Introduction of high-dimensional interpretable machine learning models and their applications

Simon Bussy

THÈSE DE DOCTORAT
DE SORBONNE UNIVERSITÉ

Spécialité : Statistique

École doctorale n°386: Sciences Mathématiques de Paris Centre

réalisée

au Laboratoire de Probabilités, Statistique et Modélisation

sous la direction de Agathe GUILLOUX, Anne-Sophie JANNOT et Stéphane GAÏFFAS

présentée par

Simon BUSSY

pour obtenir le grade de :

DOCTEUR DE SORBONNE UNIVERSITÉ

Sujet de la thèse :

Introduction of high-dimensional interpretable machine learning models and their applications

soutenue le 16/01/2019

devant le jury composé de :

Pr.	Stéphane GAÏFFAS	(1,2)	Examinateur
Pr.	Agathe GUILLOUX	(3)	Directrice de thèse
Dr.	Anne-Sophie JANNOT	(4,5)	Co-directrice de thèse
Pr.	Gregory NUEL	(1)	Examinateur
Dr.	Franck PICARD	(6)	Examinateur
Dr.	Raphaël PORCHER	(7)	Rapporteur
Pr.	Rodolphe THIEBAUT	(8,9)	Examinateur
Pr.	Jean-Philippe VERT	(10,11)	Rapporteur

[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. [6]LBBE, UMR CNRS 5558 Université Lyon 1, F-69622 Villeurbanne, France. [7]CRESS INSERM, UMR 1153, Université Paris Descartes, Paris, France. [8]SISTM, Inria Bordeaux Sud-Ouest, Epidémiologie et Biostatistique, Bordeaux, France. [9]INSERM, UMR 897, Université de Bordeaux, Bordeaux, France. [10]MINES ParisTech, Université de recherche PSL, CBIO-Centre for Computational Biology, Paris, France. [11]Google Brain, Paris, France.

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université - UPMC
Campus Pierre et Marie Curie
Tour 25, couloirs 15-25 & 15-16
$2^{\text{ème}}$ étage
4, Place Jussieu
75005 Paris

À mes parents, À Alex

Remerciements

La thèse de doctorat est un travail s'inscrivant dans la durée et qui constitue le fil conducteur de trois années intensives. De nombreuses personnes se retrouvent alors entre le doctorant et son doctorat, pour le pire ou le meilleur ! Ce sont ces personnes, m'ayant permis de mener à bien cette aventure, que j'aimerais mettre en avant dans ces quelques lignes de remerciement, sans doute trop courtes pour pouvoir toutes les citer.

Mes premiers remerciements s'adressent naturellement à mes directeurs : Agathe Guilloux, Professeur à l'université d'Évry Val d'Essonne, Anne-Sophie Jannot, docteur en génétique statistique et médecin spécialiste en santé publique, ainsi que Stéphane Gaïffas, Professeur à l'université Paris Diderot et Professeur chargé de cours à l'école polytechnique. Cela a été un véritable plaisir de travailler avec vous et de pouvoir profiter de vos connaissances pendant ces trois années. Vous vous êtes impliqués dans l'orientation de mes travaux et investis dans celle de mon avenir. C'est pourquoi ces quelques mots ne suffiront certainement pas à vous donner toute la mesure de ma gratitude.

Agathe, merci pour ton intérêt, ta confiance et ton soutien. J'espère un jour avoir ta rapidité de lecture d'un papier scientifique.

Anne-Sophie, merci pour ta disponibilité et tes nombreux conseils. J'ai découvert avec toi les rouages du monde de la recherche, merci aussi pour ces nombreuses discussions scientifiques et pour toutes les opportunités que tu m'as offertes.

Stéphane, merci de m'avoir toujours laissé des libertés d'initiative, merci aussi pour tes encouragements qui m'ont souvent mis en confiance. J'ai notamment appris avec toi à augmenter continuellement mon niveau d'exigence, et à être toujours plus ambitieux sur les objectifs à atteindre.

Merci au Professeur Jean-Philippe Vert d'avoir immédiatement accepté d'être rapporteur de mes travaux. J'en suis honoré et vous remercie pour votre investissement dans l'évaluation de mon travail et pour votre implication très positive.

Je remercie également le Professeur Raphaël Porchet d'avoir bien voulu être rapporteur. Je vous suis vraiment reconnaissant d'avoir porté votre regard expert sur mon manuscrit, merci pour vos remarques pertinentes pour l'améliorer.

Ce doctorat n'aurait pas été possible sans le soutien de l'Université Pierre et Marie Curie (UPMC, désormais Sorbonne Université), qui m'a permis, grâce à une bourse de thèse, de travailler sereinement durant ces trois ans. Je voudrais aussi remercier Gérard Biaud, Professeur à Sorbonne Université, qui a dirigé le LSTA ainsi que Lorenzo Zambotti, Professeur à Sorbonne Université et directeur actuel du LPSM, de m'y avoir accueilli avec les meilleures conditions de travail.

Je tiens ensuite à adresser mes remerciements à tous les membres du LPSM qui m'ont accompagné durant cette thèse, avec une pensée pour les doctorants qui ont fait vivre ce laboratoire et qui ont fait de Jussieu un endroit chaleureux et convivial. Je remercie également Corinne et Louise pour m'avoir permis d'organiser ma soutenance ainsi que pour leur gentillesse et leur efficacité au cours de ces trois ans.

Tout au long de ma thèse, j'ai pu rencontrer des personnes enthousiasmées par la recherche et toujours ravies de discuter et d'expliquer leurs travaux. Merci à elles, pour leur passion et pour les discussions enrichissantes qui s'en sont suivies. Je remercie tout particulièrement les personnes avec qui j'ai eu d'étroites collaborations, en commençant évidemment par Mokhtar : je te remercie pour tout ce que tu m'as apporté lorsque je débutais dans le monde des inégalités oracles, ou dans celui des processus de comptage. Je garderai toujours un super souvenir de nos nombreuses journées à reprendre au tableau les calculs des travaux que l'on a en commun.

Raphaël, ça a été un vrai plaisir de travailler avec toi. Tu m'as beaucoup apporté concernant l'interprétation clinique : merci pour ça, pour ta réactivité hors pair et pour ta bonne humeur constante.

Enfin Antoine, avec qui le courant est instantanément passé : notre collaboration naissante s'annonce fructueuse, merci de m'aider à parfaire mes connaissances sur les statistiques bayésiennes en pratique.

Au terme de ce parcours, je remercie enfin celles et ceux qui me sont chers et que j'ai pu parfois délaisser faute de temps. Leurs attentions et encouragements m'ont accompagné tout au long de ces années.

Je suis tout d'abord redevable à mes parents à qui je dédie cette thèse, pour leur soutien moral et la confiance qu'ils m'ont toujours accordé. Leur présence et leurs encouragements sont pour moi les piliers fondateurs de ce que je suis et ce que je fais. À mon père, que j'admire tant et qui m'a transmis (je l'espère !) l'humilité, ainsi que le goût des mathématiques : je continue donc à suivre tes pas ! À ma mère, dont je suis tellement fier, qui m'a transmis sensibilité et douceur : si j'ai su garder l'équilibre, c'est grâce à toi. Merci de te soucier toujours autant de mon bonheur.

À Alex : merci pour ton amour, ta patience et tes encouragements. Pour ton

soutien indéfectible, je te serai toujours infiniment reconnaissant. Cette thèse t'est aussi dédiée, car je n'y serais jamais parvenu sans toi. Je suis fier de toi, de nos projets, et je suis certain que l'avenir nous réserve de très belles surprises.

À mes sœurs qui me manquent et dont je suis si fier. Lola, profite de ta vie rêvée dans le *far west* canadien que je découvre un peu plus à chaque visite. Aurélie, on restera toujours connectés. À ton tour de connaître les joies de la thèse !

J'ai ensuite une pensée particulière pour mon grand-père dont l'optimisme et la joie de vivre sont sans égales, et pour ma grand-mère qui serait, je n'en doute pas, très fière de ce que j'ai accompli. Puis, ayant la chance d'avoir reçu un soutien important du reste de ma famille, je ne peux pas omettre de mentionner mes cousins, cousines, oncles et tantes, que je remercie chaleureusement ; avec une pensée spéciale pour mon oncle et parrain Thierry qui m'a donné le goût d'entreprendre et m'a appris très tôt la rigueur et l'exigence du travail bien fait.

Merci aussi à la famille d'Alex : Christelle, Eric, Cucu, Vincent, Toto et Nina. J'ai été très sensible à votre générosité et à votre souci de mon bien être. Votre présence a été chaleureuse et bienveillante durant toutes ces années.

Mes derniers remerciements, et non des moindres, vont bien sûr à mes amis. Tout d'abord, je me dois de donner quelques précisions sur le contexte particulier dans lequel s'est déroulé cette thèse. Plus d'un an avant le début de celle-ci, je me lançais dans une aventure ayant pour objectif de révolutionner le système d'approvisionnement en produits alimentaires frais des restaurants – jugé archaïque, non optimisé et par conséquent à grosse empreinte écologique. C'était pour moi l'occasion de satisfaire la soif d'entreprendre un projet impactant mêlant technologie, opérationnel et écologie ; soif qui m'habite toujours. Nous ne pouvions alors imaginer sereinement que 4 ans plus tard, nous aurions créé de nos mains une société structurée, avec plus de 20 personnes se donnant à 100% pour elle chaque jour. Nous pouvons être fiers de ce que nous avons accompli en si peu de temps, et la vitesse de développement ne cesse de croître, avec une multitude de nouveaux projets excitants à mener de front, notamment l'introduction de l'IA dans de nombreux pans de l'entreprise, ou autour des circuits courts.

Un immense merci donc, à mes deux associés Pierre et Ross, à qui je suis tellement reconnaissant de la confiance absolue qu'ils m'ont accordée au cours de ces dernières années. Pierre : on se connaît par cœur et grâce à toi je n'ai jamais eu peur de foncer, j'y ai toujours cru. Ross : tu m'as fait grandir personnellement et professionnellement. Merci pour ça et pour ta vision du « jeu » qui nous apporte une meilleure stabilité chaque jour.

Un grand merci à toute l'équipe Califrais dont je suis si fier. Continuez à m'époustoufler par votre implication qui nous fera aller loin !

François, mon maître musical : on n'est toujours pas prêt de se quitter, de nombreux projets nous attendent avec Sophie et toi. Pierre D., mon coloc de rêve, on devrait enfin avoir plus de temps pour faire de la musique ! Merci à Ben pour toutes ces parties de tennis endiablées qui ont tellement contribué à mon bien-être. Un petit mot aussi pour Sophie J. : même si on s'est moins vu ces derniers temps, notre lien est indélébile. Antoine R. : voilà 4 années (déjà !) depuis notre rencontre sur les bancs de l'ENS, et tu me fais toujours autant rire. C'était super pour moi d'avancer sur ma thèse en même temps que toi sur la tienne. Merci pour tes relectures ! Merci aussi évidemment à Samy qui m'a toujours motivé, et qui s'assure un beau parcours avec un début de thèse remarquable à Princeton.

Un grand merci à Lucie, Alice, Lolo, Jojo, Louis : on a un groupe tellement génial ! J'ai hâte de continuer à multiplier avec vous les vacances aux quatre coins du monde.

Enfin, ces trois années m'ont demandé beaucoup de travail et d'investissement dans les projets que j'ai entrepris. J'ai toujours eu pour habitude de mettre beaucoup de cœur et d'énergie dans mes passions, mon travail en faisant partie. C'est une période intense qui s'achève, et une nouvelle débute dans laquelle je souhaite prendre davantage de hauteur dans mon travail de recherche. Beaucoup de sujets m'intéressent dans lesquels j'aimerais me perfectionner. J'ai encore, fort heureusement, énormément à apprendre.

Avant-propos

La statistique vient de connaître une transformation importante lors des deux dernières décennies avec le développement de nouvelles méthodes d'inférence en grande dimension. Cette évolution récente découle de la nécessité de traiter les masses de données qui affluent en grande quantité dans de très nombreux domaines comme le web-marketing, la finance ou la santé pour n'en citer que quelques uns. Dans le domaine de la santé en particulier, qui sera le domaine d'application principal des méthodes développées dans cette thèse, des avancées importantes ont déjà eu lieu (en bioinformatique notamment) et d'autres sont attendues dans les années à venir grâce au volume de données de plus en plus grand (les fameuses *Big Data*), aux nouvelles techniques pour traiter ces données et à la puissance de calcul des ordinateurs. En effet, les techniques récentes (et à venir !) de machine learning pourraient bien bouleverser la médecine telle qu'on l'entend aujourd'hui, en proposant des traitements et des stratégies de prévention personnalisés, en travaillant dans l'anticipation plus que dans la curation, tout en réduisant la facture globale.

Dans de nombreuses applications médicales, on dispose d'un grand nombre de variables observées souvent plus grand que le nombre de patients dans l'échantillon : c'est dans ce contexte qu'on parle de grande dimension. Bien évidemment, les variables considérées dans un tel cadre – étant très nombreuses – ne sont pas toutes pertinentes : c'est la notion de parcimonie (sparsité/sparsity). Identifier les variables pertinentes est un enjeu fondamental pour interpréter sur le plan médical les données de grande dimension et tenter de comprendre les processus sous-jacents afin de prendre des décisions. Certaines méthodes récentes, comme le *deep learning*, peuvent se révèler très performantes en prédiction pour certaines tâches, mais disposent d'une interprétabilité limitée (à ce jour), ce qui ne permet pas toujours de comprendre en profondeur les outils utilisés et les raisons de leur succès.

Dans cette thèse, nous introduisons différentes méthodes d'apprentissage en grande dimension disposant d'un fort pouvoir d'interprétabilité. Pour chaque méthode, nous étudions leurs performances théoriques en établissant par exemple des inégalités oracles non-asymptotiques, ainsi que leurs performances pratiques sur don-

nées synthètiques et réelles avec des applications principalement en santé, tout en comparant les résultats avec l'état de l'art dans chacun des problèmes considérés.

Cette thèse a été financée par un contrat doctoral à l'Université Pierre et Marie Curie, du 1er octobre 2015 au 30 septembre 2018, et réalisée au Laboratoire de Statistique Théorique et Appliquée (LSTA), qui est devenu le Laboratoire de Probabilités, Statistique et Modélisation (LPSM) après sa fusion avec le Laboratoire de Probabilités et Modèles Aléatoires (LPMA) le 1er janvier 2018.

Organisation de la thèse

Le manuscrit comporte six chapitres, pouvant être lus indépendamment les uns des autres, dont voici une brève description.

— Le Chapitre 1 introduit les concepts fondamentaux et le formalisme utilisés dans la suite du manuscrit, à savoir des notions d'apprentissage supervisé et d'analyse de survie en grande dimension. Puis, les principaux résultats des chapitres qui suivent sont présentés et synthétisés.

— Le Chapitre 2 considère une étude de cas clinique avec une cohorte de malades atteints de drépanocytose, la maladie génétique la plus fréquente dans le monde résultant d'une mutation sur le gène codant pour l'hémoglobine. L'idée est de proposer des outils de visualisation de données longitudinales, en ne considérant que des hospitalisations sans complication particulière, de façon à décrire l'évolution "normale" des biomarqueurs et des paramètres vitaux de ces patients lors de leurs séjours à l'hôpital.

— Le Chapitre 3 est une étude de cas de grande dimension centrée sur la prédiction de la réhospitalisation précoce des patients de la cohorte étudiée au chapitre précédent (mais cette fois avec tous les patients, pas seulement ceux avec une crise vaso-occlusive non compliquée). De nombreux modèles sont considérés et comparés. Les questions abordées portent d'une part sur le choix du modèle relativement à ses performances prédictives et sa capacité à sélectionner les covariables explicatives, et d'autre part sur la comparaison du cadre de prédiction binaire (basé sur le choix arbitraire d'un délai) avec le cadre d'analyse de survie. Un modèle en particulier obtient d'excellentes performances sur ces données : le *C-mix*, introduit au chapitre suivant.

— Dans le Chapitre 4, nous proposons un modèle de mélange de durées en grande dimension, le *C-mix*, qui apprend à ordonner des patients suivant leur risque qu'un événement d'intérêt se produise rapidement, tout en déterminant des sous-groupes de pronostics différents au sein de la population. Nous introdui-

sons un algorithme efficace pour résoudre le problème convexe sous-jacent et nous illustrons notre approche sur des données simulées et des données génétiques en cancérologie.

— Le Chapitre 5 se place dans un contexte d'apprentissage supervisé en grande dimension et propose de combiner l'encodage "one-hot" de covariables continues avec l'utilisation d'une nouvelle pénalité appelée *binarsity* qui impose une régularisation par variation totale ainsi qu'une contrainte linéaire dans chaque groupe de variables binaires généré par l'encodage. Une inégalité oracle non-asymptotique en prédiction est proposée et la méthode est évaluée en pratique sur différents jeux de données.

— Le Chapitre 6 introduit une méthode pronostique appelée *binacox* qui traite du problème de la détection de multiples seuils par covariable continue dans un cadre multivarié d'analyse de survie en grande dimension. Celle-ci est basée sur le modèle de Cox et combine l'encodage "one-hot" avec la pénalité binarsity. Une inégalité oracle non-asymptotique est établie et les performances de la méthode sont examinées sur des données synthétiques et des données génétique en cancérologie.

Chaque chapitre est précédé d'un résumé permettant de le situer dans son contexte et de rendre compte des principaux points abordés. Précisons aussi que toutes les méthodes abordées aux cours de ces chapitres sont évaluées à la fois sur données simulées et sur données réelles, puis comparées à l'état de l'art. L'ensemble des codes informatiques (essentiellement développés en `Python`) utilisés pour réaliser ces études, implémenter les modèles proposés, et générer les figures de ce manuscrit sont disponibles en accès open source. Une section "Software" achève alors chaque chapitre et pointe vers un dépôt GitHub mettant à disposition le code relatif au chapitre, ainsi que des turoriels pour apprendre à utiliser les méthodes introduites.

Collaborations, publications et communications orales

Collaborations

Certains travaux présentés dans cette thèse ont été réalisés dans le cadre de collaborations (autres que celles avec mes directeurs de thèse) :

- Une première collaboration avec Mokhtar Z. Alaya a eu pour objectif de développer la méthode basée sur la pénalité binarsity dont les résultats obtenus sont présentés dans le Chapitre 5. J'ai rejoint le projet (qui avait déjà commencé) au milieu de ma première année de thèse, Mokhtar étant alors attaché temporaire d'enseignement et de recherche au LSTA. L'extension de l'utilisation de la pénalité binarsity dans le problème de détection de multiples seuils dans les covariables continues, dans un modèle de Cox en grande dimension, est un travail mené de nouveau en collaboration avec Mokhtar lors de ma dernière année de thèse, ce dernier étant alors chercheur postdoctoral au laboratoire Modal'X de l'université Paris Nanterre. Les résultats sont présentés dans le Chapitre 6.
- Une seconde collaboration avec Raphaël Veil, interne de santé publique, a permis de travailler en profondeur sur une étude de cas et les données complexes présentées au Chapitre 2, menant par la suite à l'étude comparative de modèles de prédiction de la réhospitalisation précoce présentée au Chapitre 3.
- Une troisième collaboration à débuté à la fin de ma thèse avec Antoine Barbieri, alors chercheur postdoctoral au centre de recherche des Cordeliers (INSERM), sur un projet combinant le modèle C-mix, présenté au Chapitre 4, joint avec une modélisation linéaire mixte afin de faire face à des données longitudinales multivariées. Une approche bayésienne est adoptée et quelques détails supplémentaires sur le papier en cours sont donnés dans la Section "Directions of future research", à la fin du manuscrit.

Publications

Articles parus ou acceptés pour publication dans des journaux internationaux :

- **S. Bussy**, A. Guilloux, S. Gaïffas, A.S. Jannot
  C-mix : a high dimensional mixture model for censored durations, with applications to genetic data
  Statistical Methods in Medical Research, 2017.

- M.Z. Alaya, **S. Bussy**, S. Gaïffas, A. Guilloux
  Binarsity : a penalization for one-hot encoded features
  Accepté pour publication et en révision mineure dans *Journal of Machine Learning Research*, 2018.

Articles soumis dans un journal international :

- **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
  Soumis à *BMC Medical Research Metodology*, 2018.

- R. Veil, **S. Bussy**, J.B. Arlet, A.S. Jannot et B. Ranque
  Trajectories of biological values and vital parameters : a retrospective cohort study on non-complicated vaso-occlusive crises
  Soumis à *Haematologica*, 2018.

Articles en cours de rédaction :

- **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
  Journal visé : *Journal of the Royal Statistical Society : Series B.*

- A. Barbieri, **S. Bussy**, S. Zohar et A.S. Jannot
  Mixture of joint modeling for multivariate longitudinal and survival data, with applications to metastatic colorectal cancer data
  Journal visé : *Statistics in Medicine.*

- P. Tran, A.L. Feral-pierssens, **S. Bussy**, L. Amar , G. Bobrie, G. Chatellier, A. Burgun, M. Azizi, A.S. Jannot
  Differential correlation between socio-economic status and clinical

characteristics in among men and women in patients consulting in a tertiary hypertension unit.
Journal visé : *Journal of Hypertension*.

Communications orales

- ♣ Machine Learning Summer School, Buenos Aires – Argentine, Juin 2018
  Binacox : automatic cut-points detection in high-dimensional Cox model (poster)
- ♣ I.A. & Santé, Nancy – France, Juillet 2018
  Design d'un algorithme d'IA en grande dimension pour prédire la réadmission à l'hôpital (oral)
- ♣ Séminaire des doctorants du LPSM, Paris – France, Avril 2017
  Survival model in high-dimension
- ♣ International Society for Clinical Biostatistics, Birmingham – United Kingdom, Août 2016
  C-mix : A high-dimensional mixture model for censored durations (poster)
- ♣ Séminaire du MAP5, Paris – France, Mai 2016
  Modèle pénalisé pronostique à variable latente pour des données censurées
- ♣ Paris Big Data Management Summit, Paris – France, Mars 2016
  A high dimensional mixture model for time-to-event data (poster)
- ♣ Workshop Data Initiative, École Polytechnique, Palaiseau – France, Mars 2016
  Modelling Patient Time-Series Data from EHR using Gaussian Processes
- ♣ Séminaire du département de médecine interne de l'Hôpital Européen George Pompidou (HEGP), APHP, Paris – France, Janvier 2016
  Facteurs prédictifs de la réhospitalisation précoce de drépanocytaires adultes
- ♣ Séminaire des doctorants du LPSM, Paris – France, Janvier 2016
  Risk assessment of sickle-cell anemia
- ♣ Séminaire des doctorants Centre de Recherche des Cordeliers INSERM, Paris – France, Décembre 2015
  Continuous time survival in latent variable models

♣ Workshop Data Initiative, École Polytechnique, Palaiseau – France, Octobre 2015
Forecasting Non-Stationary Time Series

Enseignement

✻ **Modèles linéaires**, Master 1 ISUP–UPMC
Chargé de TD 2017-18 (enseignante : Charlotte Dion), 2016-17 (enseignante : Claire Boyer)

✻ **Inférence statistique**, License 3 ISUP–UPMC
Chargé de TD 2015-16 ; 2016-17 ; 2017-18 (enseignant : Olivier Lopez)

✻ **Séries temporelles**, Master 1 ISUP-UPMC
Chargé de TD 2015-16 (enseignant : Vincent Lefieux)

Co-encadrement

✤ Maud De Tollenaere, co-encadrée avec A.S. Jannot, F. Pages et A. Guilloux
Stage de recherche (2018) du Master 2 Analyse, Modélisation et Ingénierie de l'Information Biologique et Médicale, Université Paris-Saclay
Techniques d'apprentissage automatique pour affiner le pronostic à l'aide de données d'imagerie immunologique

✤ Christophe Botella, co-encadré avec A.S. Jannot et A. Guilloux
Stage de recherche (2016) du Master 2 Mathématiques pour les Sciences du Vivant, Université Paris-Saclay
Modèle de régression sur données longitudinales : application à des données médicales à faible résolution

Table des matières

Chapitre 1

Introduction

Sommaire

Dans ce chapitre, nous introduisons les notions, les outils et le formalisme général implicitement utilisés dans le reste du manuscrit. Ensuite, le cheminement suivi lors de cette thèse est décrit, ainsi que les différents résultats obtenus dans les chapitres qui suivent.

1.1 Apprentissage supervisé en grande dimension

L'objectif général de l'apprentissage statistique (machine learning) est d'élaborer des procédures automatiques qui permettent de mettre en évidence des règles générales à partir d'exemples (données).

