthod

Résultats

Contexte

Résultats

C-mix

, IIIIX

Applications

Conclusion

inarsity

Méthode

Applications Conclusion

Binacox

Application

Conclusion

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

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

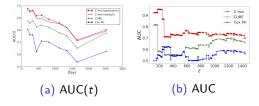


Figure 3: Résultats sur le cancer BRCA

ntroduction

Méthode

Résultats

eadmissio

Methode Pácultata

-mix

-11111

Applications

Conclusion

Binarsity

Méthode

Applications Conclusion

Binacox

Méthode Application

Conclusion

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

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

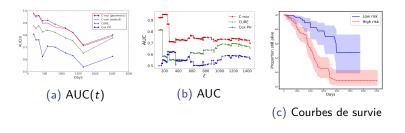


Figure 3: Résultats sur le cancer BRCA

ntroduction

Contexte Méthode

Résultats

ntexte

Résultats

C-mix

Applications

Conclusion

Binarsity

Applications
Conclusion

Binacox

Méthode Applications

Conclusion

Conclusion

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

Soutenance de thèse 17/27

Conclusion

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

Soutenance de thèse 17/27

Introduction

Trajectoires

Méthod

Réadmissions

ontexte

Résultats

-mix

Modèle Inplications

Conclusion

inarsity

Methode Applications

inacox

Méthode Applications

Conclusion

Réadmissions

Résultats

-mix

Modèle Applications

Conclusion

Sinarsity

Applications

Binacox

Méthode Applications

onclusion.

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

Article associé

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

Réadmissions

Méthode

-mix

Modèle Applications

Conclusion

Dillarsity

Méthode Applications

Binacox

Méthode

Applicatio 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

Article associé

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

Code Python

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

Méthode

Réadmissions

Méthode

C-mix

Modèle Applications

Conclusion

Dinarsity

Application Conclusion

Binacox

Méthode Application

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

Article associé

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

Code Python

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

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

Soutenance de thèse 17/27

Binarsity

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

Soutenance de thèse 18/27

Introduction

rajectoires

Méthode

Résultats

Réadmission:

Contexte

Résultat

C-mix

Modèle

pplications

Binarsity

Méthode

Applications

inacox

macox

Applications

Conclusion

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

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

Soutenance de thèse 18/27

minoduction

Trajectoires

Méthode

Réadmissions

Méthode Résultats

C miv

Modèle Applications

inarsity

Méthode

Applications

inacox

Méthode Applications

Conclusion

- ▶ Apprentissage supervisé $(x_i, y_i)_{i=1,...,n}$ 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 $k = 1, \dots, d_i$, on a

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

```
X_{\bullet,1}
\vdots
12,5
9,2
3,1
8,7
\vdots
```

Soutenance de thèse 18/27

....

Frajectoire

Méthode

Réadmissions

Contexte

Résultats

-mix

Modèle

Applications Conclusion

inarsity

Méthode

Applications

inacox

Méthode Applications

Conclusion

Conclusion

- Apprentissage supervisé $(x_i, y_i)_{i=1,...,n}$ avec x_i continues
- ► Encodage one-hot $\begin{bmatrix} 22 \end{bmatrix}$: $x_i = (x_{i,1}, \dots, x_{i,p})^{\top} \in \mathbb{R}^p$ transformé en $x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_1}^B, x_{i,2,1}^B, \dots, x_{i,2,d_2}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_n}^B)^{\top} \in \mathbb{R}^d,$ t.g. pour $i = 1, \ldots, n$ et $k = 1, \ldots, d_i$, on a

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

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

9.2

3, 1

8,7

Soutenance de thèse 18/27

Méthode

- ▶ Apprentissage supervisé $(x_i, y_i)_{i=1,...,n}$ 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.g. pour $i = 1, \dots, n$ et $k = 1, \dots, d_i$, on a

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

\vdots $i \to 12, 5$ 9, 2 3, 1 8, 7 \vdots

Soutenance de thèse 18/27

.....

Trajectoires

Méthode

Réadmissions

ontexte

Résultats

mix

Applications

Conclusion

inarsity

Méthode Application

Applications

inacox

Application

Conclusion

- Apprentissage supervisé $(x_i, y_i)_{i=1,...,n}$ 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.g. pour $i = 1, \dots, n$ et $k = 1, \dots, d_i$, on a

$$x_{i,j,k}^B = egin{cases} 1 & ext{ si } x_{i,j} \in I_{j,k} = \left(q_j(rac{k-1}{d_j}), q_j(rac{k}{d_j})
ight], ext{ (interquantiles)} \ 0 & ext{ sinon} \end{cases}$$

meroductio

Trajectoires

Méthode

Réadmissions

ontexte

Résultats

:-mix

Applications

inarsity

Méthode Application

Applications

Binacox

Application: Conclusion

Conclusion

Apprentissage supervisé $(x_i, y_i)_{i=1,...,n}$ 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 $k = 1, \dots, d_i$, on a

$$x_{i,j,k}^{\mathcal{B}} = egin{cases} 1 & ext{ si } x_{i,j} \in I_{j,k} = \left(q_j(rac{k-1}{d_j}), q_j(rac{k}{d_j})
ight], ext{ (interquantiles)} \\ 0 & ext{ sinon} \end{cases}$$

miroductio

Trajectoires

Méthode

Réadmissions

intexte

Kėsultats

-mix

Applications

narsity

Méthode

Applications

inacox

Méthode Applications

Conclusion

Conclusio

► Encodage one-hot [22]: $x_i = (x_{i,1}, \dots, x_{i,p})^{\top} \in \mathbb{R}^p$ transformé en $x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_i}^B, x_{i,2,1}^B, \dots, x_{i,2,d_2}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_p}^B)^{\top} \in \mathbb{R}^d$,

t.q. pour
$$i=1,\ldots,n$$
 et $k=1,\ldots,d_j$, on a
$$\int 1 \quad \text{si } x_{i,j} \in I_{j,k} = \left(q_j(\tfrac{k-1}{d_i}),q_j(\tfrac{k}{d_i})\right], \text{ (interquantiles)}$$

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

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

maroduction

Frajectoires

Méthode Résultats

Réadmissions

léthode ésultats

C-mix

Modèle Applications

Applications Conclusion

Méthode

Application

Applications

Binacox

Méthode

Conclusion

Conclusion

► 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.g. pour $i = 1, \dots, n$ et $k = 1, \dots, d_i$, on a

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

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

$$y|x\mapsto \exp\Big(\frac{ym^0(x)-b\big(m^0(x)\big)}{\phi}+c\big(y,\phi\big)\Big),$$
 et $\ell\big(y_1,y_2\big)=-y_1y_2+b\big(y_2\big)$