1.1.1 Généralités

En apprentissage supervisé, nous disposons (dans les cas simples) d'une base de données, également appelée échantillon d'apprentissage,

$$\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

où pour tout $i \in \{1, \dots, n\}$, les $x_i \in \mathcal{X} \subset \mathbb{R}^d$ sont les variables d'entrée (aussi appelées covariables ou *features*) et les $y_i \in \mathcal{Y} \subset \mathbb{R}$ sont les variables de sortie (ou labels) correspondantes (dans le cas de la classification, $\mathcal{Y}$ est un ensemble fini). L'échantillon $\mathcal{D}_n$ est ainsi composé de n réalisations de couples de variables aléatoires $(X_1, Y_1), \dots, (X_n, Y_n)$ que nous supposons indépendants et identiquement distribués suivant une même loi $\mathbb{P}$ inconnue.

L'objectif est de construire une fonction de prédiction à partir de $\mathcal{D}_n$, pour être en mesure de prévoir la valeur du label Y correspondant à une nouvelle réalisation d'un $X = x \in \mathcal{X}$ donné, où (X, Y) suit également la loi $\mathbb{P}$ et est indépendant des n couples qui constituent $\mathcal{D}_n$.

Une fonction de prédiction est une fonction mesurable de $\mathcal{X}$ dans $\mathcal{Y}$. En notant

$$\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$$

une fonction de perte, $\ell(y_1, y_2)$ représente la perte encourue lorsque la vraie sortie est y_1 et la sortie prédite est y_2. Par exemple,

$$\ell(y_1, y_2) = \mathbb{1}_{\{y_1 \neq y_2\}}$$

est une perte classique en classification et

$$\ell(y_1, y_2) = |y_1 - y_2|^p$$

avec $p \geq 1$ est une perte classique en régression (on parle alors de régression ℓ_p [1] et lorsque $p = 2$, ℓ est appelée perte quadratique et la tâche d'apprentissage est aussi appelée régression aux moindres carrés).

La qualité d'une fonction de prédiction $g : \mathcal{X} \to \mathcal{Y}$ est mesurée par son risque intégré [Vapnik, 1991], aussi appelé erreur de généralisation, défini par

$$R_{\mathbb{P}}(g) = \mathbb{E}_{\mathbb{P}}[\ell(Y, g(X))].$$

La "meilleure" fonction de prédiction est une fonction de l'ensemble $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ des fonctions définies sur $\mathcal{X}$ et à valeur dans $\mathcal{Y}$ minimisant le risque $R_{\mathbb{P}}$, soit

$$g^\star_{\mathbb{P}} \in \operatorname{argmin}_{g \in \mathcal{F}(\mathcal{X},\mathcal{Y})} R_{\mathbb{P}}(g).$$

Une telle fonction n'existe pas nécessairement mais existe pour les fonctions de pertes usuelles, notamment celles que nous considérerons par la suite [Vapnik, 1998]. On l'appelle fonction oracle et c'est la fonction que l'on cherche à approcher. Elle dépend de $\mathbb{P}$ et, par conséquent, est elle aussi inconnue. Nous précisons explicitement cette dépendance en $\mathbb{P}$ dans ces quelques lignes introductives, par exemple dans la notation $\mathbb{E}_{\mathbb{P}}$, ce que nous ne ferons plus dans la suite du manuscrit afin d'alléger l'écriture.

1.1.2 Minimisation du risque empirique pénalisé

Le risque inconnu $R_{\mathbb{P}}(g)$ d'une fonction de prédiction g peut être estimé par son équivalent empirique

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell\Big(Y_i, g(X_i)\Big),$$

et en supposant $\mathbb{E}_{\mathbb{P}}\Big[\ell\Big(Y, g(X)\Big)^2\Big] < +\infty$, il vient par la loi forte des grands nombres et le théorème central limite que

$$R_n(g) \xrightarrow[n\to+\infty]{p.s.} R_{\mathbb{P}}(g)$$

et

$$n^{1/2}\Big(R_n(g) - R_{\mathbb{P}}(g)\Big) \xrightarrow[n\to+\infty]{d.} \mathcal{N}\Big(0, \operatorname{Var}_{\mathbb{P}}[\ell(Y, g(X))]\Big).$$

Donc pour toute fonction de prédiction g, $R_n(g)$ effectue des déviations autour de sa moyenne $R_{\mathbb{P}}(g)$ de l'ordre de $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$. Il est alors naturel de considérer l'algorithme d'apprentissage de minimisation du risque empirique [Vapnik, 1998] défini par

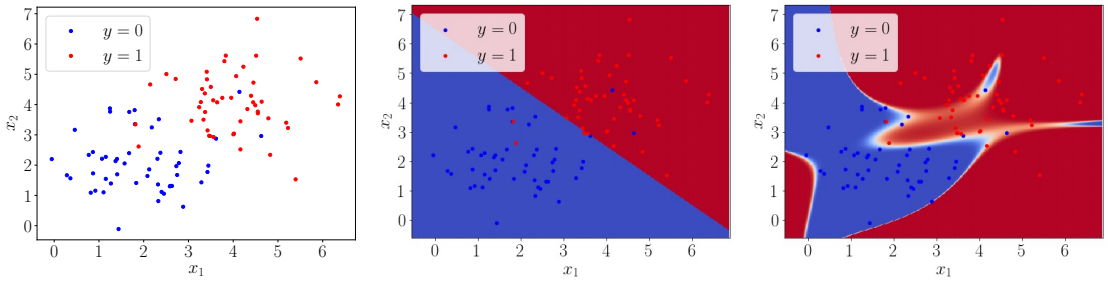
$$g_{n,\mathcal{G}} \in \operatorname{argmin}_{g \in \mathcal{G}} R_n(g), \tag{1.1}$$

1. La notation ℓ_p ici utilisée n'a rien à voir avec la notation ℓ de la fonction de perte.

avec $\mathcal{G} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$.

Choisir $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$ entraîne en général un sur-apprentissage : le risque empirique est alors nettement inférieur à $R_{\mathbb{P}}$, même lorsque $n \to +\infty$. En effet, une infinité de fonctions de prédiction peuvent minimiser le risque empirique en "apprenant par coeur" $\mathcal{D}_n$ (ce qui pose déjà un problème de choix de la fonction de prédiction !), tout en ayant de très mauvaises propriétés de généralisation lorsqu'il s'agit de prédire un nouvel exemple distinct de ceux de $\mathcal{D}_n$ utilisés en phase d'apprentissage.

L'idée est alors de choisir un $\mathcal{G}$ suffisamment complexe pour être en mesure d'approcher la fonction $g^\star_{\mathbb{P}}$, mais pas trop complexe pour éviter le sur-apprentissage. La Figure 1.1 illustre la notion de sur-apprentissage [2].



(a) Données simulées (b) Oracle $g^\star_{\mathbb{P}}$ (c) Fonction de prédiction

FIGURE 1.1 Des données de classification binaire sont générées dans $\mathbb{R}^2$ avec $n = 100$, les points bleus suivant $\mathcal{N}\Big((2,2)^\top, \Sigma\Big)$ et les rouges suivant $\mathcal{N}\Big((4,4)^\top, \Sigma\Big)$, avec $\Sigma = \mathrm{diag}(1,1)$. La perte logistique $\ell(y_1, y_2) = \log\Big(1 + \exp(-y_1 y_2)\Big)$ est utilisée et les covariables arbitrairement choisies sont des produits de puissances de x_1 et x_2, par exemple $x_1^4 x_2^2$ ou $x_1^3 x_2^3$, et on considère les fonctions de prédiction linéaires en ces covariables. De cette façon, la fonction de prédiction obtenue est suffisamment complexe pour apprendre "par coeur" les données d'apprentissage simulées, comme on l'observe sur la figure (c) où la fonction de prédiction représentée dans l'espace $\mathbb{R}^2$ initial est "loin" de l'oracle représentée sur la figure (b), qui est ici linéaire (du fait de la structure de Σ). L'erreur d'apprentissage est alors très faible, contrairement à l'erreur de généralisation.

Une approche pratique basée sur cette idée consiste à pénaliser (ou régulariser) le risque empirique à minimiser [Tikhonov and Arsenin, 1977, Breiman, 1995, Tibshirani, 1996]. Par exemple, on peut d'une part restreindre $\mathcal{G}$ à l'ensemble $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ des fonctions linéaires de $\mathcal{X}$ dans $\mathcal{Y}$ en ne considérant que des fonctions de la forme

$$g_\beta : x \mapsto x^\top \beta$$

2. Les figures de l'introduction sont générées dans un notebook `Python` disponible à l'adresse https://github.com/SimonBussy/thesis-illustrations.

avec $\beta \in \mathbb{R}^d$, et d'autre part ajouter une pénalité à R_n pour restreindre davantage $\mathcal{G}$, ce qui donne en prenant la perte quadratique

$$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 + \operatorname{pen}(\beta), \tag{1.2}$$

avec pen : $\mathbb{R}^d \to \mathbb{R}^+$ la fonction pénalité, qu'on appelle aussi fonction de régularisation. La fonction de prédiction résultant de la minimisation du risque empirique pénalisé (1.2) est alors ici $g_{\hat{\beta}}$. Il existe de nombreuses pénalités usuelles donnant des estimateurs avec différentes propriétés d'interprétabilité [Bickel et al., 2006]. Nous en présentons une partie dans la suite parmi les plus utilisées.

Quelques pénalités classiques. Restons dans le cadre de la tâche d'apprentissage proposée dans le problème (1.2). La famille de régularisation majoritairement utilisée est alors celle des normes (ou pseudo-normes) ℓ_p, à savoir

$$\operatorname{pen}(\beta) = \lambda \|\beta\|_p = \lambda \Big(\sum_{i=1}^{d} |\beta_j|^p \Big)^{1/p}$$

où $0 \leq p \leq +\infty$ et $\lambda \in \mathbb{R}^+$ est un hyper-paramètre de régularisation qui contrôle le compromis entre l'adéquation du modèle aux données et sa complexité. Pour $p \in [0, 1[$, ces fonctions de régularisation permettent de promouvoir la parcimonie du vecteur de régression β, comme le suggère la Figure 1.2. Mais elles ne sont ni différentiables, ni convexes et le problème de minimisation (1.2) peut alors se révéler difficile à résoudre en pratique. Une façon usuelle de réduire la difficulté du problème est d'utiliser la norme ℓ_1 en tant qu'approximation convexe des pénalisations, comme nous allons le voir dans la suite.

Procédures ℓ_0. La pénalité ℓ_0 est telle que

$$\operatorname{pen}(\beta) = \lambda \|\beta\|_0.$$

Elle a une interprétabilité très intuitive. En effet,

$$\|\beta\|_0 = \sum_{j=1}^{d} \mathbb{1}_{\{\beta_j \neq 0\}}$$

compte le nombre de coefficients non nuls dans le vecteur de régression [Schwarz et al., 1978]. L'objectif est donc de forcer l'estimateur obtenu dans (1.2) à avoir peu de coordonnées non nulles correspondant aux variables influentes dans la tâche prédictive. Deux pénalités classiques font intervenir la pseudo-norme ℓ_0 : le critère AIC [Akaike, 1992] où

$$\operatorname{pen}(\beta) = \frac{2}{n} \|\beta\|_0$$

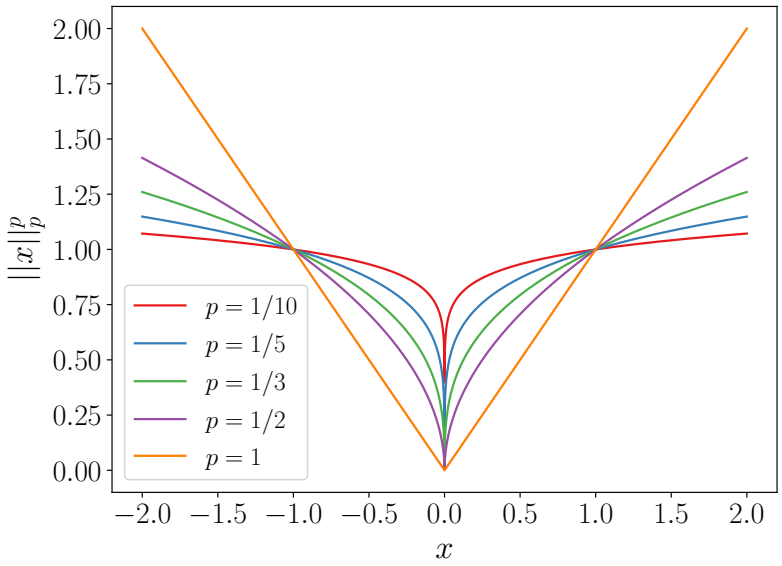


FIGURE 1.2 Illustration de l'effet de la régularisation ℓ_p avec les graphes de $x \mapsto \|x\|_p^p$ pour $x \in \mathbb{R}$ et $p \in \{1/10, 1/5, 1/3, 1/2, 1\}$. Plus p tend vers 0, moins les coefficients proches de 0 seront pénalisés.

et le critère BIC [Schwarz et al., 1978] où

$$\text{pen}(\beta) = \frac{\log d}{n} \|\beta\|_0.$$

Lorsque la dimension d du problème augmente, les procédures ℓ_0 posent des problèmes de complexité algorithmique puisque le nombre de modèles à comparer est de l'ordre de $\mathcal{O}(2^d)$ et augmente exponentiellement avec d, ce qui les rend inutilisables en pratique. On dit que le problème (1.2) est NP-dur avec cette pénalité ; en particulier, il n'est pas résolvable en un temps polynomial [Tropp, 2004].

De plus, des instabilités apparaissent lorsque d grandit [Breiman, 1995], et dans un contexte de grande dimension où d est grand devant le nombre d'exemples n (on notera $d \gg n$), ces méthodes sont à proscrire. Pour autant, la nécessité d'identifier les variables significatives et influentes est d'autant plus cruciale en grande dimension. En fait, la plupart des algorithmes et estimateurs habituels deviennent instables dès que d dépasse $\sqrt{n}$.

Pour répondre au problème algorithmique des procédures ℓ_0 qui engendrent une fonction de pénalité non convexe, une approche consiste justement à convexifier la pénalité. C'est l'idée suivie pour construire la pénalité lasso.

Lasso ℓ_1. La pénalité lasso (pour least absolute shrinkage and selection operator) introduite dans Tibshirani [1996] est une procédure classique et fondamentale de l'estimation en grande dimension. Elle est telle que

$$\text{pen}(\beta) = \lambda \|\beta\|_1$$

avec

$$\|\beta\|_1 = \sum_{j=1}^{d} |\beta_j|.$$

Le problème d'optimisation (1.2) est cette fois convexe avec cette pénalité, on peut alors faire appel aux algorithmes d'optimisation convexe pour déterminer $\hat{\beta}$. Efron et al. [2004] proposent l'algorithme LARS (Least Angle Regression Stepwise), un algorithme de selection de modèle qui permet d'obtenir un chemin de régularisation pour une plage de valeurs de λ (voir aussi Friedman et al. [2007]).

Le problème est en fait équivalent au problème de minimisation (1.1) où on restreint $\mathcal{G}$ à l'ensemble $\mathcal{L}(\mathcal{X}, \mathcal{Y})$, tout en imposant la contrainte

$$\sum_{j=1}^{d} |\beta_j| \leq s$$

avec $s \in \mathbb{R}^+$ fixé [Hebiri, 2009]. Cela revient donc à contraindre les β dans la boule de norme ℓ_1 de rayon s dans $\mathbb{R}^d$. Comme illustré dans la Figure 1.3, l'estimateur $\hat{\beta}$ obtenu est alors *sparse* : un certain nombre de ses coordonnées sont nulles, lui conférant l'interprétabilité en sélection de variables souhaitée.

Notons $\hat{\beta}^{mc}$ l'estimateur des moindres carrés obtenu sans pénalité et précisons que le choix de l'hyper-paramètre de régularisation λ influe directement sur le nombre de composantes nulles de $\hat{\beta}$: $\lambda = 0$ impliquerait que $\hat{\beta} = \hat{\beta}^{mc}$ (pas de pénalité), alors qu'une grande valeur de λ donnerait le vecteur nul pour $\hat{\beta}$. Le choix de λ est donc crucial et il existe plusieurs procédures empiriques pour en déterminer une "bonne" valeur, où l'idée est toujours de s'assurer du pouvoir de généralisation de l'estimateur obtenu, à savoir que $R_n(g_{\hat{\beta}})$ ne dévie "pas trop" de $R_{\mathbb{P}}(g_{\hat{\beta}})$. La plus connue et utilisée s'appelle la validation croisée [Tibshirani, 1996].

L'estimateur lasso a beaucoup été étudié en régression linéaire [Efron et al., 2004, Donoho et al., 2006, Zhang et al., 2008, Meinshausen et al., 2009] et plus généralement dans le cas d'un modèle de régression additive non-paramétrique [Juditsky and Nemirovski, 2000, Bunea et al., 2007, 2006, Greenshtein et al., 2004, Bickel et al., 2009]. Par exemple, sous certaines hypothèses, sa consistance est montrée dans Knight and Fu [2000] et sa signe-consistance dans Zhao and Yu [2006]. Ainsi, une littérature riche a vu le jour depuis une vingtaine d'année en s'intéressant à ses différentes propriétés théoriques dans différents cadres d'application (voir également Meinshausen et al. [2006], Candès et al. [2007], Wainwright [2009], van de Geer

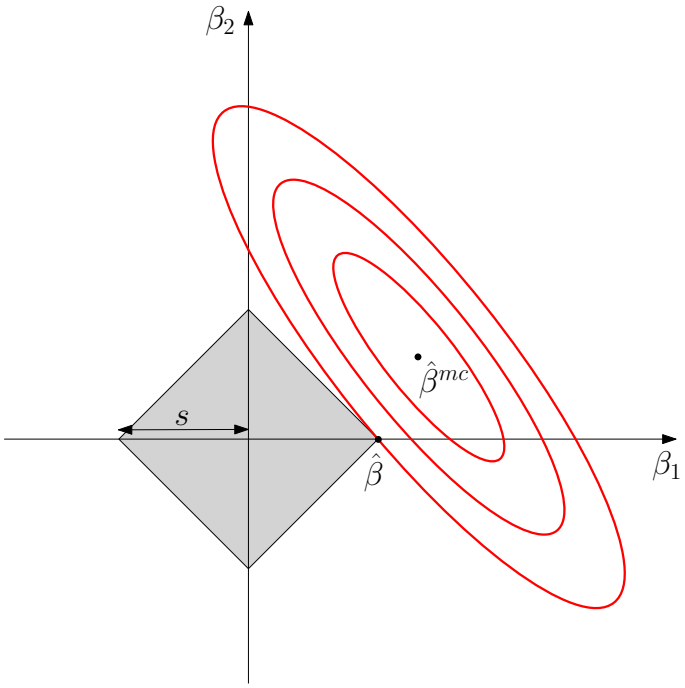


FIGURE 1.3 Illustration de l'effet de la pénalité lasso avec $d = 2$. $\hat{\beta}^{mc}$ représente ici l'estimateur des moindres carrés obtenu sans pénalité. Les ellipses rouges représentent des courbes de niveau de la fonction $\beta \mapsto R_n(g_\beta)$, avec donc la perte quadratique. Le carré grisé représente la région admissible des estimateurs lasso, soit ici $\{\beta \in \mathbb{R}^2, \|\beta\|_1 \leq s\}$. $\hat{\beta}$ représente alors l'estimateur lasso obtenu dans cet exemple, et rend compte de la sparsité induite par la pénalité puisque la seconde composante de l'estimateur est nulle.

and Bühlmann [2009], Bühlmann and van de Geer [2011], Tibshirani et al. [2013]). Par ailleurs, de nombreuses autres pénalités sont dérivées du lasso, dont quelques unes sont mentionnées ci-après.

Adaptive lasso. La pénalité adaptive lasso introduite dans Zou [2006] est une pénalité lasso pondérée telle que

$$\text{pen}(\beta) = \lambda_n \sum_{j=1}^{d} \hat{\omega}_j |\beta_j|.$$

Elle permet d'établir des propriétés oracles de l'estimateur obtenu. En prenant par exemple

$$\hat{\omega}_j = \frac{1}{|\hat{\beta}_j^{mc}|^\gamma}$$

(les $\hat{\beta}_j^{mc}$ étant $p.s.$ non nuls pour tout $j \in \{1, \dots, d\}$) et en supposant $n^{-1/2}\lambda_n \to 0$ et $n^{(\gamma-1)/2}\lambda_n \to +\infty$, alors l'estimateur adaptive lasso est consistant en sélection de variables et asymptotiquement normal [Zou, 2006].

Group lasso. La pénalité group lasso introduite dans Yuan and Lin [2006] est telle que

$$\text{pen}(\beta) = \lambda \sum_{k=1}^{K} \sqrt{p_k} \|\beta_{G_k}\|_2$$

avec

$$\|\beta_{G_k}\|_2 = \sqrt{\sum_{j \in G_k} \beta_j^2},$$

où $(G_k)_{k=1,\dots,K}$ forme une partition de $\{1, \dots, d\}$ telle que $|G_k| = p_k$. Cette pénalité favorise une sparsité par groupe G_k de covariables et généralise le lasso à ce cadre (on se ramène au lasso si $p_k = 1$ pour tout $k \in \{1, \dots, K\}$). La consistance de cet estimateur en sélection de variables est montrée dans Bach [2008], des inégalités oracles sont établies dans Nardi et al. [2008], et des extensions sont proposées par exemple dans Meier et al. [2008] où le cadre de la régression logistique est considéré (nous l'utilisons en Section 5.4 du Chapitre 5), dans Simon et al. [2013] pour le sparse group lasso, ou encore dans Alaíz et al. [2013] pour le group fused lasso.

Elastic net. La pénalité elastic net introduite dans Zou and Hastie [2005] est de la forme

$$\text{pen}(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2.$$

Elle combine la pénalité lasso avec la pénalité ridge [Hoerl and Kennard, 1970] et bénéficie ainsi des atouts des deux méthodes : elle dispose des propriétés de sélection de variables du lasso et pallie le défaut de l'estimation lasso lorsque des covariables sont fortement corrélées grâce à la partie ridge. Il y a ici deux hyper-paramètres de régularisation $(\lambda_1, \lambda_2) \in (\mathbb{R}^+)^2$.

Cette régularization est donc très utile en pratique, notamment dans des problèmes de grande dimension où le grand nombre de covariables considérées induit souvent des problèmes de collinéarité entre celles-ci. Nous en ferons particulièrement usage dans les Chapitres 3 et 4. Zou and Zhang [2009] proposent l'adaptive elastic net pour obtenir des inégalités oracles à l'aide des propriétés de l'adaptive lasso.

Fused lasso. La pénalité fused lasso introduite dans Tibshirani et al. [2005] est de la forme

$$\text{pen}(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_{\text{TV}},$$

où

$$\|\beta\|_{\text{TV}} = \sum_{j=2}^{d} |\beta_j - \beta_{j-1}|$$

est la régularisation par variation totale qui pénalise les différences successives des coefficients de β et favorise donc ce dernier à être constant par morceau [Little and Jones, 2011, Rudin et al., 1992]. Les propriétés de l'estimateur fused lasso ont par exemple été étudiées dans Rinaldo et al. [2009] et des algorithmes efficaces ont été proposés pour le calculer dans Friedman et al. [2007], Liu et al. [2010], Hoefling [2010]. Le fused lasso a notamment été utilisé avec succès pour répondre au problème de la détection de points de rupture dans des signaux [Harchaoui and Lévy-Leduc, 2010, Bleakley and Vert, 2011], ce qui sera d'un intérêt particulier dans les Chapitres 5 et 6.

Insistons sur le fait que nous n'avons présenté ici qu'une partie non exhaustive des types de régularisations proposées dans la littérature, en insistant sur les penalités liées au lasso qui nous intéressent particulièrement dans cette thèse, en rapport avec les chapitres qui suivent. Nous n'avons par exemple pas abordé les problèmes de sélection de modèle discutés dans Birgé and Massart [2001, 2007], dans un cadre d'estimation non-paramétrique. On s'intéresse alors, dans le paragraphe suivant, à la résolution algorithmique du problème de minimisation du risque empirique pénalisé.

Algorithme du gradient proximal. Rappelons que déterminer la fonction de prédiction résultant de la minimisation du risque empirique pénalisé requiert la résolution du problème d'optimisation suivant

$$\hat{\beta} \in \text{argmin}_{\beta \in \mathbb{R}^d} R_n(g_\beta) + \text{pen}(\beta), \tag{1.3}$$

où on se place ici de nouveau dans le cas classique $\mathcal{G} = \mathcal{L}(\mathcal{X}, \mathcal{Y})$. On suppose de plus que l'on choisit la fonction de perte ℓ de telle sorte que la fonction $\beta \mapsto R_n(g_\beta)$ soit convexe, différentiable et de gradient L-lipschitz, c'est-à-dire

$$\|\nabla R_n(g_u) - \nabla R_n(g_v)\|_2 \leq L\|u - v\|_2$$

pour tout $(u, v) \in (\mathbb{R}^d)^2$.