Introduction

Trajectoires

Méthode

léadmissions

léthode ésultats

-mix

Applications

narsity

Méthode

Applications

Binacox

Application: Conclusion

Conclusion

Problèmes

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

Soutenance de thèse 19/27

Introduction

Frajectoires

Méthode

resultats

Réadmissions

Contexte

Résultats

C-mi

Vlodèle

pplications onclusion

Binarsity

Méthode Applications

Applications

inacox

Méthode

Conclusion

Conclusion

Problèmes

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

Soutenance de thèse 19/27

Méthode

Problèmes

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

Soutenance de thèse 19/27

Méthode

Problèmes

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

Réponses

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

Soutenance de thèse 19/27

Introduction

Trajectoire

Méthod

Résultats

Réadmissions

Contexte

Résultats

C-mix

Modèle Applications

Applications

Binarsity

Méthode

Applications

inacox

sinacox

Application

Conclusion

Problèmes

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

Réponses

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

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

Soutenance de thèse 19/27

meroductio

Frajectoire

Méthode

Résulta

Réadmissions

Contexte

Kesultats

-mix

-11111

Applications

onclusion

inarsity

Méthode Applications

Applications

linacox

Méthode

Conclusio

Conclusion

Problèmes

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

Réponses

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

Soutenance de thèse 19/27

Introduction

Trajectoir (

Méthode

. Réadmissions

ontexte

Résultats

C-mix

Applications

inarsity

Méthode

Applications

Conclusion

Binacox

Méthode

Conclusion

Conclusion

Problèmes

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

Réponses

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

Pénalité binarsity

$$\blacktriangleright \ \mathsf{bina}(\beta) = \textstyle \sum_{j=1}^p \bigg(\sum_{k=2}^{d_j} \hat{w}_{j,k} |\beta_{j,k} - \beta_{j,k-1}| + \delta_1(\beta_{j,\bullet}) \bigg),$$

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

Soutenance de thèse 19/27

Introduction

Trajectoire:

Méthod

Réadmission

eadmissions . . .

Résultats

-mix

Applications

onclusion

Sinarsity

Méthode

Applications

Sinacox

Applications Conclusion

Conclusion

Problèmes

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

Réponses

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

Pénalité binarsity

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

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

Soutenance de thèse 19/27

Introduction

rajectoires

Méthode

Résultats

Réadmissions

ontexte

-mix

Applications

Conclusion

inarsity

Méthode

Applications

Conclusion

linacox

Applicatio

Conclusion

Résultat théorique

Fonction de risque associée [21]

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

Soutenance de thèse 20/27

Frajectoires

Méthode

Résultat

Réadmissions

Contexte

Résultats

C-mi>

Modèle

nciusion

Sinarsity

Méthode Applications

Applications

inacox

Méthode Applications

Conclusion

Conclusion

Résultat théorique

Fonction de risque associée [21]

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

$$ightharpoonup J(eta) = \left[J_1(eta), \ldots, J_p(eta)\right]$$
, avec

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

Soutenance de thèse 20/27

....

Frajectoire:

Méthode

Réadmissions

ontexte

Résultats

_-mix

Modele Applications

. .

inarsity

Méthode Applications

Applications

inacox

Méthode Applications

Conclusion

Résultat théorique

Fonction de risque associée [21]

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

 $ightharpoonup J(eta) = \left[J_1(eta), \ldots, J_p(eta)\right]$, avec

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

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

Soutenance de thèse 20/27

meroduction

Frajectoire

Méthode

Réadmissions

ntexte

Résultats

.=HIIX

Applications

Binarsity

Méthode Applications

Applications

inacox

Méthode

Conclusion

Conclusion

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

$$lacksquare$$
 $J(eta) = ig[J_1(eta), \ldots, J_p(eta)ig]$, avec

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

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

Inégalité oracle non-asymptotique à vitesse rapide

Avec grande probabilité, on a

$$R(m_{\hat{\theta}}) - R(m^{0}) \leq (1 + c_{1}) \inf_{\substack{\theta \in \mathscr{B}_{d}(R) \\ \forall j, \ 1^{\top}\theta_{j, \bullet} = 0 \\ |J(\theta)| \leq J^{*}}} \left\{ R(m_{\theta}) - R(m^{0}) + c_{2} \frac{|J(\theta)| \log d}{n} \right\}$$

avec $c_1, c_2 > 0$

Méthode

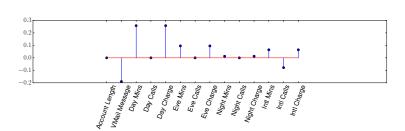


Figure 4: LR sur "Churn" (n = 3333, p = 14)

Soutenance de thèse 21/27

Introduction

Frajectoires

Méthode

resultats

Réadmissions

Contexte

Resultats

l-mix

Vloděle

Applications

inarsity

Méthode

Applications

Conclusion

Binacox

Méthode

. .

Dáfárancac

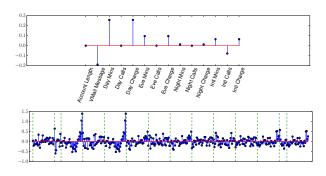


Figure 4: LR sur "Churn" (n = 3333, p = 14)

Soutenance de thèse 21/27

Introduction

Trajectoires

Méthod

Dánduninain

Contexte Véthode

. .

.-mix

Applications

inarsity

Méthode

Applications

Conclusion

Binacox

Máthada

Applications Conclusion

Conclusion

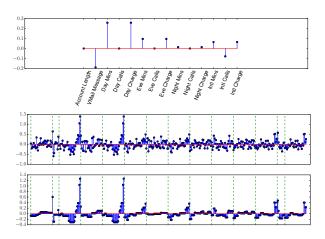


Figure 4: LR sur "Churn" (n = 3333, p = 14)

Soutenance de thèse 21/27

Introduction

raiectoires

Méthode

Réadmissions

. . .

Résultats

L-mix

Modèle

Conclusion

inarsity

Méthode

Applications Conclusion

......

Aáthoda

Applications Conclusion

Conclusion

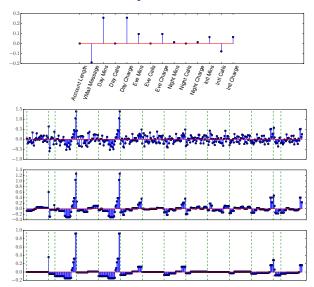


Figure 4: LR sur "Churn" (n = 3333, p = 14)

Soutenance de thèse 21/27

Introduction

.

Méthode

Kesultats

Réadmissions

ontexte

Résultats

_-mix

Modèle

Applications Conclusion

Binarsity

Méthode

Applications

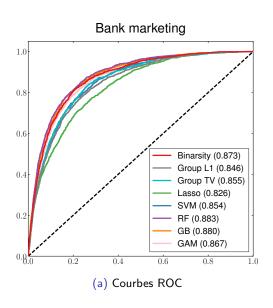
Sinacox

Méthode

onclusion

Conclusion

Applications



Soutenance de thèse 22/27

Introduction

Traiectoire

Méthoc

Résultats

Réadmissions

Contexte

resurence

mix

pplications

nclusion

narsity

léthode

Applications

Conclusion

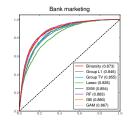
IIIaCOX

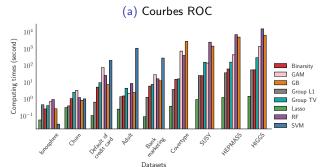
Applications

Conclusion

onclusion.

Applications





(b) Temps de calcul

Soutenance de thèse 22/27

Introduction

_ .

Máthod

Résultats

Réadmissions

eaumissions

Ivietnoo

C-mix

-IIIIX

Applications

onclusion

Binarsity

....

Applications

Conclusion

sinacox

Application Conclusion

Conclusion

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

Soutenance de thèse 23/27

Introduction

Trajectoires

Méthod

Résultats

Réadmissions

Contexte

Résultats

lodèle

nouele

Applications

Binarsity

Applications

Conclusion

Binacox

Méthode Applications

Conclusion

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

Soutenance de thèse 23/27

Conclusion

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

Soutenance de thèse 23/27

Introduction

Trajectoires

Méthod

Résultats

Réadmissions

Contexte

Résultats

C-mi

Modèle

onclusion

Binarsity

Méthode Applications

Conclusion

Binacox

Méthode

Application Conclusion

Conclusion

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

Soutenance de thèse 23/27

IIILIOGUCLIOI

Trajectoires

Contexte

Résultat

Réadmissions

Contexte

Résultats

C-mi

Modèle

pplications

linarcity

Méthode Applications

Conclusion

Binacox

Méthode

Conclusion

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** Publié dans *Journal of Machine Learning Research*, 2019.