On peut ainsi appliquer le lemme de descente proposé dans Bertsekas [1995] (Prop.A.24), soit

$$R_n(g_u) \leq R_n(g_v) + (u - v)^\top \nabla R_n(g_v) + \frac{L}{2}\|u - v\|_2^2 \tag{1.4}$$

pour tout $(u, v) \in (\mathbb{R}^d)^2$.

Pour ce qui est de la fonction de pénalité (pen), on suppose seulement sa convexité. En effet, nombre des fonctions pénalités ne sont pas différentiables, comme c'est le cas pour l'ensemble des pénalités mentionnées précédemment. C'est d'ailleurs pour cette raison qu'on ne peut pas directement résoudre le problème (1.3) à l'aide d'un algorithme classique de descente de gradient. Une idée intuitive est alors de résoudre itérativement le problème suivant

$$\beta^{(k+1)} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \Big\{ R_n(g_{\beta^{(k)}}) + (\beta - \beta^{(k)})^\top \nabla R_n(g_{\beta^{(k)}}) + \frac{L}{2} \|\beta - \beta^{(k)}\|_2^2 + \operatorname{pen}(\beta) \Big\}, \tag{1.5}$$

puisque (1.4) nous garantit que

$$R_n(g_{\beta^{(k+1)}}) + \operatorname{pen}(\beta^{(k+1)}) \leq R_n(g_{\beta^{(k)}}) + \operatorname{pen}(\beta^{(k)}).$$

L'opérateur proximal [Moreau, 1965] associé à pen est alors défini par

$$\begin{aligned} \operatorname{prox}_{\text{pen}} : \mathbb{R}^d &\to \mathbb{R}^d \\ v &\mapsto \operatorname{argmin}_{u \in \mathbb{R}^d} \Big\{ \frac{1}{2} \|u - v\|_2^2 + \operatorname{pen}(u) \Big\}. \end{aligned}$$

L'opérateur proximal généralise la notion de projection : si on prend comme pénalité la fonction indicatrice d'un ensemble convexe C, c'est-à-dire

$$\operatorname{pen}(\beta) = \delta_C(\beta) = \begin{cases} 0 & \text{si } \beta \in C, \\ +\infty & \text{sinon,} \end{cases}$$

alors $\operatorname{prox}_{\delta_C}(\beta)$ est la projection euclidienne de β sur C. En remarquant que

$$(\beta - \beta^{(k)})^\top \nabla R_n(g_{\beta^{(k)}}) + \frac{L}{2} \|\beta - \beta^{(k)}\|_2^2 = \frac{L}{2} \Big[\|\beta - \beta^{(k)} + \frac{1}{L} \nabla R_n(g_{\beta^{(k)}})\|_2^2 - \|\frac{1}{L} \nabla R_n(g_{\beta^{(k)}})\|_2^2 \Big],$$

le problème (1.5) peut s'écrire sous la forme

$$\beta^{(k+1)} = \operatorname{prox}_{L^{-1}\text{pen}} \Big(\beta^{(k)} - \frac{1}{L} \nabla R_n(g_{\beta^{(k)}}) \Big), \tag{1.6}$$

et cette écriture constitue la base de l'algorithme du gradient proximal. L'algorithme ISTA (Iterative Shrinkage Thresholding Algorithm) consiste simplement à partir d'un $\beta^{(0)} \in \mathbb{R}^d$ initial et à itérer (1.6) jusqu'à la convergence qui est garantie à une vitesse $\mathcal{O}(1/k)$ [Daubechies et al., 2004], vitesse qui est accélérée à $\mathcal{O}(1/k^2)$ dans l'algorithme FISTA [Beck and Teboulle, 2009].

Pour de nombreuses pénalités, l'opérateur proximal a une expression explicite. La Table 1.1 donne par exemple les expressions des opérateurs proximaux pour les trois pénalités qui nous intéresseront particulièrement dans la suite. La pénalité TV

TABLE 1.1 Expression des opérateurs proximaux pour les quelques pénalités évoquées précédemment qui nous intéressent pour les chapitres suivants.

Pénalité	$\text{pen}(\beta)$	$\Big(\text{prox}_{\text{pen}}(\beta)\Big)_j$
lasso	$\lambda\|\beta\|_1$	$\max(0, 1-\lambda/\|\beta_j\|)\beta_j$
elastic net	$\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2$	$1/(1+\lambda_2)\Big(\text{prox}_{\lambda_1\|\cdot\|_1}(\beta)\Big)_j$
fused lasso	$\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_{\text{TV}}$	$\Big(\text{prox}_{\lambda_1\|\cdot\|_1}(\text{prox}_{\lambda_2\|\cdot\|_{\text{TV}}}(\beta))\Big)_j$

n'admet pas d'opérateur proximal explicite, mais ce dernier peut être calculé via un algorithme proposé dans Condat [2013] par exemple.

Après avoir introduit la notion générale de minimisation du risque empirique pénalisé, les différentes pénalités qui seront considérées dans la suite du manuscrit, ainsi que la résolution algorithmique du problème d'optimisation qui en résulte (permettant d'obtenir la fonction de prédiction en pratique), nous introduisons dans la section suivante la notion d'inégalité oracle, qui sera au cœur des Chapitres 5 et 6, permettant de quantifier la vitesse à laquelle le risque de la fonction de prédiction obtenue tend vers le risque de la "meilleure" fonction de prédiction possible ("meilleure" dans un certain sens défini ci-après).

1.1.3 Inégalités oracles

Notons $g^\star_{\mathbb{P},\mathcal{G}}$ le prédicteur oracle sur $\mathcal{G}$, c'est-à-dire la fonction de prédiction minimisant le risque sur $\mathcal{G}$, soit

$$g^\star_{\mathbb{P},\mathcal{G}} \in \text{argmin}_{g\in\mathcal{G}} R_{\mathbb{P}}(g),$$

en supposant le minimum atteint pour simplifier l'écriture. Notons aussi de façon générale $g^{\text{pen}}_{n,\mathcal{G}}$ la fonction de prédiction qui minimise sur $\mathcal{G}$ le risque empirique pénalisé. On a alors naturellement

$$R_{\mathbb{P}}(g^{\text{pen}}_{n,\mathcal{G}}) \geq R_{\mathbb{P}}(g^\star_{\mathbb{P},\mathcal{G}}) \geq R_{\mathbb{P}}(g^\star_{\mathbb{P}}),$$

et l'excès de risque intégré de $g^{\text{pen}}_{n,\mathcal{G}}$ se décompose de la façon suivante

$$R_{\mathbb{P}}(g^{\text{pen}}_{n,\mathcal{G}}) - R_{\mathbb{P}}(g^\star_{\mathbb{P}}) = \underbrace{R_{\mathbb{P}}(g^{\text{pen}}_{n,\mathcal{G}}) - R_{\mathbb{P}}(g^\star_{\mathbb{P},\mathcal{G}})}_{\varepsilon_1} + \underbrace{R_{\mathbb{P}}(g^\star_{\mathbb{P},\mathcal{G}}) - R_{\mathbb{P}}(g^\star_{\mathbb{P}})}_{\varepsilon_2},$$

où ε_1 est l'erreur d'estimation (ou erreur stochastique) et ε_2 est l'erreur d'approximation (ou biais, erreur systématique). En général, plus $\mathcal{G}$ est "grand", plus l'erreur

systématique est petite et plus l'erreur stochastique est grande : c'est le compromis "biais-variance".

Pour quantifier les performances de $g_{n,\mathcal{G}}^{\text{pen}}$, on peut alors chercher à établir des bornes sur l'excès de risque intégré, ou sur l'erreur stochastique (l'erreur systématique étant souvent intrinséquement liée aux hypothèses de modélisation [3]), ou encore sur leurs versions empiriques exprimées avec le risque empirique $R_n(\cdot)$.

On peut par example avoir affaire au type d'inégalité oracle suivant

$$\forall \varepsilon \in]0,1[, \qquad \mathbb{P}[R_{\mathbb{P}}(g_{n,\mathcal{G}}) - aR_{\mathbb{P}}(g_{\mathbb{P},\mathcal{G}}^{\star}) \leq \delta_n(\varepsilon)] \geq 1 - \varepsilon, \tag{1.7}$$

qui s'exprime en terme de risque intégré, avec $a \geq 1$ une constante numérique, et où tout l'enjeu est de choisir la fonction $\delta_n :]0,1[\to \mathbb{R}^+$ qui décroit le plus vite vers 0 possible, de telle sorte que le résultat (1.7) soit le plus précis possible. On parle alors d'inégalité oracle non-asymptotique pour (1.7) précise ("sharp") si $a = 1$ et à vitesse lente (resp. rapide) si $\delta_n(\varepsilon) = \mathcal{O}(\sqrt{\log \mathcal{C}(\mathcal{G})/n})$ (resp. $\mathcal{O}(\log \mathcal{C}(\mathcal{G})/n)$), où $\mathcal{C}(\mathcal{G})$ est une mesure de la complexité de $\mathcal{G}$, par exemple la dimension de Vapnik-Chervonenkis [Vapnik and Chervonenkis, 2015] ou encore la complexité de Rademacher [Mohri and Rostamizadeh, 2009].

Mais dans la suite du manuscrit, le type d'inégalité oracle qui nous intéresse s'exprime en terme de risque empirique, et nous donnons dans le paragraphe suivant un example de résultat avec l'estimateur lasso.

Le cas du lasso. Plaçons nous dans le cadre d'un modèle de régression additif

$$Y_i = g^{\star}(X_i) + \varepsilon_i \qquad \text{pour } i \in \{1, \dots, n\} \tag{1.8}$$

où $g^{\star} : \mathbb{R}^d \to \mathbb{R}$ est la fonction inconnue à estimer et

$$\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

sont des erreurs gaussiennes. On se ramène au modèle de régression linéaire si

$$g^{\star}(x) = x^{\top}\beta^{\star}$$

pour tout $x \in \mathbb{R}^d$. Prenons l'exemple de l'estimateur lasso qui nous intéresse particulièrement dans la suite, avec donc $\hat{\beta}$ issu de (1.2) et $\text{pen}(\beta) = \lambda\|\beta\|_1$. Afin de mesurer la qualité de l'estimateur en prédiction, on va chercher à obtenir des bornes de risque avec grande probabilité pour la norme empirique associée à la perte quadratique définie par

$$\|g^{\star} - g_{\hat{\beta}}\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(g^{\star}(X_i) - g_{\hat{\beta}}(X_i)\right)^2}, \tag{1.9}$$

3. On fait dépendre $\mathcal{G}$ de n, avec une complexité croissante en n, pour obtenir la consistance du minimiseur du risque empirique pénalisé et éviter que l'excès de risque de $g_{n,\mathcal{G}}$ ne soit minoré par l'erreur systématique, mais cela ne sera pas un sujet abordé dans ce manuscrit.

voir par exemple Bunea et al. [2007], Zhang et al. [2008], Bickel et al. [2009] ; ou à borner l'espérance de (1.9), voir par exemple Massart and Meynet [2011]. On peut aussi s'intéresser à la qualité de l'estimation de $\beta^\star$ dans le cas linéaire par exemple, auquel cas on cherche naturellement à borner la quantité $\|\beta^\star - \hat{\beta}\|_p$ [Bunea et al., 2007, Meinshausen et al., 2009, Zhang et al., 2008], ou encore la qualité en sélection (particulièrement en grande dimension) où l'objectif est de retrouver le support en norme ℓ_0 de $\beta^\star$ [Zhao and Yu, 2006], soit

$$\mathcal{A}(\beta^\star) = \{j, \beta_j^\star \neq 0\}.$$

Dans le cas du modèle général de régression additif (1.8) et en prenant

$$\lambda = \lambda_n = \mathcal{O}\Big(\sqrt{\frac{\log d}{n}}\Big),$$

l'inégalité oracle non-asymptotique en prédiction pour l'estimateur lasso est de la forme

$$\|g^\star - g_{\hat{\beta}}\|_n^2 \leq a \inf_{\beta \in \mathbb{R}^d} \{\|g^\star - g_\beta\|_n^2 + \varepsilon_{n,d}(\beta)\}, \tag{1.10}$$

avec $a \geq 1$, $\|g^\star - g_\beta\|_n^2$ le terme de biais et $\varepsilon_{n,d}(\beta)$ le terme de variance. Ici, on parlera de vitesse lente (respectivement rapide) si

$$\varepsilon_{n,d}(\beta) = \mathcal{O}\Big(\sqrt{\frac{\log d}{n}}\Big)$$

(respectivement $\mathcal{O}(\log d/n)$). Pour obtenir une inégalité du type (1.10) avec grande probabilité, on fixe un $\beta \in \mathbb{R}^d$ et on part simplement du fait que, par définition de $\hat{\beta}$, on a

$$R_n(g_{\hat{\beta}}) + \mathrm{pen}(\hat{\beta}) \leq R_n(g_\beta) + \mathrm{pen}(\beta).$$

Puis en utilisant (1.8) et (1.9) on peut écrire

$$R_n(g_\beta) = \|g^\star - g_\beta\|_n^2 + \frac{2}{n}\sum_{i=1}^n (g^\star - g_\beta)(X_i)\varepsilon_i$$

et on obtient

$$\|g^\star - g_{\hat{\beta}}\|_n^2 \leq \|g^\star - g_\beta\|_n^2 + \frac{2}{n}\sum_{i=1}^n (g_{\hat{\beta}} - g_\beta)(X_i)\varepsilon_i + \mathrm{pen}(\beta) - \mathrm{pen}(\hat{\beta}).$$

Moralement, il ne reste "plus qu'à" contrôler le processus

$$\frac{1}{n}\sum_{i=1}^n X_i\varepsilon_i$$

à l'aide d'une inégalité de concentration, par exemple une inégalité de Hoeffding si les variables aléatoires de bruit ε_i sont supposées centrées et bornées, ou une inégalité de Bernstein dans le cas plus général d'erreurs sous-exponentielles. Ces inégalités, utiles dans la suite du manuscrit, sont rappelées en Annexe A.1.3.

Pour obtenir une vitesse rapide, une hypothèse supplémentaire est à faire sur la matrice

$$\boldsymbol{G} = \boldsymbol{X}^\top \boldsymbol{X}/n$$

où on note $\boldsymbol{X} = [x_{i,j}]_{1\leq i\leq n; 1\leq j\leq d}$ la matrice de design. Dans le cas de la grande dimension ($d \gg n$), le problème des moindres carrés ordinaires n'a pas de solution unique puisque la matrice $\boldsymbol{G}$ est dégénérée, soit

$$\kappa = \min_{\beta\in\mathbb{R}^d\setminus\{0\}} \frac{\sqrt{\beta^\top \boldsymbol{G}\beta}}{\|\beta\|_2} = 0,$$

où κ la plus petite valeur propre de $\boldsymbol{G}$. L'idée est alors de faire une hypothèse plus faible que la définie positivité de la matrice $\boldsymbol{G}$ qui s'appelle condition aux valeurs propres restreintes $RE(s, c_0)$ (restricted eigenvalue) définie dans Bickel et al. [2009] par

$$\min_{\substack{J\subseteq\{1,\dots,d\}\\ |J|\leq s}} \min_{\substack{\beta\in\mathbb{R}^d\setminus\{0\}\\ \|\beta_{J^\complement}\|_1\leq c_0\|\beta_J\|_1}} \frac{\sqrt{\beta^\top \boldsymbol{G}\beta}}{\|\beta_J\|_2} > 0,$$

avec $s \in \{1, \dots, d\}$ un paramètre de sparsité, $c_0 > 0$ et $(\beta_J)_k = \beta_k$ pour $k \in J$ et $(\beta_J)_k = 0$ pour $k \in J^\complement = \{1, \dots, d\}\setminus J$.

La condition $RE(s, c_0)$ assure ainsi la définie positivité de $\boldsymbol{G}$ pour les vecteurs $\beta \in \mathbb{R}^d\setminus\{0\}$ vérifiant $\|\beta_{J^\complement}\|_1 \leq c_0\|\beta_J\|_1$, et permet de lier $\|(\hat{\beta}-\beta)_{\mathcal{A}(\beta)}\|_2$ à $\|g_{\hat{\beta}} - g_\beta\|_n$ pour tout $\beta \in \mathbb{R}^d$ afin d'obtenir la vitesse rapide.

1.2 Le cas de l'analyse de survie

L'analyse de survie est un ensemble de méthodes statistiques utilisées dans divers champs d'application où l'intérêt est porté sur l'apparition d'un événement donné. Il peut s'agir d'études dans des domaines aussi variés que la médecine, la démographie, la biologie, la sociologie ou encore l'économétrie. L'évènement d'intérêt peut être le décès, mais aussi l'apparition d'une maladie (par exemple, le temps avant une rechute, un rejet de greffe, ou le diagnostic d'un cancer) ou d'une guérison, la naissance d'un enfant, la panne d'une machine, la survenue d'un sinistre (en actuariat), ou encore le désabonnement d'un client pour un service donné. Ces quelques exemples, loin d'être exhaustifs, sont tous sujets à un intérêt scientifique dès lors que l'on tente de comprendre leur cause et d'établir des facteurs de risque.

1.2.1 Généralités

Pour plus de clarté et sans perte de généralité, nous utiliserons dans la suite le terme de "survie", et le vocabulaire relatif à la survie, plutôt "qu'événement d'intérêt". En analyse de survie, les données ont (au minimum) trois particularités : (1) la variable de sortie à expliquer (*label*) est le temps d'attente jusqu'à la survenue du décès : on appelle ce temps d'attente la durée de survie (qui est toujours positive ou nulle), (2) les observations sont censurées, dans le sens où pour certains exemples de l'échantillon d'apprentissage $\mathcal{D}_n$, le décès ne s'est pas produit au moment où les données sont analysées, et (3) nous avons à notre disposition des variables explicatives (covariables) pour lesquelles on questionne l'effet sur la survie.

La fonction de survie. Notons $T \geq 0$ la variable aléatoire modélisant la durée de survie que l'on suppose absolument continue, de densité notée classiquement f et de fonction de répartition notée F. Ainsi, $F(t) = \mathbb{P}[T \leq t]$ donne la probabilité que le décès se soit déjà produit avant le temps t. La fonction de survie, qui est d'un intérêt particulier dans ce contexe, est alors simplement définie par

$$S(t) = \mathbb{P}[T > t] = 1 - F(t) = \int_t^{+\infty} f(u)\mathrm{d}u \tag{1.11}$$

et donne donc la probabilité d'être toujours en vie à l'instant t.

Le risque instantané. Une autre fonction caractéristique de la distribution de T est la fonction de risque instantané, aussi appelée hasard ou intensité, et définie par

$$\lambda(t) = \lim_{\mathrm{d}t \to 0} \frac{\mathbb{P}[t \leq T < t + \mathrm{d}t | T \geq t]}{\mathrm{d}t}. \tag{1.12}$$

Le numérateur de cette expression est la probabilité conditionnelle que le décès se produise dans l'intervalle de temps $[t, t+\mathrm{d}t[$ sachant qu'il ne s'est pas produit avant, et le dénominateur est la taille de l'intervalle. En divisant l'un par l'autre, on obtient un risque de décès par unité de temps et en prendre la limite lorsque l'intervalle tend vers le singleton $\{t\}$ donne bien intuitivement le risque instantané de décès à l'instant t, conditionnellement au fait d'être toujours vivant jusqu'au temps t. En remarquant que

$$\mathbb{P}[t \leq T < t + \mathrm{d}t | T \geq t] = \frac{\mathbb{P}[t \leq T < t + \mathrm{d}t]}{S(t)},$$

on obtient l'expression

$$\lambda(t) = \frac{f(t)}{S(t)},$$

qui peut aussi faire office de définition.

La dernière égalité de (1.11) suggère que $-f$ est la dérivée de S et permet d'écrire

$$\lambda(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t).$$

En intégrant de 0 à t et en introduisant la condition naturelle $S(0) = 1$ (il est d'usage de supposer que le décès n'a pas eu lieu au temps initial), on peut alors exprimer la probabilité de survie jusqu'au temps t comme une fonction de l'intensité à tous les instants jusqu'au temps t, soit

$$S(t) = \exp\{-\int_0^t \lambda(u)\mathrm{d}u\} = \exp\{-\Lambda(t)\} \tag{1.13}$$

en notant $\Lambda(t)$ le risque cumulé entre l'instant initial et t.

La notion de censure. La seconde caractéristique de l'analyse de survie est le phénomène de censure : le fait que pour certains exemples, le décès s'est produit dans la fenêtre d'observation finie qui est intrinséque à $\mathcal{D}_n$, et pour d'autres le décès ne s'est pas produit.

Pour être un peu plus précis, différents méchanismes peuvent mener à cette notion de censure : il se peut que le décès d'un individu ne soit pas observé avant la fin de l'étude, que l'individu sorte de l'étude – pour une raison ou une autre – avant qu'on ait pu observer son décès, ou encore que l'individu décède d'une autre cause que de la maladie étudiée par exemple. Tout ce qu'on sait pour ces individus, c'est que le temps de survie excède le temps d'observation. On aimerait alors prendre en compte et modéliser cette information.

Suivant la censure dite de *type I*, on suit n individus sur une fenêtre temporelle finie notée τ (elle peut aussi dépendre de chaque individu $i \in \{1, \ldots, n\}$, ce qui ne change pas grand chose). Une façon de modéliser le phénomène consiste à associer à chaque individu i une variable aléatoire modélisant le temps de censure potentielle en plus de la variable aléatoire T_i, et de supposer que l'on observe le minimum entre ces deux temps, conjointement à un indicateur de censure qui renseigne si le temps observé est censuré ou non. Cette modélisation permet d'écrire la vraisemblance en prenant en compte l'information apportée par chaque individu, comme nous allons le voir dans la suite après avoir introduit quelques notions fondamentales.

1.2.2 Formalisme

Le cadre général dans lequel se trouve l'analyse de survie, introduit dans Aalen [1980], est celui des processus de comptage et inclut plusieurs contextes comme les processus de Poisson marqués, les processus de Markov, ou encore les données censurées qui nous intéressent particulièrement [Andersen et al., 1993].

Nous donnons pour commencer quelques notions introductives, qui sont complétées dans l'Annexe A.1.1. Notons $C_i \geq 0$ la variable aléatoire modélisant la censure pour l'individu i,

$$Z_i = \min\{T_i, C_i\}$$

le temps censuré observé et

$$\Delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$$

l'indicateur de censure correspondant. Z_i vaut donc T_i dans le cas de l'observation du décès de l'individu i ($\Delta_i = 1$) et C_i si le décès n'a pas été observé ($\Delta_i = 0$). Dans la suite, nous ferons l'hypothèse classique que la censure est non informative, c'est-à-dire qu'elle n'apporte pas d'information sur l'état de santé du patient, ce qui se traduit par l'indépendance entre T_i et C_i conditionnellement aux covariables X_i. Cela sera particulièrement utile lors de l'écriture de la vraisemblance, cette hypothèse étant faite dans la plupart des modèles usuels [Klein and Moeschberger, 2005].

On considère l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et la filtration $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$ définie par

$$\mathcal{F}_t = \sigma\Big\{X_i, N_i(s), Y_i(s) : s \in [0,t], i \in \{1,\dots,n\}\Big\},$$

avec pour l'individu i la notation $X_i \in \mathbb{R}^d$ pour le vecteur aléatoire des covariables $\mathcal{F}_0$-mesurables, N_i le processus de comptage marqué [4] qui saute lorsque le $i^{\text{ème}}$ individu décède, c'est-à-dire

$$N_i(t) = \mathbb{1}_{\{Z_i \leq t, \Delta_i = 1\}},$$

et Y_i le processus aléatoire modélisant la présence de risque tel que

$$Y_i(t) = \mathbb{1}_{\{Z_i \geq t\}}.$$

On note Λ_i le compensateur du processus N_i par rapport à $\mathcal{F}$ tel que $M_i = N_i - \Lambda_i$ soit une $\mathcal{F}$-martingale. On parle de modèle à intensité multiplicative d'Aalen pour N_i lorsque

$$\Lambda_i = \int_0^t \lambda(s, X_i) Y_i(s) \mathrm{d}s$$

avec λ l'intensité du modèle définie dans (1.12), que l'on pourra également noter $\lambda^\star$ dans la suite pour insister sur sa non-connaissance et sur le fait qu'il s'agit de la "vraie" intensité.

Avant d'aller plus en avant dans l'élaboration d'une vraisemblance dans ce contexte, arrêtons nous un instant pour utiliser les notations qui viennent d'être introduites afin de répondre à deux questions pratiques fondamentales (qui seront souvent considérées dans la suite du manuscrit) : *comment estimer la fonction de survie d'un groupe d'individus ? Comment comparer des fonctions de survie issues de différents groupes d'individus ?*

4. Les notions utilisées sans définition sont définies en Annexe A.1.1.

Estimateur de Kaplan-Meier. L'estimateur de Kaplan-Meier introduit dans Kaplan and Meier [1958] est un estimateur non paramétrique de la fonction de survie qui découle de la simple idée que si on a deux temps ordonnés $t' < t$, alors

$$\mathbb{P}[T > t] = \mathbb{P}[T > t | T > t'] \times \mathbb{P}[T > t'].$$

En notant $(z_{(i)})_{i\in\{1,\dots,n\}}$ les temps censurés ordonnés de l'échantillon (*i.e.* rangés par ordre croissant), on a alors

$$\mathbb{P}[T > z_{(i)}] = \prod_{k=1}^{i} \mathbb{P}[T > z_{(k)} | T > z_{(k-1)}]$$

avec la convention $z_{(0)} = 0$. Or, la probabilité p_k de mourir dans l'intervalle $]z_{(k-1)}, z_{(k)}]$ sachant que l'on était vivant en $z_{(k-1)}$, soit

$$p_k = \mathbb{P}[T \leq z_{(k)} | T > z_{(k-1)}],$$

est naturellement estimée par

$$\hat{p}_k = \frac{d_k}{\sum_{i=1}^{n} Y_i(z_{(k)})},$$

où on note d_k le nombre de décès en $z_{(k)}$ (qui vaut δ_k sous l'hypothèse théorique classique que les temps de décès sont distincts si T est une variable absoluement continue). En effet, $\sum_{i=1}^{n} Y_i(z_{(k)})$ représente le nombre d'individus à risque juste avant le temps $z_{(k)}$. L'estimateur de Kaplan-Meier est alors simplement donné par

$$\hat{S}(t) = \prod_{\substack{i=1,\dots,n \\ y_{(i)} \leq t}} (1 - \hat{p}_k).$$

On obtient ainsi une fonction en escalier décroissante et continue à droite. Il existe des estimateurs de la variance de $\hat{S}(t)$ (estimateur de Greenwood [Klein, 1991] par exemple), on peut exprimer des intervalles de confiance [Rothman, 1978], et sous certaines conditions on montre que l'estimateur de Kaplan-Meier est uniformément consistant et asymptotiquement normal quand le nombre d'individus à risque est grand.

Test du logrank. Le test du logrank [Bland and Altman, 2004] est une approche non paramétrique pour comparer les fonctions de survie S_A et S_B de deux groupes A et B d'individus. Le principe du test consiste à comparer le nombre de décès observés dans chaque groupe au nombre de décès attendus, calculés sous l'hypothèse nulle d'égalité des distributions de survie, soit

$$H_0 : S_A = S_B.$$

Il permet de prendre en compte toute l'information sur l'ensemble du suivi sans nécessité de faire des hypothèses sur la distribution des temps de survie.

Pour chaque temps (ordonné) $z_{(i)}$, on note D_A^i (respectivement D_B^i) la variable aléatoire modélisant le nombre de décès au temps $T_{(i)}$ dans le groupe A (respectivement B), et d_i le nombre de décès total observés à cet instant (somme des deux réalisations). De même, on note y_A^i (respectivement y_B^i) le nombre d'individus à risque observés au temps $z_{(i)}$ dans le groupe A (respectivement B), et y_i le nombre total de sujets à risques observés à cet instant.

L'idée est de comparer les pourcentages de décès parmi les sujets à risque dans chacun des groupes en utilisant le test du Chi-2 [Moore, 1976]. On peut en effet vérifier que D_A^i suit une loi hypergéométrique d'espérance

$$\mathbb{E}[D_A^i] = \frac{d_i y_A^i}{y_i}$$

et de variance

$$\mathrm{Var}[D_A^i] = \frac{y_i - d_i}{y_i - 1} \times \frac{d_i y_A^i y_B^i}{y_i^2},$$

et que sous H_0, $D_A^i - \mathbb{E}[D_A^i]$ suit asymptotiquement une loi $\mathcal{N}(0, \mathrm{Var}[D_A^i])$, et donc que

$$\frac{\left(D_A^i - \mathbb{E}[D_A^i]\right)^2}{\mathrm{Var}[D_A^i]} \xrightarrow[n\to+\infty]{d.} \chi^2(1).$$

En sommant sur tous les individus et en ajoutant des pondérations ω_i on obtient la statistique suivante

$$\mathcal{S} = \frac{\left(\sum_{i=1}^n \omega_i (D_A^i - \mathbb{E}[D_A^i])\right)^2}{\sum_{i=1}^n \omega_i^2 \mathrm{Var}[D_A^i]}$$

et sous H_0, par indépendance de D_A^i et D_A^j pour $i \neq j$, on a également

$$\mathcal{S} \xrightarrow[n\to+\infty]{d.} \chi^2(1).$$

Le test du logrank considère alors que chaque décès a le même poids quel que soit l'instant où il survient, soit $\omega_i = 1$ pour tout $i \in \{1, \ldots, n\}$. Choisir une pondération différente amène à un test différent : le test de Gehan considère par exemple que les poids sont plus élevés pour les décès précoces en prenant $\omega_i = y_i$.

Bien entendu, le test se généralise au cas où il y a $K \geq 2$ groupes, la statistique de test correspondante suivra alors asymptotiquement une distribution $\chi^2(K-1)$.

Les deux questions posées ayant été traitées, revenons au problème de l'écriture de la vraisemblance d'un modèle en analyse de survie.

Construction de la vraisemblance. On considère l'échantillon *i.i.d.*

$$\Big\{\Big(X_i, N_i(t), Y_i(t)\Big) : t \in [0,\tau], i \in \{1,\dots,n\}\Big\},$$

avec $\tau > 0$ la durée de l'étude, ce qui se résume en pratique dans le cas des données censurées à droite aux données

$$\mathcal{D}_n = \{(x_1, z_1, \delta_1), \dots, (x_n, z_n, \delta_n)\}.$$

Oublions dans un premier temps les covariables x_i [5], et supposons que l'on observe le temps z_i pour l'individu i. Si le temps est non censuré, alors sa contribution à la vraisemblance est donnée par la valeur de la densité évaluée au temps $z_i = t_i$, soit

$$L_i = f(t_i) = S(t_i)\lambda(t_i).$$

En revanche, si $z_i = c_i$ et que l'individu i est toujours vivant à l'instant z_i (donnée censurée), alors sous l'hypothèse de censure non informative, on sait seulement que sa survie dépasse z_i et sa contribution à la vraisemblance est donnée par

$$L_i = S(c_i).$$

La vraisemblance peut alors s'écrire

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} S(z_i)\lambda(z_i)^{\delta_i}.$$

Sans prendre en compte les covariables $x_i \in \mathbb{R}^d$, on obtient une population homogène, dans le sens où les durées de survie de tous les individus sont gouvernées par la même fonction de survie S. Mais la troisième caractéristique de l'analyse survie est la présence de covariables explicatives pouvant affecter le temps de survie, l'idée est alors de modéliser leurs effets.

Différents modèles ont été proposés dans la littérature à ce sujet. On trouve par exemple les modèles de vie accélerée [Bagdonavicius and Nikulin, 2001] (accelerated life models) qui proposent d'exprimer le logarithme de la durée de survie par un modèle linéaire classique, à savoir

$$\log T = X^\top \beta^\star + \varepsilon,$$

où ε est une erreur avec une certaine loi. En notant S_0 la fonction de survie "de base", soit pour un individu tel que $X = 0$, donc $S_0(t) = \mathbb{P}[\exp(\varepsilon) > t]$, la fonction de survie d'un individu i s'exprime

$$S_i(t|X_i = x_i) = S_0\Big(t\exp(-x_i^\top \beta^\star)\Big).$$

5. Précisons qu'une notation en lettre minuscule, qui signifie qu'on parle de la réalisation de la variable aléatoire, sera préférée pour des quantités comme la vraissemblance, mais qu'on préférera une notation en lettre majuscule (faisant intervenir l'échantillon théorique) pour les estimateurs et les considérations théoriques.

On comprend ainsi le nom donné à cette famille de modèles, puisque le terme $\exp(-x_i^\top \beta^\star)$ est un facteur d'accélération : un changement dans les covariables modifie l'échelle de temps. Ces modèles sont par exemple utilisés en économétrie ou en démographie. Mais le modèle le plus populaire et utilisé en analyse de survie est sans doute le modèle à risques proportionnels introduit dans Cox [1972]. Il sera d'un intérêt particulier dans la suite et nous en présentons les principes généraux dans la section suivante.

1.2.3 Le modèle de Cox

Le modèle de Cox [Cox, 1972], dit à risques proportionnels, suppose que l'intensité a la forme

$$\lambda_i(t|X_i = x_i) = \lambda_0^\star(t) \exp\Big(f^\star(x_i)\Big), \tag{1.14}$$

où $\lambda_0^\star$ est le risque de base décrivant le risque pour les individus tels que $x_i = 0$ qui servent de référence, et $\exp\Big(f^\star(x_i)\Big)$ est le risque relatif associé aux covariables de l'individu i. Le modèle sépare ainsi les effets du temps et ceux des covariables à travers une structure multiplicative (effet multiplicatif des covariables sur la fonction de risque) et on parle de risques proportionnels puisque le rapport des fonctions de risque (1.14) de deux individus distincts i et j ne dépend pas du temps : les fonctions de risque sont donc proportionnelles. En effet, on a

$$\frac{\lambda_i(t|X_i = x_i)}{\lambda_j(t|X_j = x_j)} = \exp\Big(f^\star(x_i) - f^\star(x_j)\Big)$$

pour tout $t \in [0, \tau[$. C'est une conséquence du modèle mais c'est surtout une hypothèse forte qu'il convient de vérifier, ce qui sera brièvement discuté en Annexe A.1.2.

En général, aucune forme *a priori* n'est imposée au risque de base. On peut en faire de même avec $f^\star$ dans une modélisation de Cox non-paramétrique [Hastie and Tibshirani, 1990]. Mais la modélisation classique, que l'on choisira dans la suite du manuscrit, est semi-paramétrique et suppose que $f^\star$ soit linéaire, c'est-à-dire

$$f^\star(x_i) = x_i^\top \beta^\star$$

avec $\beta^\star \in \mathbb{R}^d$. Il est aussi possible d'obtenir différents modèles à risques proportionnels en faisant différentes hypothèses sur le risque de base, un choix classique étant celui d'une loi de *Weibull* avec

$$\lambda_0^\star(t) = \frac{\mu}{\phi}\Big(\frac{t}{\phi}\Big)^{\mu-1},$$

où $\mu > 0$ est un paramètre de forme et $\phi > 0$ un paramètre d'échelle. Précisons que cette famille de lois est la seule donnant lieu à un modèle appartenant simultanément à la famille des modèles à risques proportionnels et à celle des modèles

de vie accélerée [Kalbfleisch and Prentice, 2011]. Elles sont très utiles pour modéliser des risques monotones (croissants si $\mu \in]0,1[$ et décroissants si $\mu > 1$), mais deviennent mal adaptées lorsque les risques sont en forme de cloche, ce qui est courant dans divers cas pratiques. Une alternative est l'utilisation de lois de Weibull généralisées [Mudholkar and Kollia, 1994].

Précisons aussi qu'il existe des extensions directes du modèle de Cox (1.14), en supposant par exemple que les covariables x_i peuvent varier au cours du temps [Sueyoshi, 1992] (on perd alors la propriété de proportionnalité des risques), ou que les effets (c'est-à-dire le vecteur de coefficients $\beta^\star$) peuvent varier au cours du temps [Tian et al., 2005], ou bien même les deux simultanément [Therneau and Grambsch, 2013] tel que

$$\lambda_i(t) = \lambda_0^\star(t) \exp\Big(x_i(t)^\top \beta^\star(t)\Big).$$

D'autres extensions classiques sont par exemple le modèle de Cox stratifié [Therneau and Grambsch, 2013], utile si une variable qualitative ne vérifie pas l'hypothèse des risques proportionnels : on peut alors considérer que le risque de base est différent dans les différentes "strates" définies par la variable en question ; ou encore les modèles de fragilité (frailty) [Duchateau and Janssen, 2007] utiles pour prendre en compte l'hétérogénéité des observations avec une dépendance entre les temps de survie d'individus de différents groupes : le risque pour l'individu i appartenant au groupe $k \in \{1, \dots, K\}$ s'exprime

$$\lambda_{k,i}(t|X_{k,i} = x_{k,i}, Z_k = z_k) = \lambda_0^\star(t) z_k \exp(x_{k,i}^\top \beta^\star)$$

où z_k est la réalisation d'une variable latente Z_k appelée "fragilité" du groupe k et supposée *i.i.d.*, souvent d'espérance nulle et de variance θ (pour des questions d'identifiabilité du modèle) mesurant l'hétérogénéité entre les groupes. Cette dernière extension sera par exemple considérée parmi les modèles de référence dans le Chapitre 4.

Ainsi, deux paramètres sont inconnus dans le modèle de Cox : le vecteur de régression $\beta^\star \in \mathbb{R}^d$ et la fonction λ_0. Beaucoup d'études ne s'intéressent qu'à l'estimation de $\beta^\star$, ce qui sera parfois notre cas, ce qui permet de déterminer les facteurs pronostiques qui influencent la durée de survie, et d'ordonner les individus suivant leurs risques relatifs. En revanche, pour pouvoir prédire la durée de survie d'un individu i sachant ses covariables x_i, il est indispensable d'estimer également la fonction λ_0 puisque d'après l'Équation (1.13), on a

$$S_i(T_i > t|X_i = x_i) = \exp\Big(-\int_0^t \lambda_0^\star(u) \exp(x_i^\top \beta^\star) \mathrm{d}u\Big).$$

Présentons alors les méthodes d'estimation classiques de ces deux paramètres lorsque $d < n$.

Estimation. Concernant le paramètre de régression $\beta^\star \in \mathbb{R}^d$, l'estimation se fait en minimisant l'opposé de la log-vraisemblance partielle introduite dans [Cox, 1972] qui s'exprime par

$$\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \Big(x_i^\top \beta - \log \sum_{i' : z_{i'} \geq z_i} \exp(x_{i'}^\top \beta) \Big).$$

Lorsqu'on s'intéressera à des propriétés théoriques, on utilisera davantage l'écriture

$$\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \Big(X_i^T \beta - \log S_n(t,\beta) \Big) \mathrm{d}N_i(t)$$

avec

$$S_n(t,\beta) = \frac{1}{n}\sum_{i=1}^{n} \exp(X_i^T \beta) Y_i(t) \tag{1.15}$$

qui utilise l'échantillon aléatoire (lettres capitales), le passage d'une écriture à l'autre étant montrée en Annexe A.1.2 où on détaillera aussi comment obtenir la log-vraisemblance partielle à partir de la log-vraisemblance du modèle.

Ainsi, la log-vraisemblance partielle ne fait pas intervenir le risque de base $\lambda_0^\star$ et l'estimateur naturel de $\beta^\star$ est donné par

$$\hat{\beta} \in \mathrm{argmin}_{\beta \in \mathbb{R}^d} -\ell_n(\beta). \tag{1.16}$$

Le problème (1.16) n'admet pas de solution explicite en général, mais étant convexe, on a recours aux algorithmes d'optimisation convexe. L'estimateur obtenu est consistant et asymptotiquement normal [Andersen and Gill, 1982].

Concernant l'estimation du risque de base $\lambda_0^\star$, un estimateur classique est obtenu à partir d'un estimateur du risque de base cumulé défini par

$$\Lambda_0(t) = \int_0^t \lambda_0^\star(u) \mathrm{d}u.$$

L'estimateur de cette quantité, de type estimateur de Nelson-Aalen [Aalen, 1978], est défini pour un $\beta \in \mathbb{R}^d$ fixé par

$$\hat{\Lambda}_0(t,\beta) = \int_0^t \frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u,\beta)} \mathrm{d}\bar{N}(u),$$

avec

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{et} \quad \bar{N} = \frac{1}{n}\sum_{i=1}^{n} N_i.$$

L'estimateur de Breslow [Breslow, 1972], très utilisé en pratique, est alors simplement donné par $\hat{\Lambda}_0(t,\hat{\beta})$. Un estimateur de $\lambda_0^\star$ obtenu à partir de l'estimateur de Breslow est alors donné par

$$\hat{\lambda}_h(t) = \frac{1}{nh}\sum_{i=1}^{n} \int_0^\tau K\Big(\frac{t-u}{h}\Big) \frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u,\hat{\beta})} \mathrm{d}N_i(u),$$

avec K une fonction noyau d'intégrale 1 et $h > 0$ la longueur de la fenêtre.

Le cas $d \gg n$. Les estimateurs qui viennent d'être présentés ont de bonnes propriétés théoriques lorsque $d < n$. Par exemple, on a $\|\hat{\beta} - \beta^\star\|_2^2 = \mathcal{O}(d/n)$, mais cela implique aussi que cet estimateur n'est plus consistant dans le cas de la grande dimension. Une procédure lasso appliquée à la log-vraisemblance partielle de Cox a alors été proposée dans Tibshirani [1997] et différents résultats théoriques ont ensuite été montrés pour cet estimateur : une inégalité asymptotique en estimation en norme ℓ_2 [Bradic et al., 2011], des inégalités oracles en prédiction [Kong and Nan, 2014] ainsi que des résultats en prédiction en norme ℓ_p [Huang et al., 2013]. L'adaptive lasso a également été considéré pour le modèle de Cox, avec des résultats asymptotiques d'estimation [Zhang and Lu, 2007] et de consistance en sélection [Zhang and Lu, 2007, Zou, 2008].

Précisons que la procédure lasso a également été considérée dans d'autres modèles d'analyse de survie comme le modèle additif d'Aalen [Martinussen and Scheike, 2009] avec des inégalités oracle non-asymptotique démontrées dans ce contexte [Gaïffas and Guilloux, 2012].

1.3 Cheminement et contributions

Dans cette section, nous présentons le cheminement suivi tout au long de cette thèse, correspondant à l'ordre des chapitres du manuscrit. Chaque chapitre est un projet mené de façon distincte et aboutissant à un papier de recherche, ces différents projets s'étant naturellement enchaînés lors de la thèse. Précisons que chaque chapitre comporte un paragraphe "Software" pointant vers un dépôt GitHub qui met à disposition le code open source ayant généré les différentes figures et permis les analyses du chapitre correspondant.

Chapitre 2 : Trajectories of biological values and vital parameters, a retrospective cohort study on non-complicated vaso-occlusive crises

Contexte. La drépanocytose est la maladie génétique la plus fréquente dans le monde, avec environ 310000 naissances concernées chaque année [Piel et al., 2013]. Elle est due à une mutation génétique de l'hémoglobine qui est une protéine des globules rouges assurant le transport de l'oxygène dans le sang. Cette mutation favorise l'apparition de globules rouges rigides en forme de faucille susceptibles de bloquer la circulation sanguine au niveau des capillaires de certains organes (os, reins ou cerveau par exemple) provoquant des douleurs aigües [Stuart and Nagel, 2004]. Cette manifestation clinique est appelée crise vaso-occlusives (CVO) et chez la plupart des patients, l'injection intraveineuse d'opiacés à intervalles réguliers est requit pour calmer la douleur, jusqu'à la fin de la crise qui est traitée par hydratation.

Il n'existe pas de biomarqueur fiable pour diagnostiquer une COV, bien qu'il ait été observé que les COVs sont souvent (mais pas toujours) associées à une hémolyse accrue (élévation de la lactate déshydrogénase (LDH) et bilirubine), à une augmentation de l'anémie ou encore de la protéine C-réactive (CRP). Ces biomarqueurs inflammatoires sont surveillés pour détecter la survenue d'une complication infectieuse au cours des COVs, mais l'évolution "normale" des paramètres vitaux et des résultats biologiques de laboratoire effectués de façon récurrente au moment du diagnostique et de l'évolution des COVs non compliquées est jusqu'alors inconnue.

Le premier objectif de ce chapitre est alors de décrire le comportement et l'évolution des biomarqueurs et des paramètres vitaux lors d'une CVO dite non compliquée (c'est-à-dire sans apparition d'une infection ou d'une complication quelconque), dans le but d'aider au diagnostic d'une part, et à détecter la présence d'une possible complication d'autre part. Les résultats sont obtenus à partir d'une étude rétrospective sur une cohorte de l'hôpital universitaire européen George Pompidou (HEGP) à Paris, dont le service de médecine interne est un des centres de référence pour les patients drépanocytaires adultes. Le second objectif est d'identifier quel(s) biomarqueur(s) et/ou paramètre(s) vital(aux) est à surveiller avec attention dans les jours suivant une admission pour CVO, et orienter le choix des tests de laboratoires à effectuer ainsi que les moments opportuns pour les réaliser.

Méthode. Les données recueillies sont principalement de type longitudinales (séries temporelles). Elles ont été extraites de l'entrepôt de données de l'HEGP et concernent les patients y étant admis pour CVO entre le 1er janvier 2010 au 31 décembre 2015. De nombreux séjours sont exclus afin de ne concerver que les CVOs dites non compliquées (voir Section 2.2.2), il reste alors 329 séjours pour 164 patients.

Pour décrire les trajectoires moyennes des données longitudinales, la méthodologie suivie (détaillée en Section 2.C) est telle que pour chaque variable longitudinale, on génére une grille uniforme de temps t_k, puis un *spline* du premier ordre f_i est ajusté pour les trajectoire individuelle de chaque séjour i permettant ensuite de calculer les valeurs $f_i(t_k)$ pour chaque temps de la grille. On obtient ainsi, pour chaque variable longitudinale, une matrice dont le nombre de colonnes est la taille de la grille et où le nombre de lignes correpond au nombre de séjours, à partir de laquelle on déduit une trajectoire moyenne avec intervalles de confiance gaussiens.

Résultats. De nombreuse statistiques descriptives sont présentées dans la Section 2.3.1, ainsi que les résultats de différents tests statistiques univariés. Puis, les résultats les plus intéressants cliniquement sont présentés et analysés. La Figure 1.4 donne un aperçu des différents graphiques obtenus. Plusieurs biomarqueurs et paramètres vitaux pertinents ont été identifiés comme particulièrement importants à surveiller lors d'une COV, à savoir l'hémoglobine, les leucocytes (et plus précisément les éosinophiles), la CRP et la température. En ce qui concerne l'hémoglobine, elle

est rarement revenue au niveau de base avant la sortie de l'hôpital, mais une baisse au cours du séjour peut prédire une réadmission précoce. Les trajectoires des leucocytes ont montré des tendances spécifiques, en particulier pour les neutrophiles et les éosinophiles, qui pourraient aider à mieux évaluer le diagnostic de COV. Une analyse des proportions inférieures ou supérieures à certains seuils a par exemple montré que plus de 95% des séjours pour COV non compliquée présentaient une valeur de référence de la CRP inférieure à 100 mg/L le premier jour après l'admission, et pas de fièvre pendant toute la durée du séjour, ce qui suggère que la fièvre et/ou qu'une CRP supérieure à 100 mg/L à l'admission ne doivent pas être attribuées à la COV elle-même.

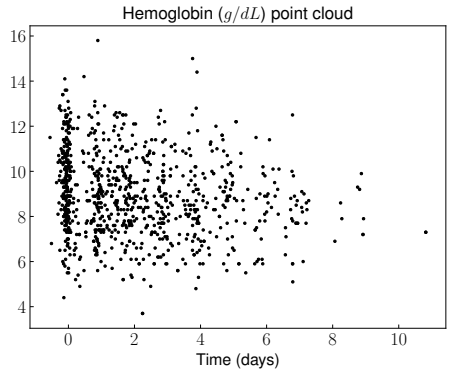
Ainsi, ce travail illustre notamment comment extraire de l'information pertinente à partir d'un entrepôt de données cliniques de grande dimension, en travaillant sans hypothèse *a priori* et en considérant différentes méthodes.

Article associé.
R. Veil, **S. Bussy**, J.B. Arlet, A.S. Jannot et B. Ranque
Trajectories of biological values and vital parameters : a retrospective cohort study on non-complicated vaso-occlusive crises
Soumis à *Haematologica*, 2018.

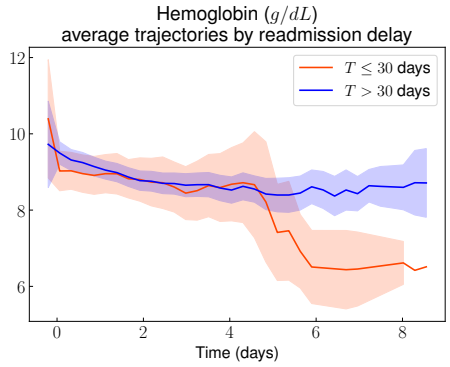
Code. Tous des codes implémentés (principalement en `Python`) pour ce chapitre sont disponibles à l'adresse `https://github.com/SimonBussy/redcvo` sous la forme de programmes annotés et l'ensemble des figures y sont disponibles (pour tous les biomarqueurs et paramètres vitaux recueillis lors d'un séjour à l'hôpital pour CVO, pas seulement ceux mentionnés dans le chapitre).