Soutenance de thèse 23/27

Introduction

Trajectoire:

Méthod

.....

Réadmissions

Contexte

Résultats

-mix

1odèle

Conclusion

onclusion

....

Applications

Conclusion

Binacox

Dillaco,

Applicatio Conclusion

Conclusion

Resultats

Réadmissions

ontouto

Résultats

C-mix

Modèle

Applications

Conclusion

omarsity

Methode

Conclusion

Conclusion

Binacox

Méthode

Applicatio

onclucion

Dáfárancac

 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** Publié dans *Journal of Machine Learning Research*, 2019.

Code Python/C++

▶ Disponible à https://github.com/SimonBussy/binarsity

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 Publié dans Journal of Machine Learning Research, 2019.

Code Python/C++

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

Soutenance de thèse 23/27

IIIIIOductioi

rajectoires

Méthod

Réadmissions

Contexte

Résultat

_ .

Modèle

Applications

Binarsity

Méthode Applications

Applications Conclusion

Binacox

Méthode Applications

Conclusion

Références

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

lacktriangle Biomarqueurs ightarrow décision clinique : choix de seuils

Soutenance de thèse 24/27

IIItroductioi

Traiectoire:

Contexte

Résultat

Réadmission

Contexte

Récultate

/loděle

A 11 ...

Applications

onclusion

Binarsity

Méthode Applications

Conclusion

Binacox

Méthode

Application

onclusion

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

Soutenance de thèse 24/27

Introductio

Trajectoires

Méthode

Réadmission

ontexte

Résultats

C-mi>

Modele Applications

Sinarsity

Applications

Binacox

Méthode

Applications Conclusion

Conclusion

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

Soutenance de thèse 24/27

IIIIIOductioi

Trajectoires

Contexte

Résultats

Réadmissions

Contexte

Résultats

C-mix

Aodele Applications

pplications

Binarsity

Applications

Binacox

Méthode

Applications

Conclusion

- ightharpoonup Biomarqueurs ightarrow 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,

Soutenance de thèse 24/27

meroductio

Trajectoires

Méthode

Résultat

Réadmissions

Contexte

Résultats

miv

1odèle

onclusion

Sinarsity

Méthode Applications

Conclusion

inacox

Méthode

Application

onclusion

- ightharpoonup Biomarqueurs ightarrow 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,

Soutenance de thèse 24/27

.....

rajectoires

Méthode

Résultats

Réadmissions

ontexte éthode

Résultats

Ľ-mix

Modèle Applications

onclusion

Binarsity

Applications

Binacox

Méthode

Application: Conclusion

Conclusion

- ▶ 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

Soutenance de thèse 24/27

Introduction

rajectoires

Méthode

Dándmissis

Réadmissions

Lontexte Vléthode

Résultats

C-mi>

Modèle Applications

Applications Conclusion

inarsity

Applications

Binacox

Méthode

Conclusion

Conclusion

- ▶ 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
- $ightharpoonup Z = T \wedge C$, $\Delta = \mathbb{1}(T \leq C)$ et $X \in \mathbb{R}^p$

Soutenance de thèse 24/27

Introduction

Trajectoires

Méthode

Réadmission:

Contexte

Résultats

C-mix

Modèle Applications

onclusion

inarsity

Applications

Binacox

Méthode

Conclusion

Conclusion

- ▶ 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
- $ightharpoonup Z = T \wedge C, \ \Delta = \mathbb{1}(T \leq C) \ \text{et} \ X \in \mathbb{R}^p$
- Risque instantané pour un patient i donné par

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

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

Soutenance de thèse 24/27

Introduction

Trajectoires

Méthode

Réadmissions

ontexte

Résultats

-mix

Applications

onclusion

ilaisity

Applications Conclusion

Binaco

Méthode

Conclusion

Conclusion

- ightharpoonup Biomarqueurs ightarrow 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
- $ightharpoonup Z = T \wedge C$, $\Delta = \mathbb{1}(T \leq C)$ et $X \in \mathbb{R}^p$
- Risque instantané pour un patient i donné par