Chapitre 3 : Early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework

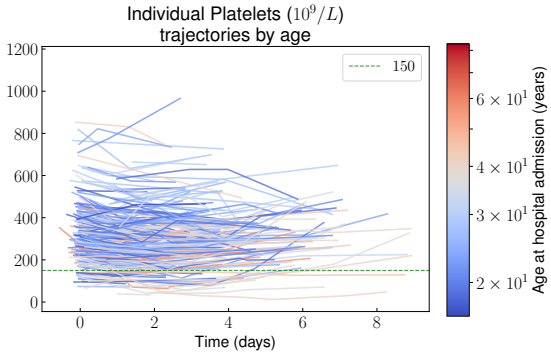
Contexte. À partir de l'étude de la cohorte présentée au Chapitre 2, la question qui se pose naturellement est celle de l'utilisation de ces données complexes et riches pour tenter de prédire la réhospitalisation précoce pour ces patients. En effet, ces patients sont sujets à des rechutes, c'est-à-dire qu'une autre CVO peut se déclencher dans les quelques jours suivant la sortie d'une hospitalisation pour CVO [Bunn, 1997, Platt et al., 1991]. Les cliniciens parlent alors de rechute, dans le sens où la seconde CVO a de fortes chances d'être "liée" à la précédente. D'où l'idée de prédire le risque d'une réhospitalisation précoce pour un séjour donné. Bien qu'il existe des études orientées sur les facteurs de risque [Brousseau et al., 2010, Rees et al., 2003], très peu d'entres elles sont centrées sur cette question de la prédiction de la réhospitalisation précoce.



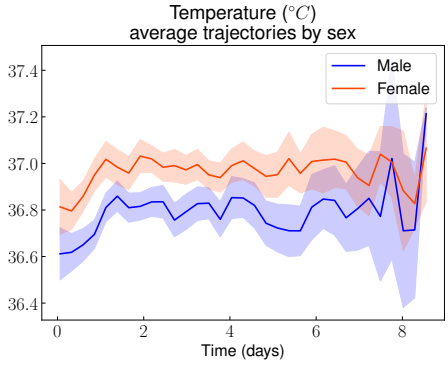
(a) Nuage de points pour l'hémoglobine



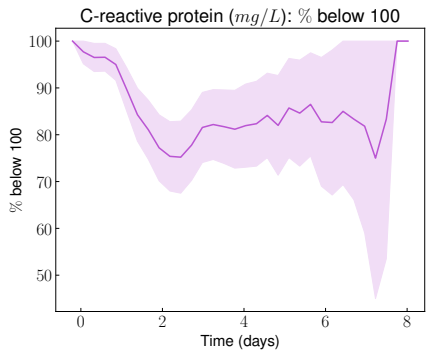
(b) Trajectoires moyennes d'hémoglobine pour les patients avec rechute dans les 30 jours vs. les autres



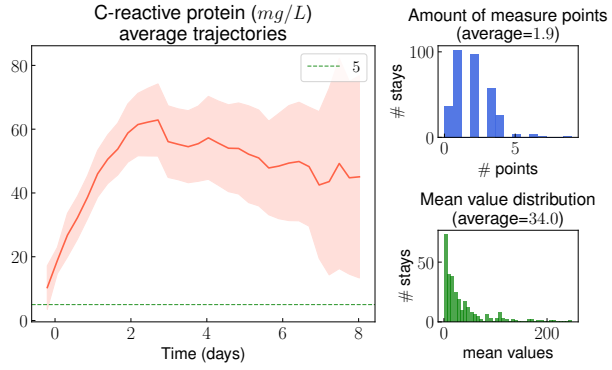
(c) Trajectoires individuelles des plaquettes avec gradient de couleurs suivant l'âge



(d) Trajectoires moyennes de la température suivant le sexe



(e) Évolution du % de patients avec une CRP $< 100mg/L$



(f) Trajectoire moyenne de CRP et informations complémentaires

FIGURE 1.4 Échantillon de résultats graphiques provenant de la Section 2.3.

Le but de ce chapitre est de comparer différentes approches statistiques pour prédire la réhospitalisation et d'identifier les variables importantes pour cette prédiction, en considérant simultanément les performances en prédiction et l'aspect de sélection de variables dans un contexte de grande dimension, aspects qui sont trop souvent traîtés séparément. L'évènement réhospitalisation sera considéré pour certaines approches comme une variable binaire et pour d'autres dans un contexte d'analyse de survie.

Dans le cadre de prédiction binaire, l'objectif est de prédire si le patient va déclencher une CVO dans un délai ϵ fixé suivant sa sortie de l'hôpital. On prendra par exemple dans cette étude une valeur de 30 jours pour ϵ, qui est un délai classique dans les études concernant la drépanocytose [Brousseau et al., 2010]. Le choix *a priori* de ce délai est arbitraire, et toute conclusion tirée dans ce contexte, en terme de prédiction ou de sélection de variables, est entiérement conditionnée par lui : il est donc périlleux d'inférer des conclusions générales. Il y a, de plus, une perte d'information temporelle puisque la granularité des temps de réhospitalisation des patients est réduite à une information binaire : lors de la phase d'apprentissage, les modèles de classification binaire ont la même information, en terme de risque, pour deux patients avec des temps de retour de 31 jours et 6 mois par exemple.

L'alternative est l'analyse de survie, introduite dans la Section 1.2, qui considère comme variable explicative la durée jusqu'à la réhospitalisation et "prend en compte" l'intégralité de l'information temporelle. Afin de comparer les deux approches en terme de prédiction, l'idée est d'utiliser les fonctions de survie estimées des modèles d'analyse de survie, et de les évaluer en ϵ pour obtenir une prédiction binaire.

Méthode. Une méthodologie d'extraction de covariables à partir des données longitudinales est proposée dans la Section 3.2.2, avec par exemple la pente d'une régression linéaire ajustée sur les 48 heures précédent la sortie de l'hôpital, ou encore les hyper-paramètres des noyaux de processus gaussiens [Pimentel et al., 2013] ajustés pour chaque trajectoire individuelle.

Dans le contexte de prédiction binaire, on considère les modèles de l'état de l'art suivant : la régression logistique (LR) [Hosmer Jr et al., 2013], les machines à vecteur de support (SVM) [Schölkopf and Smola, 2002], les forêts aléatoires (RF) [Breiman, 2001], le gradient boosting (GB) [Friedman, 2002] ainsi que les réseaux de neuronnes (NN) [Yegnanarayana, 2009].

Pour ce qui est du contexte d'analyse de survie, les modèles considérés sont les suivant : le modèle de Cox [Cox, 1972], le modèle de CURE [Farewell, 1982] et enfin le modèle C-mix [Bussy et al., 2018] introduit au Chapitre 4, qui a précisément été conçu face au problème rencontré dans le Chapitre 3, à savoir celui de la prédiction d'une sous-population à fort risque de réhospitalisation précoce.

De plus amples détails concernant les méthodes statistiques utilisées sont donnés dans la Section 3.2.3.

Résultats. La Table 1.2 rend compte des résultats de prédiction binaire obtenus pour le choix du seuil $\epsilon = 30$ jours, en terme d'AUC [Bradley, 1997]. La Figure 1.5 donne, quant à elle, un aperçu des résultats graphiques obtenus et détaillés dans la Section 3.3.

Ainsi, l'étude suggère notamment qu'entraîner des modèles d'analyse de survie utilisant toute l'information temporelle pour ensuite se servir des fonctions de survie estimées afin de prédire le risque d'apparition de l'événement d'intérêt avant un certain délai, peut nettement améliorer les prédictions obtenues par des modèles directement entraînés dans un cadre binaire. Cela semble faire sens intuitivement, et il serait intéressant de confirmer ce phénomène sur d'autres jeux de données, par exemple dans un contexte de prédiction du désabonnement de clients (*churn prediction*) pour un service donné : cette application est très courante et quasiment toujours traîtée dans un cadre de prédiction binaire.

Enfin, précisons que le modèle C-mix, conçu pour répondre spécifiquement au problème, obtient les meilleurs résultats et dispose de surcroit d'avantages attrayants pour l'interprétabilité des résultats.

TABLE 1.2 Résultats des prédictions binaires des différents modèles considérés en terme d'AUC.

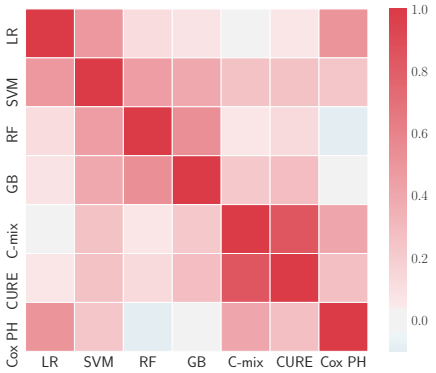
SVM	GB	LR	NN	RF	CURE ($\epsilon = 30$)	Cox PH ($\epsilon = 30$)	C-mix ($\epsilon = 30$)
0.524	0.561	0.616	0.707	0.738	0.831	0.855	**0.940**

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
Soumis à *BMC Medical Research Metodology*, 2018.

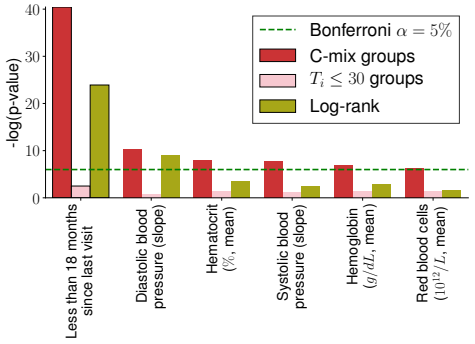
Code. Tous des codes implémentés (principalement en `Python`) pour ce chapitre, permettant notamment la génération des figures qu'on y trouve, sont disponibles à l'adresse `https://github.com/SimonBussy/early-readmission-prediction` sous la forme de notebooks annotés.

Chapitre 4 : C-mix, a high dimensional mixture model for censored durations

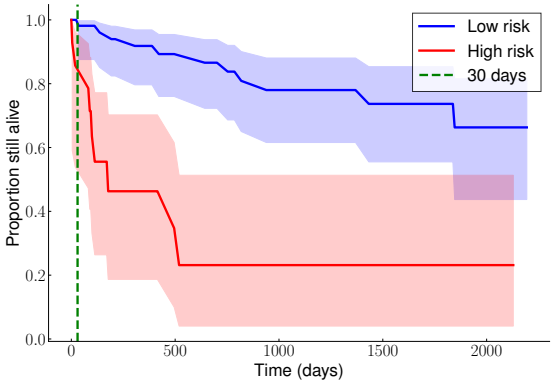
Contexte. Une question récurrente en médecine personnalisée est celle d'identifier des sous-groupes de patients de différents pronostics, en se basant par exemple sur



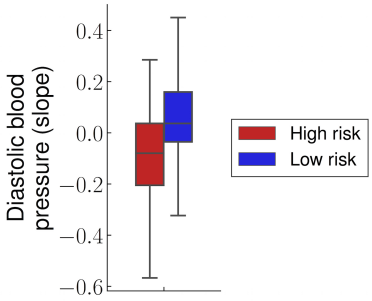
(a) Matrice de corrélation en sélection de variables des modèles considérés.



(b) Comparaison de tests basés sur les groupes prédis par le C-mix, sur le délai de $\epsilon = 30$ jours et sur les temps de survie (logrank).



(c) Estimation des courbes de survie des deux groupes prédis par le C-mix.



(d) Boîte à moustache de la pression diastolique pour les deux groupes prédis par le C-mix.

FIGURE 1.5 Échantillon de résultats graphiques provenant de la Section 3.3.

leurs expressions génétiques [Rosenwald et al., 2002]. Prédire le sous-groupe d'un patient est classique dans un contexte de classification, lorsque les sous-groupes sont connus à l'avance [Hastie et al., 2001a, Tibshirani et al., 2002]. Dans ce chapitre, on suppose au contraire que les groupes ne sont pas connus.

Dans ce contexte, une approche répandue est de détecter des sous-groupes dans l'espace des covariables de façon non-supervisée (avec des méthodes dites de *clustering*) [Bhattacharjee et al., 2001, Beer et al., 2002], mais cela ne permet pas de construire des sous-groupes en fonction du pronostic des patients. À l'inverse, une autre approche consiste à identifier les sous-groupes en se basant exclusivement sur les temps de survie mais sans faire intervenir les covariables [Shipp et al., 2002, Van't Veer et al., 2002], ce qui est d'un intérêt pratique restreint. La méthode que nous proposons, le C-mix, utilise à la fois les covariables et les temps de survie de

façon supervisée dans un contexte de grande dimension. Il s'agit d'un modèle de mélange de $K \geq 1$ lois sur les temps de survie, où les probabilités d'appartenance aux sous-groupes sont déterminées par les covariables.

Il se différencie des modèles de l'état de l'art sous différents aspects. Le modèle le plus utilisé en analyse de survie est le modèle de Cox [Cox, 1972]. Ce dernier a été étendu à la grande dimension [Simon et al., 2011], mais ne permet pas de stratifier la population en sous-groupes à risques homogènes et n'offre pas d'outil simple pour la pratique clinique. Il repose de plus sur l'hypothèse forte des risques proportionnels (voir Section 1.2.3), et s'avère moins performant que le C-mix sur les données simulées et réelles considérées. Des modèles de mélanges de lois des temps de survie ont déjà été considérés [De Angelis et al., 1999, Kuo and Peng, 2000], mais aucun d'entre eux ne s'applique dans un contexte de grande dimension. Un autre exemple classique est le modèle de CURE [Farewell, 1982] qui considère qu'une fraction de la population n'est plus à risque, ce qui est rarement vérifié en pratique.

Le modèle. On introduit une variable latente $Z \in \{0, \dots, K-1\}$ qui modélise l'appartenance à un sous-groupe, et la quantité d'intérêt pour un patient avec des covariables observées $X = x \in \mathbb{R}^d$ est sa probabilité conditionnelle d'appartenance au k-ième groupe, que l'on suppose de la forme

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

et tel que

$$\sum_{k=0}^{K-1} \pi_{\beta_k}(x) = 1.$$

La densité conditionnelle de T sachant $X = x$ est alors donnée par le mélange

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

de K densitées f_k, avec $t \geq 0$, $\alpha_k \in \mathbb{R}^{d_k}$ et $\beta_k \in \mathbb{R}^d$ le vecteur de coefficients qui quantifie l'impact des covariables sur la probabilité d'appartenance au groupe k. On note $\theta = (\alpha_0, \dots, \alpha_{K-1}, \beta_0, \dots, \beta_{K-1})^\top$ le vecteur de paramètres à estimer.

On se place dans le cadre classique de la censure aléatoire de type I rappelée dans la Section 1.2, avec les notations $C \geq 0$ pour la variable de censure, puis Y et Δ le temps censuré et l'indicateur de censure respectivement définis par

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

On considère un échantillon d'apprentissage *i.i.d.* de n patients

$$\mathcal{D}_n = \Big\{(x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n)\Big\},$$

et en supposant de façon classique que T et C sont indépendants sachant Z et X [Klein and Moeschberger, 2005], et que C et Z sont indépendants [Kuo and Peng, 2000], on peut écrire la vraissemblance du modèle

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

où on note F la fonction de répartition correspondant à la densité f, $\bar{F} = 1 - F$ et $F(y^-) = \lim_{\substack{u \to y \\ u \leq y}} F(u)$.

Inférence. La fonction objectif à minimiser lors de l'inférence utilise la pénalité elastic net [Zou and Hastie, 2005] présentée dans la Section 1.1.2. Elle est donnée par

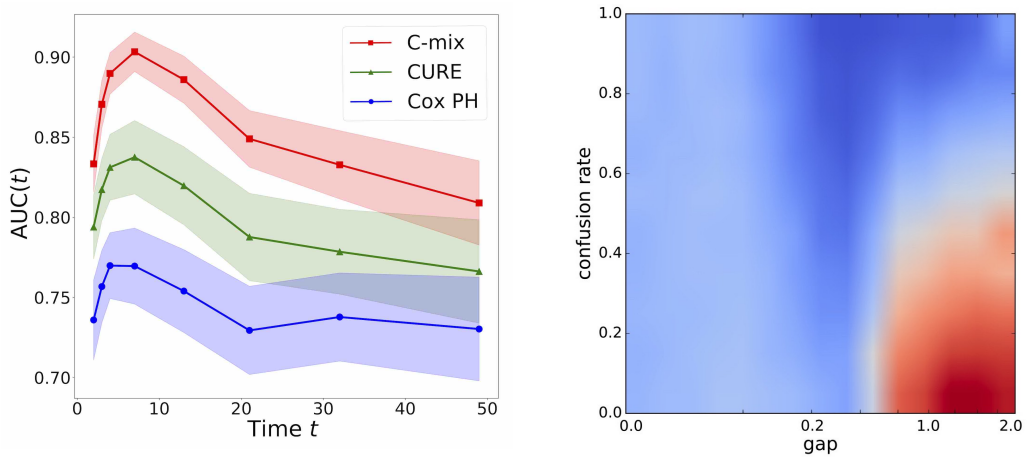
$$\ell_n^{\text{pen}}(\theta) = -\ell_n(\theta) + \sum_{k=0}^{K-1} \gamma_k \Big((1-\eta) \|\beta_k\|_1 + \frac{\eta}{2} \|\beta_k\|_2^2 \Big), \tag{1.17}$$

avec $\eta \in [0,1]$ fixé et $\gamma_k \geq 0$ les hyper-paramètres de régularisation. Nous introduisons alors l'algorithme QNEM (Quasi-Newton Expectation Maximization) qui combine les algorithmes EM [Dempster et al., 1977] et L-BFGS-B [Zhu et al., 1997] afin de minimiser (1.17). Sous certaines hypothèses classiques (concernant la régularité et la convexité des fonctions d'intérêts), le Théorème 4.3.1 établit la convergence de l'algorithme QNEM vers un minimum local de la fonction objectif définie dans (1.17).

Applications. On s'intéresse dans les applications au cas où $K = 2$, avec $Z = 1$ pour les patients ayant un risque élevé de décès rapide et $Z = 0$ pour les autres. Différentes paramétrisation sont essayées pour les densités f_k, et il s'avère que prendre de simples lois géométriques donne les meilleurs résultats. On remarque empiriquement qu'eu égard aux métriques utilisées lors de la validation croisée (pour le choix de l'hyper-paramètre de régularisation) ainsi que pour évaluer les performances – à savoir le C-index [Harrell et al., 1996] et l'AUC(t) [Heagerty et al., 2000], soit des métriques où l'ordre relatif des prédictions importe – il est crucial d'imposer un ordre stochastique entre les fonctions de survie des différents groupes. Cet ordre existe naturellement avec le choix des lois géométriques (deux courbes de survies issues de lois géométriques avec des paramètres distincts ne se croisent pas), et celles-ci donnent de surcroit des mises à jours explicites des paramètres lors de l'étape "M" de l'algorithme QNEM (voir Section 4.3.3).

Le modèle est alors évalué en pratique au travers d'une étude en simulation et comparé avec les modèles de l'état de l'art, à savoir le modèle de Cox [Cox, 1972],

le modèle de CURE [Farewell, 1982] ou encore les modèles de vie accélerée [Kalbfleisch and Prentice, 2011]. Les performances en prédiction sont examinées, mais aussi la stabilité en sélection de variables. Pour ce faire, deux hyper-paramètres sont introduits : un pour régler la sparsité des données et un pour modéliser la présence plus ou moins importante de facteurs de confusion (le "confusion rate"). Un hyper-paramètre est également introduit pour quantifier la "distance" entre les deux groupes au sein des covariables (le "gap"), influant directement sur la difficulté du problème sous-jacent de *clustering* de la population. Les données sont générées suivant les différents modèles en compétition, et le C-mix obtient les meilleurs résultats dans la plus grande partie des nombreuses configurations considérées. La Figure 1.6 donne alors un aperçu des graphiques résultant de l'étude en simulation. Le modèle



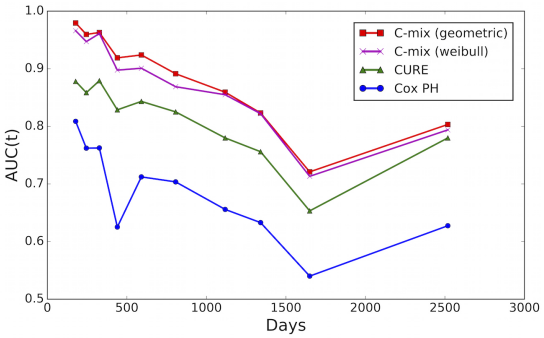
(a) AUC(t)

(b) Sélection de variables

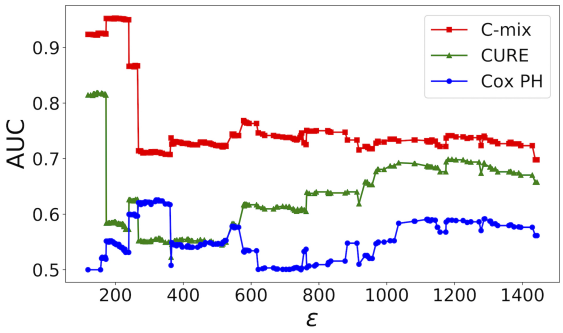
FIGURE 1.6 Échantillon de résultats graphiques provenant de la Section 4.4, avec à droite la Figure 1.6a et les performances des modèles considérés en terme d'AUC(t) sur des données simulées suivant le C-mix ; et à gauche la Figure 1.6b et les performances en sélection de variables du C-mix sur des données simulées suivant le modèle de Cox, pour différentes configurations ("confusion rate", "gap"), où la couleur rouge signifie que le support du "vrai" vecteur de coefficients est parfaitement retrouvé, ce qui est de moins en moins le cas à mesure qu'on tend vers le bleu foncé. Les différentes transitions de phases observées sont interprétées en Section 4.4.4.

est ensuite utilisé sur trois jeux de données publiques de génétique en cancérologie (dont une description est données dans la Section A.2.2) et obtient également de bonnes performances, comme l'illustre la Figure 1.7 qui donne un aperçu des résultats sur le cancer du sein.

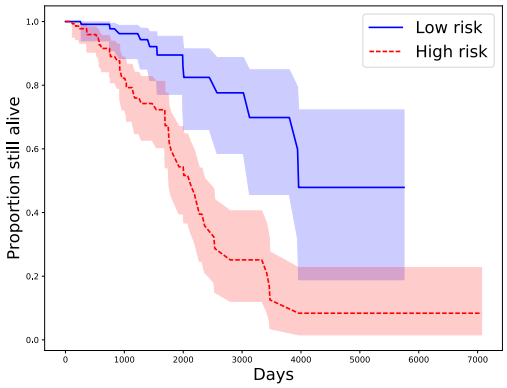
Ainsi, le C-mix surpasse largement les autres modèles considérés en terme de



(a) AUC(t)



(b) AUC



(c) Courbes de survie des deux groupes

FIGURE 1.7 Échantillon de résultats graphiques provenant de la Section 4.5 concernant l'application des modèles considérés sur un jeu de données génétiques de patients atteints d'un cancer du sein. La Figure 1.7a donne les résultats en terme d'AUC(t), la Figure 1.7b en terme d'AUC en utilisant, pour un patient i donné, la fonction de survie estimée et évaluée au temps ϵ – soit $\hat{S}_i(\epsilon|X_i = x_i)$ – pour prédire la quantité binaire $T_i > \epsilon$ pour différents ϵ, et la Figure 1.7c donne les estimateurs de Kaplan-Meier des deux groupes identifiés par le C-mix.

performances prédictives, en sélection de variables et en temps de calcul (voir Section 4.5). Mais surtout, il dispose d'un fort pouvoir d'interprétation en identifiant des sous-groupes de différents pronostics basés sur les covariables : il constitue ainsi un nouvel outil prometteur pour la médecine personnalisée, notamment en cancérologie. Une étude est par ailleurs menée dans la Section 4.G concernant les gènes sélectionnés par les différents modèles dans l'application en cancérologie, et il s'avère que parmi les gènes qui ressortent, de nombreux sont bien connus alors que d'autres méritent sans doute une attention plus poussée.

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. Tous des codes implémentés (principalement en `Python`) pour ce chapitre sont disponibles à l'adresse https://github.com/SimonBussy/C-mix sous la forme de programmes annotés et de tutoriels pour apprendre à utiliser le C-mix en pratique.