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

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

But : estimer simultanément

Soutenance de thèse 24/27

Introduction

Trajectoires

Méthode

Réadmissions

.____

Résultats

:-mix

Applications

Conclusion

narsity

Applications Conclusion

Binaco

Méthode

Conclusio

Conclusion

- ightharpoonup Biomarqueurs ightarrow 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
- $ightharpoonup Z = T \wedge C$, $\Delta = \mathbb{1}(T \leq C)$ et $X \in \mathbb{R}^p$
- Risque instantané pour un patient i donné par

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

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

► But : estimer simultanément

Soutenance de thèse 24/27

Introduction

Trajectoires

Méthode

Réadmissions

eadmission

Résultats

-mix

Applications

Conclusion

marsity

Applications Conclusion

Binaco

Méthode Applicatio

Conclusion

D.(((....

- ightharpoonup Biomarqueurs ightarrow 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
- $ightharpoonup Z = T \wedge C$, $\Delta = \mathbb{1}(T \leq C)$ et $X \in \mathbb{R}^p$
- Risque instantané pour un patient i donné par

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

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

► But : estimer simultanément

$$\qquad \qquad \mu^{\star} = (\mu_{1,1}^{\star}, \dots, \mu_{1,K_{\bullet}^{\star}}^{\star}, \dots, \mu_{p,1}^{\star}, \dots, \mu_{p,K_{p}^{\star}}^{\star})^{\top} \in \mathbb{R}^{K^{\star}}$$

$$\beta^{\star} = (\beta_{1,1}^{\star}, \dots, \beta_{1,K_1^{\star}+1}^{\star}, \dots, \beta_{p,1}^{\star}, \dots, \beta_{p,K_p^{\star}+1}^{\star})^{\top} \in \mathbb{R}^{K^{\star}+p}$$

Soutenance de thèse 24/27

Introduction

Trajectoires

Méthode

Réadmissions

eadmissions

Résultats

-mix

Modele Applications

Conclusion

narsity

Applications Conclusion

Binacox

Méthode Application

Conclusion

Conclusion

- ightharpoonup Biomarqueurs ightarrow 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
- $ightharpoonup Z = T \wedge C, \ \Delta = \mathbb{1}(T \leq C) \ \text{et} \ X \in \mathbb{R}^p$
- Risque instantané pour un patient i donné par

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

où
$$I_{i,k}^{\star} = (\mu_{i,k-1}^{\star}, \mu_{i,k}^{\star}]$$
 pour $k \in \{1, \dots, K_i^{\star} + 1\}$

But : estimer simultanément

$$\beta^{\star} = (\beta_{1,1}^{\star}, \dots, \beta_{1,K_1^{\star}+1}^{\star}, \dots, \beta_{p,1}^{\star}, \dots, \beta_{p,K_p^{\star}+1}^{\star})^{\top} \in \mathbb{R}^{K^{\star}+p}$$

$$K^* = \sum_{i=1}^p K_i^*$$

Soutenance de thèse 24/27

Introduction

rajectoires

Méthode

Résultats

éadmissions

ntexte

c

-IIIIA

Applications

onclusion

al-a-

Applications Conclusion

Binacox

Méthode

Conclusion

Conclusion

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

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

οù

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

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

Soutenance de thèse 25/27

.....

Frajectoire

Méthode Résultats

Réadmissions

Méthode

~ miv

∕lodèle

Applications

narsity

Méthode Applications Conclusion

пасох

Méthode Applications

Conclusion

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

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

Soutenance de thèse 25/27

minoduction

Trajectoire

Méthod

Résultats

Réadmissions

Contexte Véthode

Résultats

C-mi

Modèle Applications

Conclusion

inarsity

Méthode Applications

Binacox

Méthode

Application

Conclusion

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

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

Soutenance de thèse 25/27

Introduction

Trajectoires

Méthode

Résultats

Réadmissions

Contexte

.....