Chapitre 5 : Binarsity, a penalization for one-hot encoded features in linear supervised learning

Dans ce chapitre, on se place dans un cadre d'apprentissage supervisé en grande dimension, où on suppose avoir affaire à des covariables continues. Un pré-traitement classique (et souvent nécessaire) consiste à standardiser les covariables. Une autre approche consiste à les discrétiser [Dougherty et al., 1995], par exemple par processus de binarisation appelé encodage "one-hot" évoqué ci-après.

Binarisation. Pour un exemple i, l'idée consiste ici à transformer son vecteur de covariables continues de dimension p en un vecteur de dimension $d \gg p$ donné par

$$x_i^B = (x_{i,1,1}^B, \dots, x_{i,1,d_1}^B, \dots, x_{i,p,1}^B, \dots, x_{i,p,d_p}^B)^\top \in \mathbb{R}^d,$$

où

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

avec $d = \sum_{j=1}^p d_j$ et $I_{j,l} = (\mu_{j,l-1}, \mu_{j,l}]$ où on peut choisir une grille uniforme $\mu_{j,l} = l/d_j$, ou encore les quantiles empiriques $\mu_{j,l} = q_j(l/d_j)$ d'ordre l/d_j pour la j-ième covariable.

Inférence. On se place dans le cadre des modèles linéaires généralisés [Green and Silverman, 1994] où on suppose que la distribution conditionnelle de la sortie sachant les covariables en entrée, c'est-à-dire de $Y|X = x$, appartient à la famille exponentielle à un paramètre, soit de densité donnée par

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

avec les fonctions $b(\cdot)$ et $c(\cdot)$ connues, et $m^0(\cdot)$ la fonction inconnue que l'on cherche à estimer. On considère alors le risque empirique

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, m_\theta(X_i)\Big),$$

où $m_\theta(X_i) = \theta^\top X_i^B$ et $\theta \in \mathbb{R}^d$ le paramètre à estimer, avec la fonction de perte $\ell(y_1, y_2) = -y_1 y_2 + b(y_2)$. Le problème de minimisation du risque empirique pénalisé que l'on considère est le suivant

$$\hat{\theta} \in \operatorname{argmin}_{\theta\in\mathbb{R}^d} \Big\{R_n(\theta) + \mathrm{bina}(\theta)\Big\},$$

avec

$$\mathrm{bina}(\theta) = \sum_{j=1}^{p}\Big(\sum_{k=2}^{d_j} \hat{w}_{j,k}|\theta_{j,k} - \theta_{j,k-1}| + \delta_1(\theta_{j,\bullet})\Big)$$

la pénalité appelée *binarsity* permettant de faire face aux problèmes de conditionnement de la matrice binarisée d'une part, et de sélection de variables en grande dimension d'autre part (voir Section 5.2) ; où les poids $\hat{w}_{j,k} > 0$ sont définis dans la Section 5.3, et où

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

Résultat. On note

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

la fonction de risque associée [van de Geer, 2008], et $J(\theta) = \Big[J_1(\theta), \ldots, J_p(\theta)\Big]$ la concaténation des supports relativement à la pénalité par variation totale (TV), à savoir $J_j(\theta) = \Big\{k : \theta_{j,k} \neq \theta_{j,k-1}, \text{ for } k = 2, \ldots, d_j\Big\}$. En notant maintenant

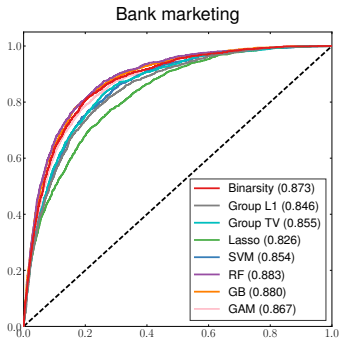
$$B_d(\rho) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq \rho\}$$

la boule de norme ℓ_2 de rayon $\rho > 0$ dans $\mathbb{R}^d$, le Théorème 5.3.1 établit alors l'inégalité oracle non asymptotique à vitesse rapide qui est synthétisée par l'équation suivante

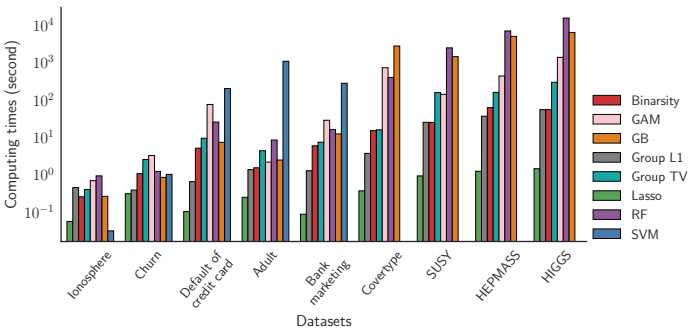
$$R(m_{\hat{\theta}}) - R(m^0) \leq (1 + c_1) \inf_{\substack{\theta\in B_d(\rho) \\ \forall j,\ \mathbf{1}^\top\theta_{j,\bullet}=0 \\ |J(\theta)|\leq J^*}} \left\{R(m_\theta) - R(m^0) + c_2\frac{|J(\theta)|\log d}{n}\right\},$$

qui est vraie avec grande probabilité, avec c_1 et c_2 deux constantes positives (c_2 provenant d'une condition aux valeurs propres restreintes sur la matrice binarisée). Dans le cas de la régression aux moindres carrés, l'inégalité est *sharp*, soit $c_1 = 0$.

Applications. La méthode est est utilisée dans le cadre d'une régression logistique sur 9 jeux de données standards de classification binaire [Lichman, 2013] et comparée avec les méthodes de l'état de l'art suivantes : la régression logistique avec une pénalité lasso [Tibshirani, 1996], group lasso [Meier et al., 2008] ou group TV, les machines à vecteurs de support (SVM) [Schölkopf and Smola, 2002], les modèles additifs généralisés (GAM) [Hastie and Tibshirani, 1990], les forêts aléatoires (RF) [Breiman, 2001] ainsi que le gradient boosting [Friedman, 2002]. La Figure 1.8 donne un aperçu des résultats obtenus dans la Section 5.4. La pénalité binarsity surpasse les performances prédictives de celles des pénalités lasso, group lasso, group TV ainsi que celles des modèles linéaires généralisés, et est compétitive avec les forêts aléatoires et le boosting, avec des temps de calcul nettement inférieurs. Et surtout, elle offre une interprétabilité puissante, qui sera d'ailleurs au cœur du chapitre suivant : en plus de la sélection de variables, la méthode identifie des seuils significatifs dans les covariables continues initiales (aux positions des sauts du vecteur de régression, voir Figure 5.1), ce qui procure une compréhension plus précise et profonde du modèle que celle fournie par le lasso.



(a) Courbes ROC des méthodes considérées sur un des jeux de données.



(b) Comparaison des temps de calcul des différentes méthodes sur les jeux de données considérés.

FIGURE 1.8 Échantillon de résultats graphiques provenant de la Section 4.5.

Article associé.
M.Z. Alaya, **S. Bussy**, S. Gaïffas, A. Guilloux
Binarsity : a penalization for one-hot encoded features
Accepté pour publication et en révision mineure dans *Journal of Machine Learning Research*, 2018.

Code. Tous des codes implémentés (principalement en `Python/C++`) pour ce chapitre sont disponibles à l'adresse `https://github.com/SimonBussy/binarsity` sous

la forme de programmes annotés et de tutoriels pour apprendre à utiliser la pénalité binarsity en pratique.

Chapitre 6 : Binacox, automatic cut-points detection in a high-dimensional Cox model

Contexte. Traduire en décision clinique les valeurs prises par des biomarqueurs requiert souvent de choisir des seuils. Nous introduisons dans ce chapitre une méthode pronostique appelée *binacox* afin de traiter le problème de détection de multiples seuils par covariable continue de façon multivariée, dans un contexte de grande dimension. On note T le temps de décès, C le temps de censure, $X \in \mathbb{R}^p$ le vecteur de covariables, $Z = T \wedge C$ le temps censuré à droite et $\Delta = \mathbb{1}(\{T \leq C\})$ l'indicateur de censure. On suppose alors que le risque instantané pour un patient i est donné par

$$\lambda^\star(t|X_i = x_i) = \lambda_0^\star(t) e^{f^\star(x_i)},$$

avec $\lambda_0^\star(t)$ le risque de base, et

$$f^\star(x_i) = \sum_{j=1}^{p} \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \mathbb{1}(x_{i,j} \in I_{j,k}^\star),$$

où $I_{j,k}^\star = (\mu_{j,k-1}^\star, \mu_{j,k}^\star]$ pour $k \in \{1, \dots, K_j^\star + 1\}$. Le but est d'estimer simultanément le vecteur de seuils

$$\mu^\star = ({\mu_{1,\bullet}^\star}^\top, \dots, {\mu_{p,\bullet}^\star}^\top)^\top = (\mu_{1,1}^\star, \dots, \mu_{1,K_1^\star}^\star, \dots, \mu_{p,1}^\star, \dots, \mu_{p,K_p^\star}^\star)^\top \in \mathbb{R}^{K^\star}$$

et le vecteur de coefficients correspondant

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

en notant $K^\star = \sum_{j=1}^p K_j^\star$ le nombre total de seuils.

Méthode. On commence par utiliser la même méthode de binarisation que dans le chapitre précédent. On obtient ainsi

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

où

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

et $I_{j,l} = (\mu_{j,l-1}, \mu_{j,l}]$ avec un choix de grille uniforme par exemple, soit $\mu_{j,l} = l/(d_j + 1)$. On définit alors

$$f_\beta(x_i) = \beta^\top x_i^B = \sum_{j=1}^{p} f_{\beta_{j,\bullet}}(x_i)$$

où

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

pour tout $j \in \{1, \ldots, p\}$, de telle sorte que f_β est un estimateur de $f^\star = f_{\beta^\star}$. On définit alors la log-vraissemblance négative partielle binarisée (normalisée par n^{-1}) comme

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

et on considère le problème d'optimisation suivant

$$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathscr{B}_{p+d}(R)} \Big\{ \ell_n(f_\beta) + \operatorname{bina}(\beta) \Big\}$$

avec

$$\mathscr{B}_{p+d}(R) = \{\beta \in \mathbb{R}^{p+d} : \|\beta\|_2 \leq R\}$$

la boule de norme ℓ_2 de rayon $R > 0$ dans $\mathbb{R}^{p+d}$ et

$$\operatorname{bina}(\beta) = \sum_{j=1}^{p} \Big(\sum_{l=2}^{d_j+1} \omega_{j,l} |\beta_{j,l} - \beta_{j,l-1}| + \delta_1(\beta_{j,\bullet}) \Big),$$

avec

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

et où les poids sont ici de telle sorte que

$$\omega_{j,l} = \mathcal{O}\left(\sqrt{\frac{\log(p+d)}{n}} \right),$$

voir Section 6.B.1 pour plus de détails. En notant alors

$$\mathcal{A}_j(\hat{\beta}) = \Big\{ l : \beta_{j,l} \neq \beta_{j,l-1}, \text{ for } l = 2, \ldots, d_j + 1 \Big\} = \{\hat{l}_{j,1}, \ldots, \hat{l}_{j,s_j}\},$$

on obtient l'estimateur

$$\widehat{\mu}_{j,\bullet} = (\mu_{j,\hat{l}_{j,1}}, \ldots, \mu_{j,\hat{l}_{j,s_j}})^\top,$$

avec $s_j = |\mathcal{A}_j(\hat{\beta})| = \widehat{K}_j$.

Résultats. On définit la divergence de Kullback-Leibler empirique entre $f^\star$ et une fonction candidate f comme

$$KL_n(f^\star, f) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left\{\frac{e^{f^\star(X_i)}\sum_{i=1}^n Y_i(t)e^{f(X_i)}}{e^{f(X_i)}\sum_{i=1}^n Y_i(t)e^{f^\star(X_i)}}\right\}Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}\mathrm{d}t,$$

où $Y_i(t) = \mathbb{1}(Z_i \geq t)$ est le processus modélisant la présence de risque. Le Théorème 6.3.1 établit alors une inégalité oracle non-asymptotique à vitesse rapide qui est synthétisée par l'équation suivante

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

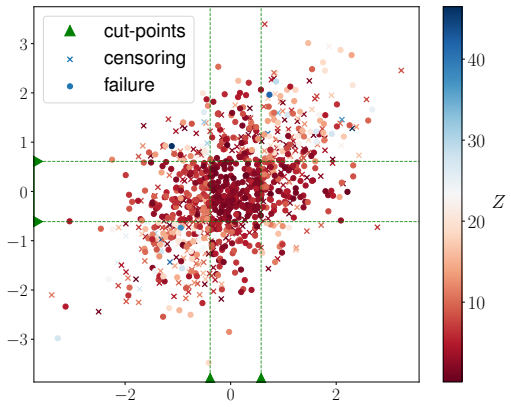
qui est vraie avec grande probabilité, avec c_1 et c_2 deux constantes positives (et c_2 qui dépend d'un facteur de compatibilité d'une certaine matrice, voir Section 5.3 pour plus de détails).

Le modèle est alors évalué en pratique au travers d'une étude en simulation et comparé avec les méthodes de l'état de l'art pour la détection de seuils en analyse de survie, à savoir des méthodes basées sur des tests du logrank multiples [Budczies et al., 2012]. Les performances en détection sont examinées, mais aussi la stabilité en sélection de variables. La Figure 1.9 donne un aperçu des résultats obtenus dans la Section 6.4 concernant l'étude en simulation. La méthode binacox surpasse alors largement les méthodes univariées existantes en terme de détection, et davantage encore en terme de temps de calcul.

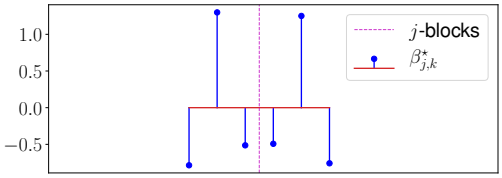
Les méthodes considées sont ensuite appliquées sur trois jeux de données publiques de génétique en cancérologie (dont une description est données dans la Section A.2.2) et le binacox obtient de bonnes performances, en détectant des seuils intéressants pour certains gènes connus (voir la Section 6.5). La Figure 1.10 donne un apperçu des résultats graphiques obtenus.

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
Journal visé : *Journal of the Royal Statistical Society : Series B.*

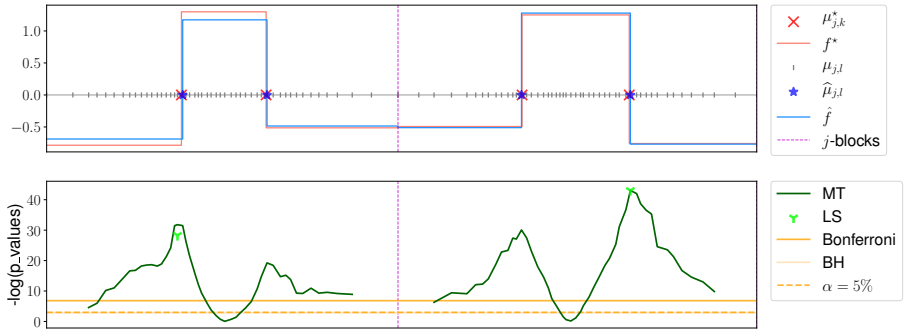
Code. Tous des codes implémentés (principalement en `Python/C++`) pour ce chapitre sont disponibles à l'adresse `https://github.com/SimonBussy/binacox` sous la forme de programmes annotés et de tutoriels pour apprendre à utiliser la méthode en pratique.



(a) Données générées suivant le modèle binacox.

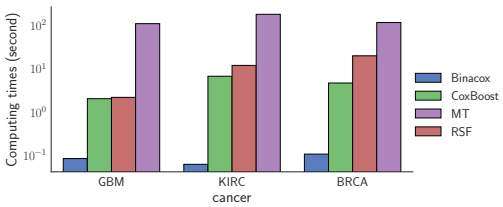


(b) Illustration du $\beta^\star$ utilisé pour obtenir la Figure 1.9a.

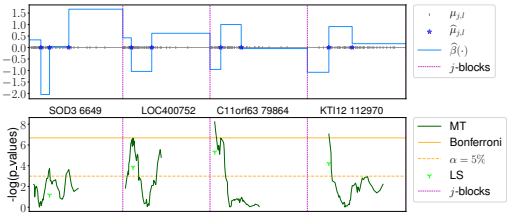


(c) Estimation des différentes méthodes considérées sur les données de la Figure 1.9a.

FIGURE 1.9 Échantillon de résultats graphiques provenant de la Section 6.4.



(a) Comparaison des temps de calcul des méthodes considérées sur les 3 cancers.



(b) Résultats en détection des différentes méthodes sur le cancer GBM.

FIGURE 1.10 Échantillon de résultats graphiques provenant de la Section 6.5, à consulter pour plus de détails.

* * *

Note. All chapters are self-contained and can be read independently. All notations used within a given chapter are defined in the chapter, and some notations may differ between chapters for convenience reasons (for instance, Z denotes a latent variable in Chapter 4, while it denotes the censored times in Chapter 6).

Les différents chapitres peuvent être lus indépendamment les uns des autres. L'ensemble des notations utilisées dans un chapitre y sont précisément définies, et certaines notations peuvent différer d'un chapitre à l'autre pour des raisons de commodité d'écriture (par exemple, Z représente une variable aléatoire latente dans le Chapitre 4, alors qu'il s'agit de la variable aléatoire des temps censurés dans le Chapitre 6).

Chapitre 2

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

Sommaire

Abstract. Hospital admission of sickle-cell disease (SCD) patients presenting a non-complicated vaso-occlusive crisis (VOC) is justified both by refractory pain to ambulatory available drugs and by possible complications, such as infections or acute chest syndrome. But, the "normal" range of values for vital parameters and usual laboratory results, both at the diagnosis and throughout the non-complicated VOC, have yet to be established, in order to detect such complications and/or differential diagnosis. In this observatory retrospective study, we describe the behavior of biomarkers and vital parameters throughout a non-complicated VOC hospital stay. We included 329 non-complicated VOC-related stays of 164 SCD patients, which took place between 2010 and 2015 in the internal medicine department of the GPUH in Paris, France. We identified several relevant biomarkers and vital parameters when monitoring a VOC episode, namely hemoglobin, leucocytes (and more specifically eosinophils), CRP and temperature. Regarding hemoglobin, it rarely returned to baseline levels before discharge, but a drop during the stay may be predictive of early readmission. Leucocytes trajectories showed specific trends, especially for neutrophils and eosinophils, that could help further assess the diagnosis of VOC. Under/above threshold proportion analysis showed that over 95% of non-complicated VOC stays displayed a baseline CRP value of under 100 mg/L within the first day following admission, and had no fever throughout the entire stay, suggesting that fever and/or CRP over 100 mg/L at admission should not be attributed to the VOC itself.

Résumé. L'admission à l'hôpital pour les patients atteints de drépanocytose lors d'une crise vaso-occlusive (COV) non compliquée est justifiée par des douleurs réfractaires aux médicaments disponibles en ambulatoire et par d'éventuelles complications, comme des infections ou un syndrome thoracique aigu. Cependant, l'évolution "normale" des valeurs pour les paramètres vitaux et les résultats de laboratoire habituels, à la fois lors du diagnostique et au cours de la COV non compliquée, n'est pas encore suffisamment établi pour pouvoir détecter sereinement de telles complications. Dans cette étude rétrospective, nous décrivons le comportement des biomarqueurs et des paramètres vitaux lors d'un séjour hospitalier non compliqué pour CVO. 329 séjours sont inclus correspondant à 164 patients drépanocytaires entre 2010 et 2015 admis dans le département de médecine interne de l'Hopitâl Universitaire George Pompidou à Paris. Nous avons identifié plusieurs biomarqueurs et paramètres vitaux pertinents lors de la surveillance d'un épisode de COV, à savoir l'hémoglobine, les leucocytes (et plus précisément les éosinophiles), la CRP et la température. En ce qui concerne l'hémoglobine, elle est rarement revenue au niveau de base avant la sortie de l'hôpital, mais une baisse au cours du séjour peut prédire une réadmission précoce. Les trajectoires des leucocytes ont montré des tendances spécifiques, en particulier pour les neutrophiles et les éosinophiles, qui pourraient aider à mieux évaluer le diagnostic de COV. Une analyse des proportions inférieures

ou supérieures à certains seuils a montré que plus de 95% des séjours pour COV non compliquée présentaient une valeur de référence de la CRP inférieure à 100 mg/L le premier jour après l'admission, et pas de fièvre pendant toute la durée du séjour, ce qui suggère que la fièvre et/ou qu'une CRP supérieure à 100 mg/L à l'admission ne doivent pas être attribuées à la COV elle-même.

2.1 Introduction

Background. Sickle-cell disease (SCD) is the most common monogenic disorder worldwide, resulting from a range of recessive mutations. Over 5% of the world population is carrying a variant, and 2.55 births out of 1000 are affected with the disease [Modell and Darlison, 2008]. Although multiple genotypes can lead to the phenotypical trait, homozygote S-hemoglobin, otherwise known as sickle-cell anemia, is accountable for roughly 70% of cases [Rees et al., 2010].

The inherited mutated variants lead to a defective β-hemoglobin sub-unit, which predisposes the sickling of erythrocytes [Pauling et al., 1949, Bunn, 1997, Stuart and Nagel, 2004]. Consequently, these misshaped and rigid red blood cells will, under certain conditions, obstruct capillaries, thus inducing acute ischemia to downstream organs and tissues [Bunn, 1997, Stuart and Nagel, 2004].

Such episodes, called vaso-occlusive crisis (VOC), are responsible for acute pain syndromes and ultimately result in increased morbidity and mortality [Diggs, 1965, Platt et al., 1991, 1994, Prasad et al., 2003, Darbari et al., 2013, Vichinsky et al., 2000]. SCD also results in chronic hemolysis leading to vasculopathy and ultimately to organ damages, as well as an increased susceptibility to infection due to functional asplenia.

Rationale. Hospital admission for non-complicated VOC is justified both by refractory pain to ambulatory available drugs and by possible complications, such as infections or acute chest syndrome. The diagnosis of VOC relies on the occurrence of an acute bone pain that is usually located on one or several limbs or the spine and is usually straightforward. However, it is impossible to differentiate VOC from pain of other origin, especially from opioid addiction, which is a possible complication of recurrent pain killer use.

There is currently no reliable diagnostic biomarker of ongoing VOC, although it has been observed that VOC are often, but not always, associated with increased hemolysis (elevated lactate dehydrogenase (LDH) and bilirubin), increased anemia (low hemoglobin level) and moderate systemic inflammation (elevated C-Reactive Protein (CRP) and hyperleukocytosis). These common inflammatory biomarkers are monitored to detect the occurrence of an infectious complication during VOC, but the "normal" range of values for vital parameters and usual laboratory results at the diagnosis and during the evolution of non-complicated VOC are unknown.

In hospitalized VOC episodes, opioid-based patient-controlled analgesia (PCA) is a first choice treatment to control pain ; therefore the resolution of the VOC is corroborated by the decrease of both pain intensity and opioid requested doses. In this context, establishing the common range of vital parameters and usual laboratory results at admission and during the course of a non-complicated VOC episode could help the clinician either infirm the diagnosis of VOC, or detect the presence of a

possible complication. As a result, it could be a first step towards a specific guideline for monitoring VOC episode, both in term of which laboratory tests to perform, as well as when to perform them. Additionally, combining VOC monitoring data with readmission delays could help assess the predictability of early readmissions, which is also a challenge in the management of VOC episodes [Frei-Jones et al., 2009, Brousseau et al., 2010].

Today, clinical data warehouses (CDW) can automatically store heterogeneous real-life data from electronic health records (EHR), allowing researchers to use clinical, administrative, and biological retrospective data to answer this kind of questions. The Georges Pompidou University Hospital (GPUH), in Paris, France, set up such a data warehouse [Zapletal et al., 2010]. The GPUH internal medicine department is also a center of expertise for SCD adult patients, and manages over 150 VOC-related hospitalizations per year [bnd]. Reuse of heath data from this department, facilitated by the CDW, could help describe how biomarkers and vital signs behave throughout a hospitalized non-complicated VOC.

Objectives. The main objective for this study was to describe the behavior of biomarkers and vital parameters throughout a non-complicated VOC hospital stay. The secondary objective was to identify which biomarker(s) and/or vital sign(s) should be monitored in the days following a hospital admission for VOC in order to help identifying stays with high risk of early readmissions after hospital discharge.

2.2 Method

2.2.1 Study design

This is a monocentric retrospective cohort study. Data was extracted from de GPUH CDW which uses the i2b2 star-shaped standard [Uzuner et al., 2011, Murphy et al., 2010]. It contains routine care data divided into several categories, including demographics (*e.g.* date of birth, sex, etc.), vital signs (*e.g.* blood pressure, temperature, etc.), diagnoses (ICD-10), procedures (French CCAM classification), clinical data (structured questionnaires from EHR), free text reports, biological test results (LOINC), and Computerized Provider Order Entry (CPOE) drug prescriptions.