-mix

Applications

Conclusion

inarsity

Méthode

Applications Conclusion

inacox

Méthode

Applicatio

Conclusion

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

Soutenance de thèse 25/27

Introduction

Trajectoire

Contexte

Résultats

Réadmissions

Contexte

Résultats

C-mix

Modèle

Applications

inarcity

Méthode

Applications Conclusion

inacox

Méthode

Applicatio

Conclusion

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

Soutenance de thèse 25/27

Introduction

Trajectoire

Méthode

D.C. Jackson

-

Máthoda

Kesultats

-mix

Modèle Applications

Applications

inarsity

Applications Conclusion

3inacox

Méthode

Applicatio

Conclusion

La méthode

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

Soutenance de thèse 25/27

Introduction

Trajectoire

Méthode

Réadmissions

Contexte

Résultats

-mix

Applications

Conclusion

inarsity

Applications Conclusion

Binaco

Méthode

Applicatio Conclusion

Conclusion

- ▶ 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-vraissemblance négative partielle $\ell_n(f_eta)$
- $\qquad \qquad \hat{\beta} \in \mathsf{argmin}_{\beta \in \mathscr{B}_{p+d}(R)} \left\{ \ell_n(f_\beta) + \mathsf{bina}(\beta) \right\}$
- $\blacktriangleright \ J_j(\hat{\beta}) = \left\{I: \beta_{j,l} \neq \beta_{j,l-1}, \mathsf{pour}\ I = 2, \ldots, d_j + 1\right\} = \{\hat{l}_{j,1}, \ldots, \hat{l}_{j,s_j}\}$
- ightharpoons $\widehat{\mu}_{j,ullet} = (\mu_{j,\widehat{l}_{j,1}},\dots,\mu_{j,\widehat{l}_{j,s_j}})^{ op}$ avec $s_j = |J_j(\hat{eta})| = \widehat{K}_j$

Inégalité oracle non-asymptotique à vitesse rapide

Avec grande probabilité, on a

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

avec $c_1, c_2 > 0$

Introduction

rajectoires

Méthode

Résultat

Réadmissions

thode

.

lodèle

Applications

marsity

Applications Conclusion

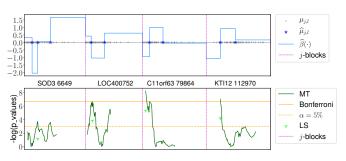
Binaco

Méthode

Applicatio

Conclusion

Applications sur données réelles (TCGA)



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

Soutenance de thèse 26/27

Introduction

Trajectoire

Méthode

Réadmissions

ontexte

Résultats

-mix

LINE

applications

Conclusion

inarsity

Applications

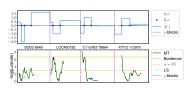
Binacox

Méthode

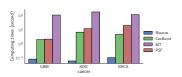
Applications Conclusion

Conclusion

Applications sur données réelles (TCGA)



(a) Résultats sur le cancer GBM

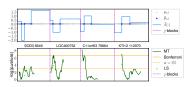


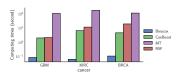
(b) Temps de calcul

Soutenance de thèse 26/27

Applications

Applications sur données réelles (TCGA)