2.2.2 Setting and studied population

The sample included all stays from patients admitted to the internal medicine department for VOC (ICD-10 D57.0) between January 1st 2010 and December 31st 2015. We excluded patients encoded as opioid addicts (ICD-10 F11) as well as those

who were treated with either Methadone or Buprenorphine, both confirmed by hospitalization reports and drug prescriptions. We also excluded complicated VOC, defined as stays for which :

— The patient stayed in the ICU at any point during the hospitalization.

— The stay's severity was rated as 3 or 4 on the 4-level CMA ("Complications ou Morbidité Associée") scale of severity (see Section 2.A for details).

— The patient received at least one red blood cell transfusion during his stay

— The stay was associated with a diagnosis of complication (*e.g.* bacterial infection, thrombosis, gout, etc. See Section 2.B for details. Acute chest syndromes where excluded *de facto* by previous criteria).

— The stay duration was higher than 90^{th} percentile of the duration of remaining stays after applying all of the excluding criteria listed above.

To ensure complete reporting of our routinely collected health data, we followed the REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement [Benchimol et al., 2015].

This study received approval the Institutional review board from Georges Pompidou University Hospital (IRB 00001072 - project n°CDW_2014_0008) and the French data protection authority (CNIL - n° 1922081).

2.2.3 Covariates

The following covariates were extracted :

— Demographic data (sex, date of birth).

— Dates and timestamps for hospital admission (which occurs after the patient has been thoroughly evaluated and is considered stable by the emergency department), as well as hospital discharge.

— Laboratory values measured at least once for over 75% of patients from both the emergency department (prior to hospital admission) and the subsequent hospital stay.

— Vital parameters measured at least once for over 75% of patients after admission (vital parameters from the emergency department could not be retrieved).

— Opioid prescriptions (*e.g.* molecule, starting and ending timestamps, pharmaceutical form, etc.).

— Baseline hemoglobin level and comorbidities from every available free text reports from the patients' EHR regardless of the source department and the stay.

In order to facilitate facilitate the extraction of variables from such textual reports, we used a locally developed browser-accessible tool called FASTVISU [Escudié et al., 2015]. This software is linked with the CDW, and allowed us to rapidly check throughout these textual reports for highlighted information and to vote for baseline hemoglobin level value or comorbidities. Key words using regular expressions are used to focus on specific mentions within the text (*e.g.* "h.m1,2oglobine" would highlight the text parts containing the French word "hémoglobine", even with common misspelling).

2.2.4 Data derivation

We derived new covariates as follows :

— The age of the patient at hospital admission, determined from both the patient's date of birth and the timestamp for the admission.

— The duration of the stay, from the admission and discharge timestamps. Note that the duration of the stay does not include the time the patient spent in the emergency department, prior to admission.

— The post-opioid observation period, from the opioid prescription ending timestamps and the patient's discharge timestamp.

— The stay's readmission delay (*i.e.* the length of time between discharge and the next readmission), determined from admission and discharge timestamps (including for VOC stays which were excluded from this study), which we then dichotomized on the commonly accepted 30 day post-discharge readmission threshold [Frei-Jones et al., 2009, Brousseau et al., 2010].

— The hemoglobin gap to baseline (*i.e.* the difference between each punctual hemoglobin measure and the patient's baseline hemoglobin value).

2.2.5 Missing data imputation

We imputed the missing variables as follows :

— For the patients' baseline hemoglobin value, we imputed with the last hemoglobin value measured prior to discharge, from the first included stay of the patient.

— For the post-opioid observation period, the data was obviously missing for patients who didn't receive any opioids in the internal medicine department. Knowing that the emergency department follows a specific protocol for VOC, which systematically includes the use of opioid drugs at admission, we assumed that every patient admitted for VOC received opioids at admission. Thus, we imputed this derived variable with the total duration of the stay.

2.2.6 Statistical methods

For descriptive purposes, we calculated the overall statistics among patients (gender, baseline hemoglobin, age and medical history at the first included stay). We also described stay-specific statistics (*e.g.* stay duration, post-opioid observation period, and the molecule and pharmaceutical form of the opioid prescription during the stay), as well as the number of included stays per patient.

Additionally, we grouped patients by genotype and performed univariate testing for differences between the groups : we used the Chi-squared test (or Fisher's exact test) for categorical variables and Wilcoxon's sum-rank test for quantitative variables.

Regarding selected biomarkers and vital parameters, because in routine care they are measured for each patient on a specific pattern, described each variable's mean trajectory and its confidence interval (see Section 2.C for details). For each parameter, we also generated a point cloud of all measures and plotted each individual stay's raw trajectory for descriptive purposes.

2.3 Results

2.3.1 Descriptive statistics

As precised in Figure 2.2, 329 hospital stays for non-complicated VOC were included for a total of 164 patients with an average of 2 stays per patient, varying from 1 to 10, see Figure 2.1a. The duration of the stay ranges from 1 to 10 days, with an average of 4.4 days as shows Figure 2.1b.

Patients statistics, presented in Table 2.1, at first included admission show some differences between the sickle-cell anemia (SCA) group vs. other genotypes : sickle-cell anemia was associated with younger age at admission (28 vs. 33 years old), lower baseline hemoglobin level (8.5 vs. 10.3 g/dl), more frequent medical history of acute chest syndrome (77 vs. 42%) and priapism (22 vs. 0% of men).

On the other hand, sickle-cell anemia seemed associated with less frequent medical history of avascular bone necrosis (19 vs. 33%), without reaching significance threshold. Stays statistics show no difference between the early readmission group (less than 30 days after discharge) vs. the late or no readmission group, see Table 2.2.

2.3.2 Laboratory values trends

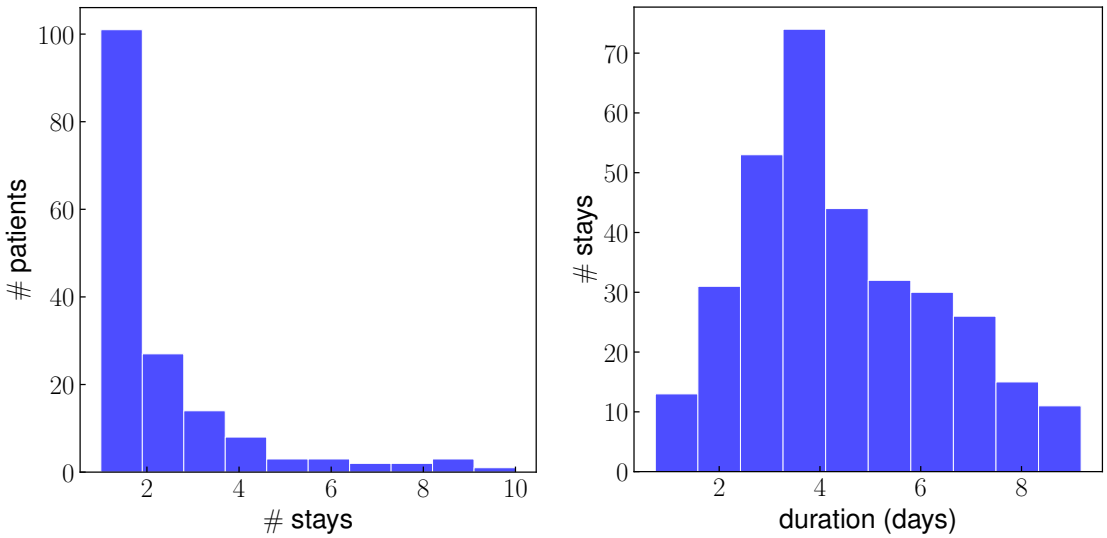
Figure 2.3 shows hemoglobin point cloud with a high concentration of measures performed prior to hospital admission, while the patients are still in the emergency department. Hemoglobin values range from 4 to 16 g/dL over the stay and slowly

TABLE 2.1 Patients statistics for basic covariates. The p-values correspond to univariate testing for differences between the groups based on each modalities. n is the number of patients.

Covariate (p-value)	**Modality**	**Whole sample** $n = 164$ (100%)	**SCA** $n = 121$ (73.78%)	**Non-SCA** $n = 43$ (26.22%)
Gender (0.724)	Female	87 (53.05%)	63 (52.07%)	24 (55.81%)
	Male	77 (46.95%)	58 (47.93%)	19 (44.19%)
Age at first hospital admission (0.041)	Mean (sd)	29.42 (10.24)	28.2 (9.27)	32.85 (12.03)
	Median [Q1 ; Q3]	27.5 [21.8 ; 34.5]	26.5 [21.9 ; 31.9]	33.1 [20.9 ; 38.8]
	Min ; Max	16.35 ; 83.08	6.35 ; 83.08	17.29 ; 57.13
Baseline haemoglobin in g/dL ($<$ 0.001)	Mean (sd	8.98 (1.59)	8.49 (1.3)	10.35 (1.52)
	Median [Q1 ; Q3]	9 [8 ; 10]	8 [8 ; 9]	10 [9.5 ; 11]
	Min ; Max	5 ; 13.5	5 ; 13.4	7 ; 13.5
Household situation (0.019)	Lives by oneself	98 (80.99%)	76 (86.36%)	22 (66.67%)
	Lives with others	23 (19.01%)	12 (13.64%)	11 (33.33%)
	NA	43	33	10
Professional activity (1)	Active	123 (86.62%)	89 (86.41%)	34 (87.18%)
	Inactive	19 (13.38%)	14 (13.59%)	5 (12.82%)
	NA	22	18	4
Acute chest syndrom syndrom ($<$ 0.001)	Present	111 (67.68%)	93 (76.86%)	18 (41.86%)
	Absent	53 (32.32%)	28 (23.14%)	25 (58.14%)
Avascular bone necrosis (0.089)	Present	37 (22.56%)	23 (19.01%)	14 (32.56%)
	Absent	127 (77.44%)	98 (80.99%)	29 (67.44%)
Dialysis (0.457)	Present	2 (1.22%)	1 (0.83%)	1 (2.33%)
	Absent	162 (98.78%)	120 (99.17%)	42 (97.67%)
Heart failure (1)	Present	0 (0%)	0 (0%)	0 (0%)
	Absent	164 (100%)	121 (100%)	43 (100%)
Ischemic stroke (0.186)	Present	6 (3.66%)	3 (2.48%)	3 (6.98%)
	Absent	158 (96.34%)	118 (97.52%)	40 (93.02%)
Leg skin ulceration (0.457)	Present	10 (6.1%)	9 (7.44%)	1 (2.33%)
	Absent	154 (93.9%)	112 (92.56%)	42 (97.67%)
Known nephropathy (0.433)	Present	8 (4.88%)	5 (4.13%)	3 (6.98%)
	Absent	156 (95.12%)	116 (95.87%)	40 (93.02%)
Pulmonary hypertension (1)	Present	3 (1.83%)	2 (1.65%)	1 (2.33%)
	Absent	161 (98.17%)	119 (98.35%)	42 (97.67%)
Known retinopathy (0.275)	Present	19 (11.59%)	12 (9.92%)	7 (16.28%)
	Absent	145 (88.41%)	109 (90.08%)	36 (83.72%)
***we consider only male patients**		$n = 77$ (100%)	$n = 58$ (75.32%)	$n = 19$ (24.68%)
Priapism (0.030)	Present	13 (16.88%)	13 (22.41%)	0 (0%)
	Absent	64 (83.12%)	45 (77.59%)	19 (100%)

TABLE 2.2 Basic stays statistics. The p-values correspond to univariate testing for differences between the groups based on each modalities. N is the number of stays.

Covariate (p-value)	**Modality**	**Whole sample** $N = 329$ (100%)	**Early readmission** (≤ 30 **days)** $N = 49$ (14.89%)	**Late or no readmission** $N = 280$ (85.11%)
Length of hospital in days (0.553)	Mean (sd)	4.44 (1.92)	4.22 (1.55)	4.48 (1.98)
	Median [Q1 ; Q3]	4.01 [2.91 ; 5.78]	3.99 [3.04 ; 5.29]	4.05 [2.88 ; 5.9]
	Min ; Max	0.73 ; 9.18	0.92 ; 7.9	0.73 ; 9.18
Post-opioid observation period in hours (0.282)	Mean (sd)	15.66 (37.38)	12.19 (28.79)	16.26 (38.69)
	Median [Q1 ; Q3]	1.15 [0.74 ; 1.91]	1.11 [0.74 ; 1.45]	1.15 [0.76 ; 1.92]
	Min ; Max	0 ; 212.42	0 ; 103.98	0.12 ; 212.42
Received orally administered opioids (0.239)	Yes	14 (4.26%)	0 (0%)	14 (5%)
	No	315 (95.74%)	49 (100%)	266 (95%)
Received Oxycodone (1)	Yes	7 (2.55%)	1 (2.38%)	6 (2.58%)
	No	268 (97.45%)	41 (97.62%)	227 (97.42%)
	NA	54	7	47



(a) Repartition of the number of stays per partient (# for "number of").

(b) Histogram for the duration in days (# for "number of").

FIGURE 2.1 Basic data description.

decrease on average from 9.5 down to 8.5 g/dL over the first week after hospital admission, see Figure 2.4. Females seem to have lower hemoglobin values throughout their stay than males, see Figure 2.5a. Stays followed by an early readmission after discharge also show a clear decreasing trend around the 5^{th} day whereas other stays remain stable, see Figure 2.5b. Hematocrit values trend is similar.

White blood cell count shows an early spike over $12{\times}10^9/L$, around admission, and then decreases and stabilizes around $10{\times}10^9/L$ after the second day of the stay, see Figure 2.6. Average neutrophil count shows a similar trend in Figure 2.7, whereas average lymphocyte, monocyte, and basophil counts show no clear trend throughout the stay. Average eosinophil count increases from $2{\times}10^8/L$ at admission to $4{\times}10^8/L$ around the 5^{th} day of the stay, see Figure 2.8.

Figure 2.9 shows that individual platelets trajectories mostly remain stable throughout the stay, with an average value between $300{\times}10^9/L$ and $350{\times}10^9/L$, see Figure 2.10. Average CRP rapidly increases in the first 2 days after admission, and then stabilizes around 60 mg/L before slowly decreasing, see Figure 2.11.

It is worth noting that after the second day following hospital admission, less than 5% of VOC show a normal CRP level, see Figure 2.12a. Additionally, within the first 24 hours after admission, less than 5% of patients reach the 100 mg/L CRP threshold, see Figure 2.12b. Most individual LDH trajectories remain stable throughout the stay as shown in Figure 2.13, with an average value between 400 U/L and 450 U/L, see Figure 2.14. No clear trend emerges for electrolytes, liver

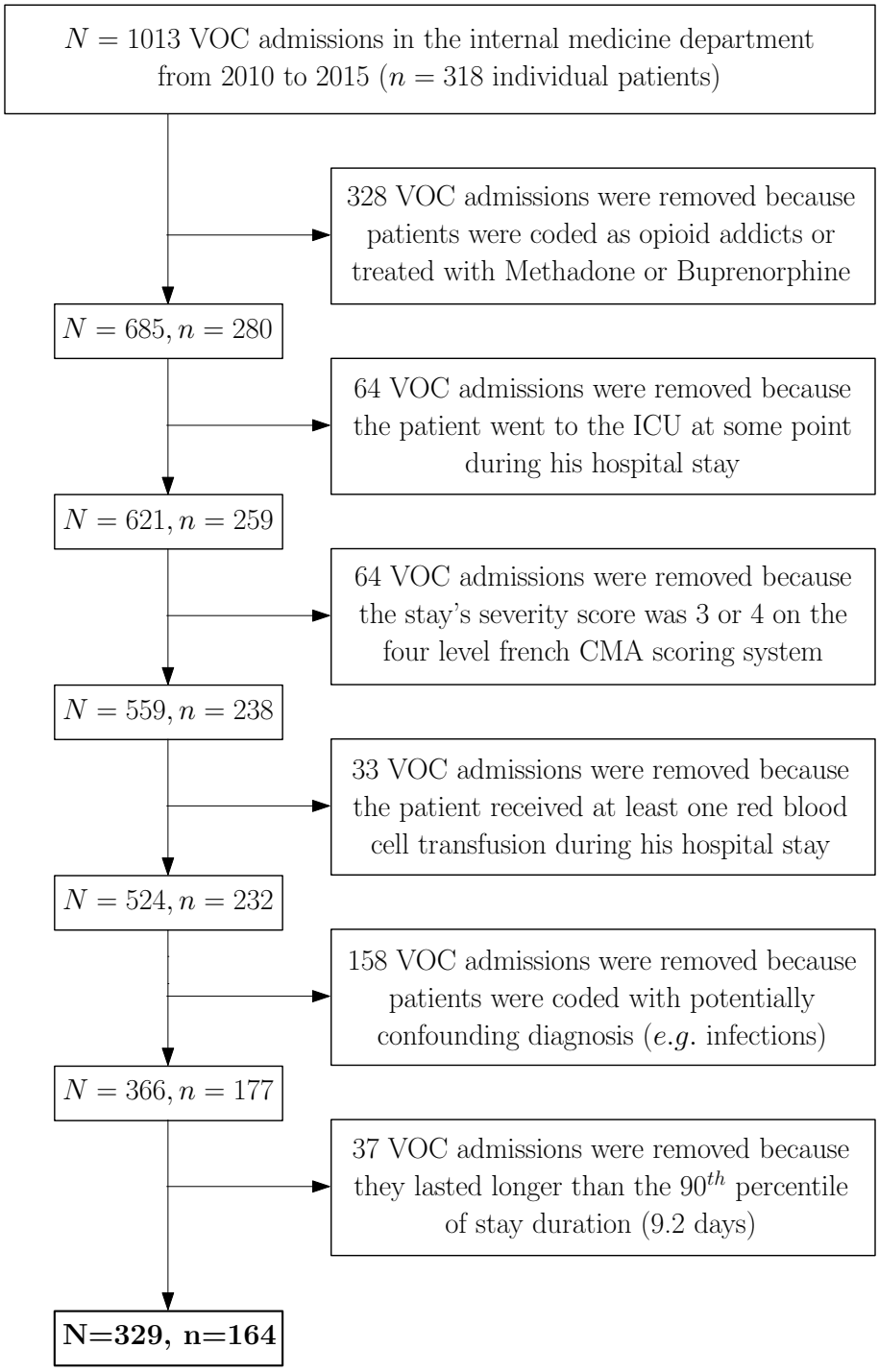


FIGURE 2.2 Illustration of the different steps followed in the patients selection phase. n is the number of patients and N the number of stays.

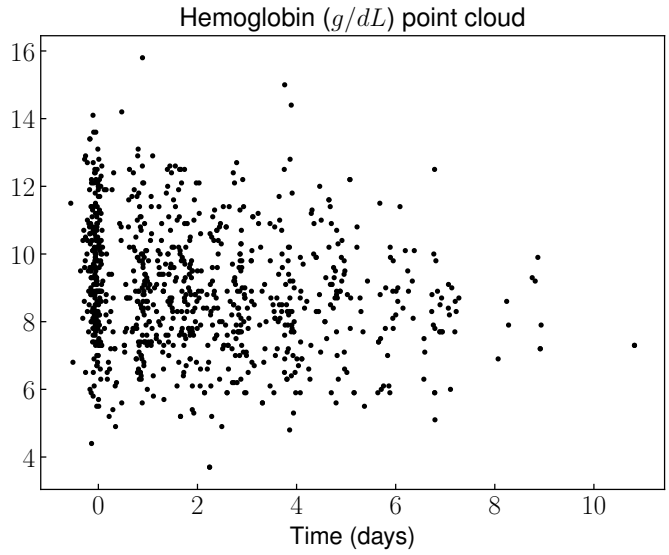


FIGURE 2.3 Hemoglobin point cloud (in g/dL) with all points of all patients.

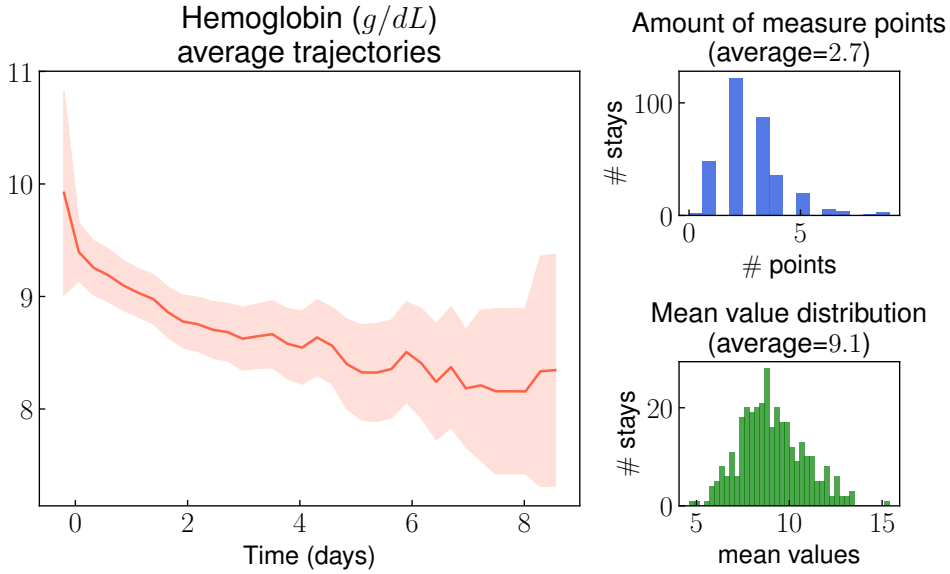
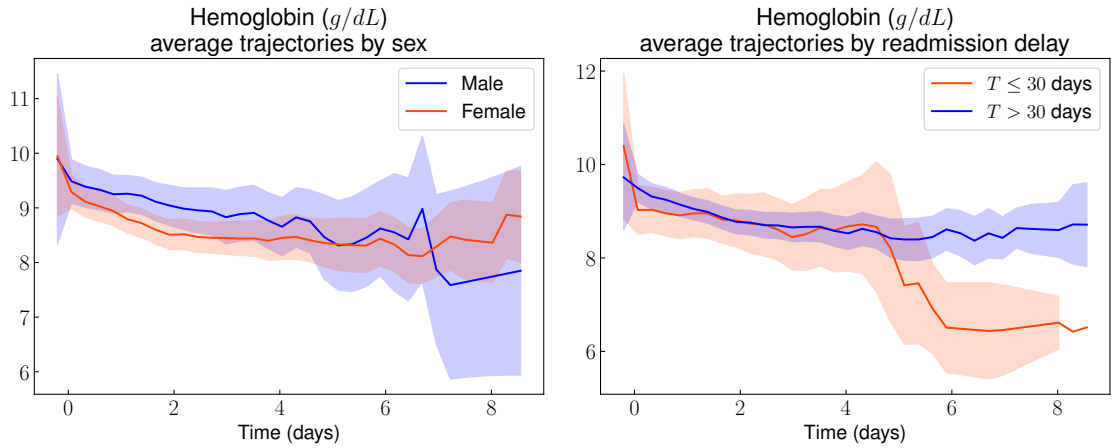


FIGURE 2.4 Left : hemoglobin average kinetics in g/dL (bold line) with 95% Gaussian confidence interval (bands). Top right : repartition of the number of hemoglobin measurement per visit ; bottom right : histogram of the hemoglobin mean.



(a) Patients are grouped by sex.

(b) Patients are grouped according to the fact that $T \leq 30$ or not.

FIGURE 2.5 Hemoglobin average kinetics in g/dL (bold line) with 95% Gaussian confidence interval (bands) for different subpopulations of patients.

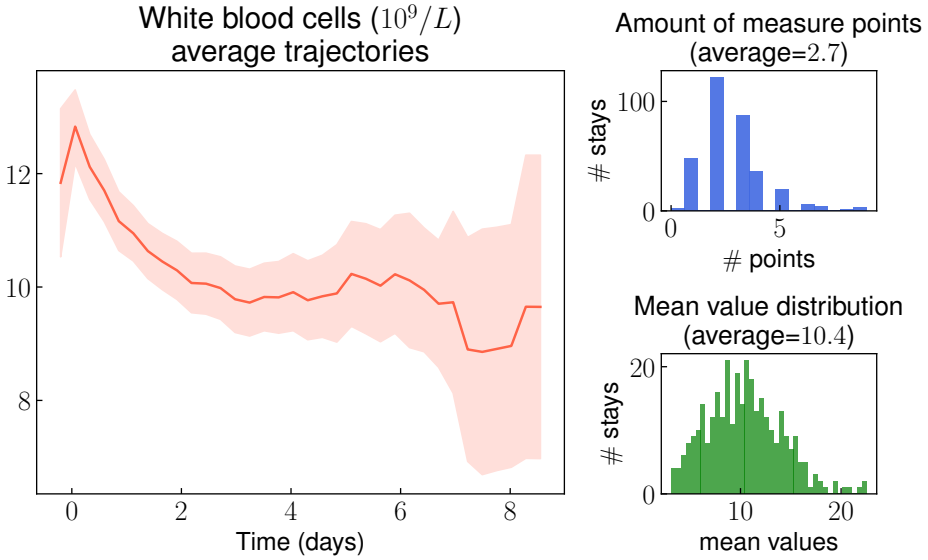


FIGURE 2.6 Left : White blood cell count average kinetics in $10^9/L$ (bold line) with 95% Gaussian confidence interval (bands). Top right : repartition of the number of measurement per visit ; bottom right : histogram of the white blood cell count mean.