(a) Résultats sur le cancer GBM

(b) Temps de calcul

Table 3: Comparison des C-index

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

Soutenance de thèse 26/27

Applications

Conclusion

 Détection multivariée de plusieurs cut-points par covariable continue

Soutenance de thèse 27/27

Introduction

rajectoires

M/AL-J-

Résultats

Réadmissions

Contexte

Résultats

-

lodèle

Applications

Rinarsity

Méthode Applications

Applications Conclusion

Binacox

Applications

Conclusion

Conclusion

Conclusion

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

Soutenance de thèse 27/27

Introduction

Trajectoires

Méthod

Résultats

Réadmissions

Contexte

Résultats

C-mix

/lodèle

pplications

Binarsity

Applications

inacox

Méthode Applications

Conclusion

onclusion

Conclusion

- Soutenance de thèse 27/27
- Introduction
 - rajectoire
- Méthod
 - Réadmissions
 - Contexte Méthode
-
- J-IIIIX
- Applications
- Conclusion
- tananatan i

inarsity

- Application: Conclusion
- Rinacov

DINACOX

Applications

Conclusion

Concidiion

Références

 Détection multivariée de plusieurs cut-points par covariable continue

► Implémentation rapide et bons résultats obtenus en estimation sur données simulées et réelles

Article associé

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

En revision dans Biometrics, 2020.

Réadmissions

Méthode

Modèle

Applications

Rinarcity

Méthode

Applications Conclusion

Binaco

Méthode

Applications

. . .

Dáfárancac

 Détection multivariée de plusieurs cut-points par covariable continue

 Implémentation rapide et bons résultats obtenus en estimation sur données simulées et réelles

Article associé

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

En revision dans Biometrics, 2020.

Code Python/C++

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

Soutenance de thèse 27/27

IIIIIOduction

ajectoires

Contexte

Résultat

Réadmissions

Contexte

Résultat

C-r

Modèle

Applications

Méthode

Applications

.

JIIIaCOX

Application

Conclusion

Références

VI. Conclusion générale

Soutenance de thèse 28/27

milioduction

rajectoires

Méthode

Résultats

Réadmissions

Contexte

Résultats

C-mix

/lodèle

Conclusion

Rinarsity

Méthode Applications

Applications

inacox

Méthode Applications

Conclusion

Références

Nouvelles méthodes

▶ Différentes méthodes de machine learning proposées

Soutenance de thèse 28/27

Conclusion

Nouvelles méthodes

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

Soutenance de thèse 28/27

Introduction

Trajectoires

Méthod

Resultats

Réadmissions

Contexte

Résultats

L-IIIIX Moděle

Applications

Binarsity

Méthode

Applications Conclusion

inacox

Méthode Applications

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.)

Soutenance de thèse 28/27

Introduction

Trajectoires

Méthode

B/ 1 1 1

Réadmissions

Méthode

Résultat

L-MIX

Modèle Applications

Sinarsity

Applications

inacox

Méthode Applications

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

C-mix Modèle

Applications

Binarsity

Méthode

Applications Conclusion

inacox

∕léthode

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

Implémentation

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

Références I

- [1] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7):1145-1159, 1997.
- Leo Breiman, Random forests, Machine learning, 45(1):5-32, 2001.
- 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.
- 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,
- David R Cox. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2):187-220, 1972.
- Vern T Farewell. The use of mixture models for the analysis of sureval data with long-term survivors. Biometrics, 38(4):1041-1046, 1982.
- 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.
- 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.
- T. Hastie and R. Tibshirani. Generalized additive models. Wiley Online Library, 1990. [10]
- [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.
- David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. Applied logistic regression, volume 398. John Wiley & Sons. 2013.
- J. P. Klein and M. L. Moeschberger. Survival analysis: techniques for censored and truncated data. Springer Science & Business Media, 2005.

Soutenance de thèse 28/27

- [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.
- L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. Journal of the [16] 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.
- 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.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267-288, 1996.
- S. van de Geer. High-dimensional generalized linear models and the Lasso. Ann. Statist., 36(2): [21] 614-645, 2008.
- J. Wu and S. Coggeshall. Foundations of Predictive Analytics (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series). Chapman & Hall/CRC, 1st edition, 2012.
- B Yegnanarayana. Artificial neural networks. PHI Learning Pvt. Ltd., 2009. [23]