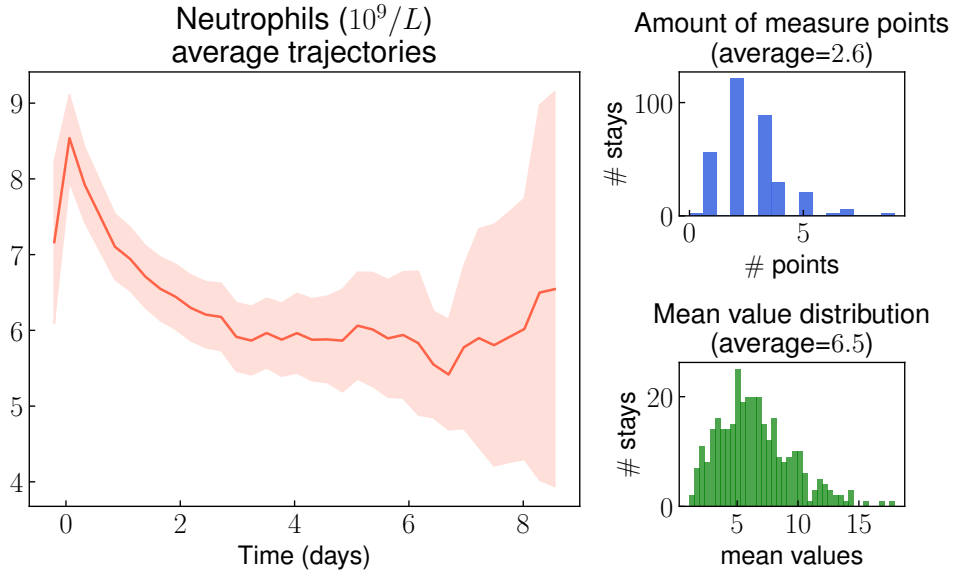


FIGURE 2.7 Left : neutrophils average kinetics in $10^9/L$ (bold line) with 95% Gaussian confidence interval (bands). Top right : repartition of the number of neutrophils measurement per visit ; bottom right : histogram of the neutrophils mean.

function and renal function markers.

2.3.3 Vital parameters trends

Figure 2.15 shows that average temperature, despite showing clear day/night cycles, remains stable throughout the stay, around 37° Celsius, with a slight gap between males and females, see Figure 2.16. Additionally, throughout the entire stay, less than 5% of patients ever reach the 38° Celsius temperature threshold, see Figure 2.17. No clear trend emerges for blood pressure and heart rate.

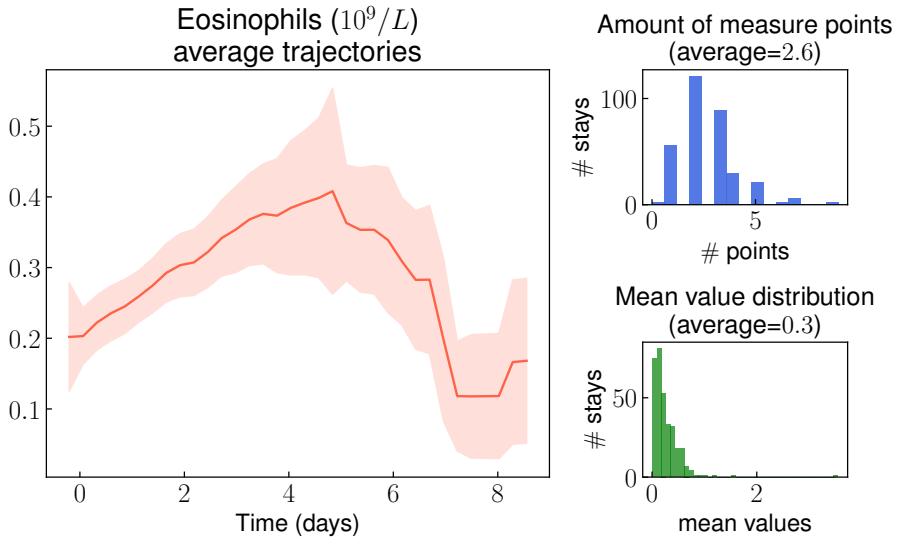


FIGURE 2.8 Left : eosinophils average kinetics in $10^9/L$ (bold line) with 95% Gaussian confidence interval (bands). Top right : repartition of the number of eosinophils measurement per visit ; bottom right : histogram of the eosinophils mean.

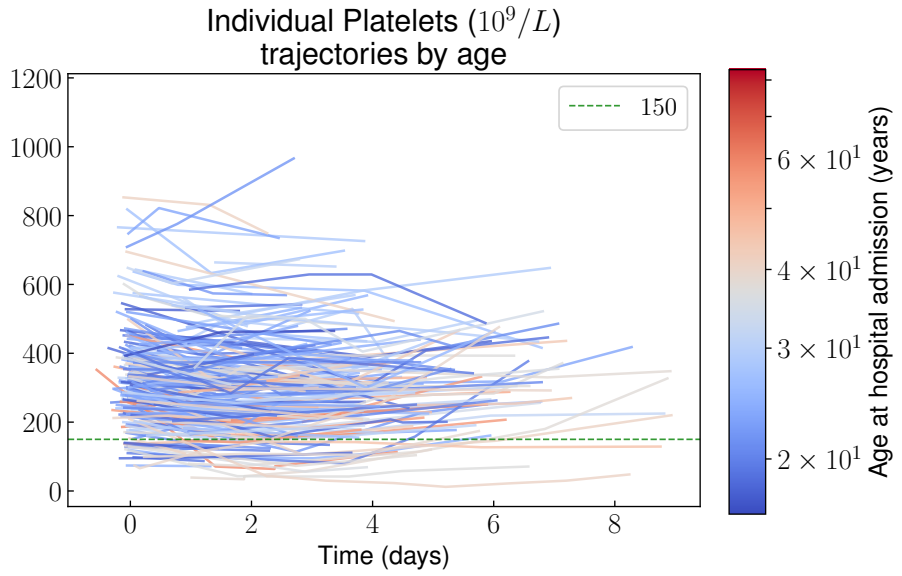


FIGURE 2.9 Individual platelets trajectories with the color gradient corresponding to the patient age : blue means young and red means old.

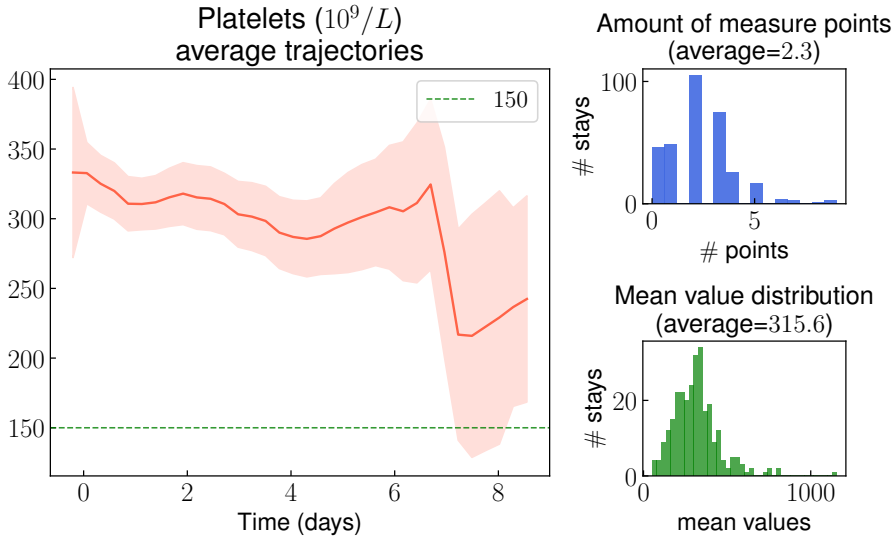


FIGURE 2.10 Left : platelets average kinetics in $10^9/L$ (bold line) with 95% Gaussian confidence interval (bands). Top right : repartition of the number of platelets measurement per visit ; bottom right : histogram of the platelets mean.

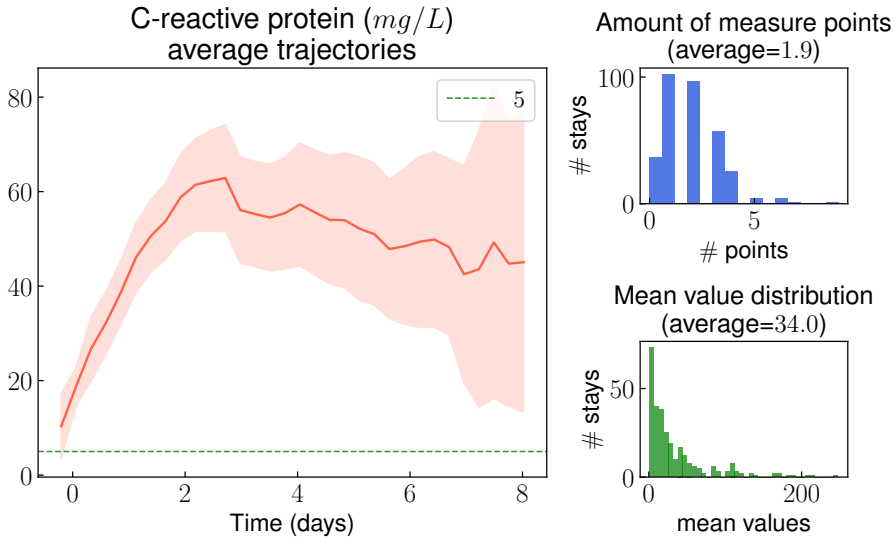
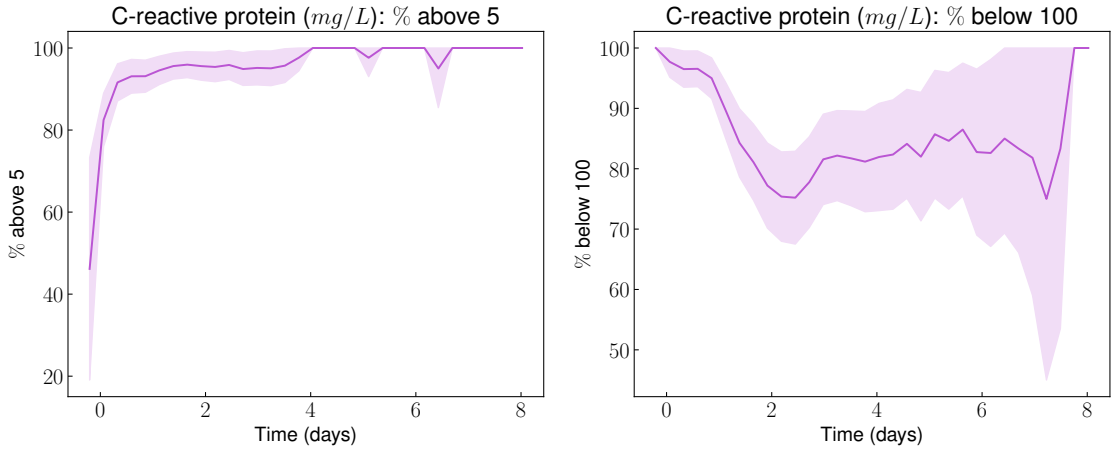


FIGURE 2.11 Left : CRP average kinetics in mg/L (bold line) with 95% Gaussian confidence interval (bands). Top right : repartition of the number of CRP measurement per visit ; bottom right : histogram of the CRP mean.



(a) Average kinetics (bold line) of the percentage of patients with a CRP below 100 mg/L, with 95% Gaussian confidence interval (bands).

(b) Average kinetics (bold line) of the percentage of patients with a CRP above 5 mg/L, with 95% Gaussian confidence interval (bands).

FIGURE 2.12 Percentage of patients with CRP above or below a given threshold according to time.

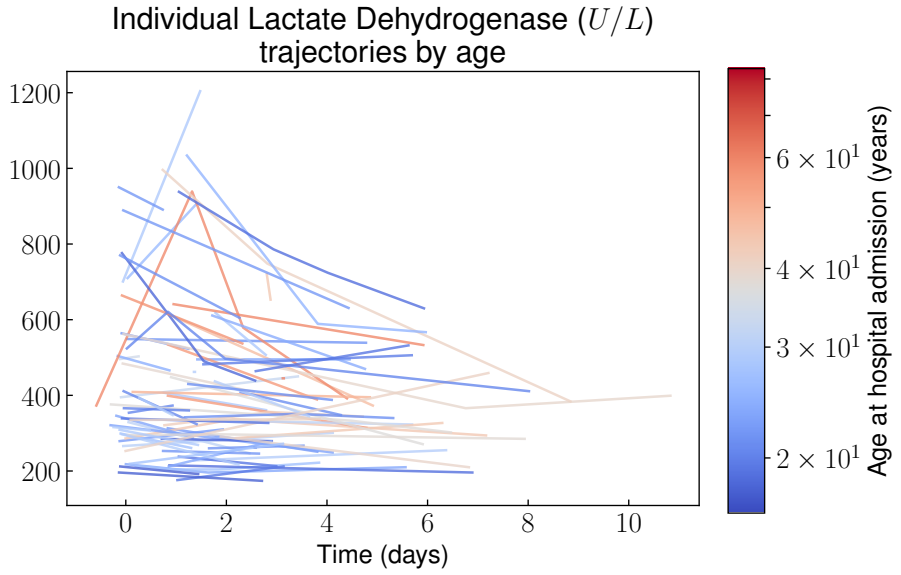


FIGURE 2.13 Individual LDH trajectories with the color gradient corresponding to the patient age : blue means young and red means old.

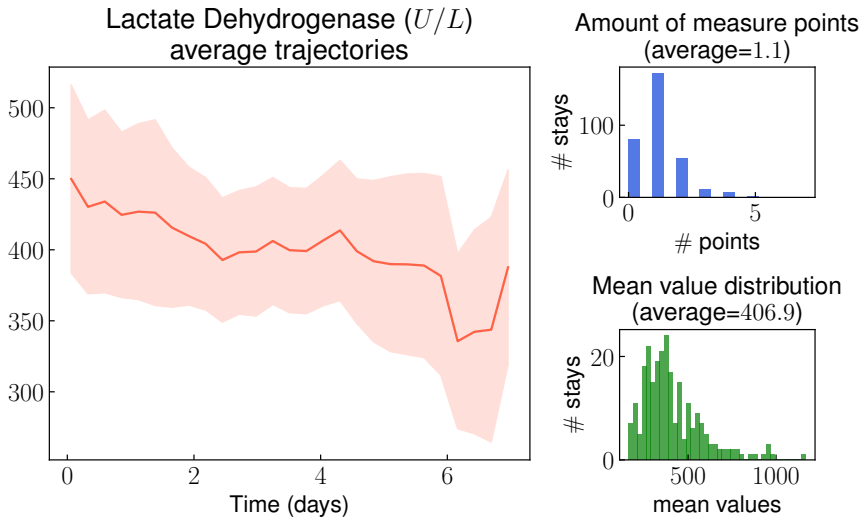


FIGURE 2.14 Left : LDH average kinetics in mg/L (bold line) with 95% Gaussian confidence interval (bands). Top right : repartition of the number of LDH measurement per visit ; bottom right : histogram of the LDH mean.

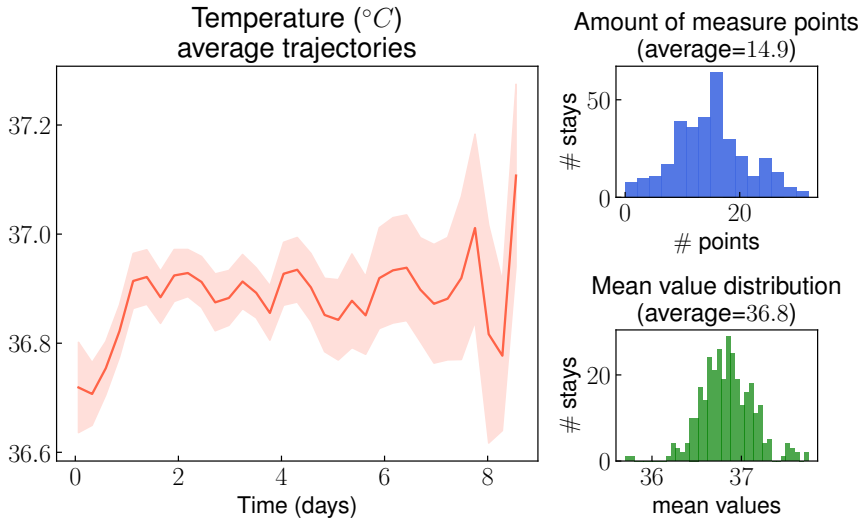


FIGURE 2.15 Left : temperature average kinetics in ° Celsius (bold line) with 95% Gaussian confidence interval (bands). Top right : repartition of the number of temperature measurement per visit ; bottom right : histogram of the temperature mean.

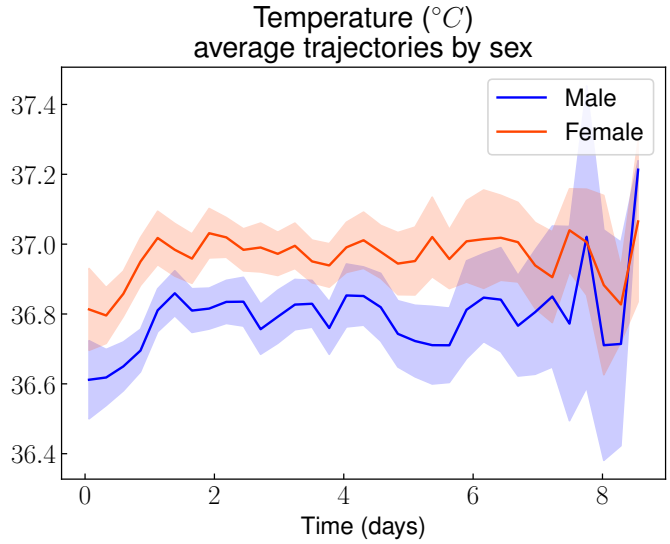


FIGURE 2.16 Temperature average kinetics in ° Celsius (bold line) with 95% Gaussian confidence interval (bands) with patients grouped according the their sex.

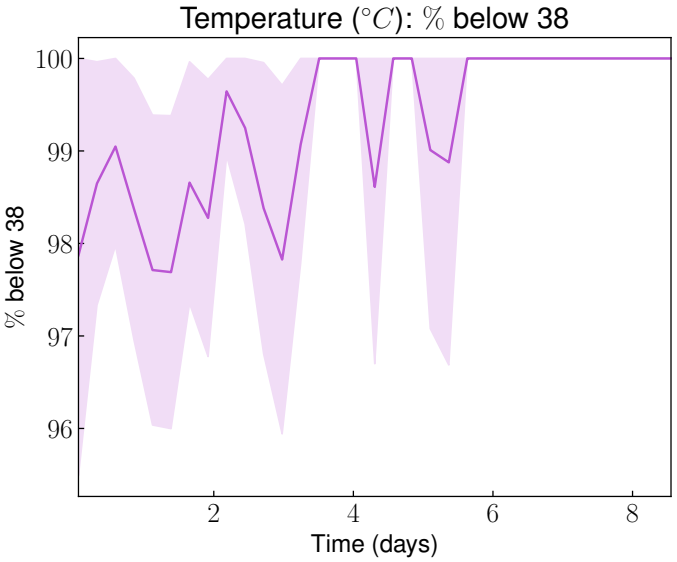


FIGURE 2.17 Percentage of patients with temperature below 38 ° Celsius according to time.

2.4 Discussion

In this chapter, we described classical trends of laboratory results and vital parameters during non-complicated VOC-related stays, using longitudinal data analysis through the reuse of routine care data. We were able to highlight slow decrease of hemoglobin level, white blood cell and neutrophil counts, and the rise of eosinophil count. We also display the 95% confidence interval of the usual laboratory results at admission and during the course of clinically unequivocal non-complicated VOC.

Although there are no specific biological diagnostic markers of VOC, the knowledge of these ranges may help the clinician infirm the diagnosis of VOC, in particular when it comes to distinguish between an authentic VOC and an unrelated pain (*e.g.* pain from morphine addiction or somatoform pain due to psychological stress). These ranges could also be useful to help diagnose a complication (*e.g.* infection or acute chest syndrome) during the course of a VOC if the patient's vital parameters or laboratory results lay outside of them.

GPUH applies a no-paper policy and our CDW pulls data from a large variety of sources. Moreover, our data were enriched with both manually extracted as well as derived covariates. Thus, we included a large number of covariates, whether they were important from an expert point of view or just routinely monitored. Unfortunately, the ranges of standard clinical and biological parameters during VOC are very large and the chapter shows that no individual marker has sufficient sensitivity or specificity by itself to either infirm or confirm the diagnosis of VOC with certainty. However several conclusions can be driven from our results.

2.4.1 Conclusions

Regarding VOC diagnosis. Markers of hemolysis (LDH ; bilirubin) do not reach levels that can be differentiated from baseline hemolysis values observed in SCD patients. Accordingly, the mean hemoglobin level was not differentiable from the mean hemoglobin level at baseline. However, it is worth noting that, retrospectively, our imputation choice for baseline hemoglobin level seems incorrect (since we used the last available level of hemoglobin during the previous hospitalization whereas our data show that the hemoglobin did not usually return to baseline at the end of hospitalization). Such choice might have interfered with these results.

Most patients displayed a significant inflammatory syndrome with elevated CRP and hyperleukocytosis, but these markers were neither sensitive nor specific enough to exclude of confirm the diagnosis of VOC. The eosinophil count rose significantly, which is in line with the previously described activated state of eosinophils in SCD patients [Canalli et al., 2004]. Finally, it is worth mentioning that after the first 48 hours following admission, our results show that normal CRP value is to be expected in fewer than 5% of VOC episodes.

Regarding VOC complication. The CRP increase observed in the first days following admission is in line with a previous study that described CRP trajectory in such context [Bargoma et al., 2005]. However, within the first 24 hours after admission, a CRP value over the 100 mg/L threshold is to be expected in less than 5% of non-complicated VOC ; such result suggests that when baseline CRP level is above this threshold, it might be reasonable to suspect and search for an associated infection. The same reasoning goes with temperature levels, which remained under 38° Celsius from admission to discharge for 95% of non-complicated VOC-related stays, showing that fever should not be attributed to the VOC episode itself.

Regarding VOC resolution. Hemolysis and anemia markers do not get back to baseline levels before hospital discharge ; thus, they are of no help to assess VOC resolution. It is worth noting that CRP level starts decreasing after the second day following admission, although it does not return to normal value before hospital discharge. The absence of anemia resolution back to baseline hemoglobin level could be a consequence of lowered the reticulocytes production due to systemic inflammation, which delays the renewal of red blood cells after hemolysis.

Regarding the risk of early readmission. Most other studies rather focused on the link between the stay's severity and steady state parameters at admission [Curtis et al., 2015, Garadah et al., 2016, Rogovik et al., 2009]. Several studies showed that elevated steady state white blood cell count and low hemoglobin level at admission were associated with higher recurrence of VOC episodes in both adults [Curtis et al., 2015, Garadah et al., 2016] and children [Krishnan et al., 2010].

In this chapter, both of these markers show a significant decreasing trend throughout the stay, which is consistent with previous findings [Ballas and Smith, 1992]. Interestingly, after stratifying by readmission delays, distinctive hemoglobin trend was observed between stays followed by an early readmission vs. other stays. Although it was shown that LDH and platelets baseline values are important markers of crisis recurrence in children [Krishnan et al., 2010], we did not highlight any trend for these parameters, neither any difference in trends between stays followed by an early readmission vs. other stays. Eosinophil count rises significantly, which is in line with the previously described activated state of eosinophils in SCD patients [Canalli et al., 2004].

This chapter therefore suggests that, for a given parameter, baseline values, steady state values, and trends throughout a hospital stay might be differentially implicated in VOC early readmission, namely because of a possible confusion between prolonged VOC (due to premature hospital discharge) and authentic recurrence. This problem is tackled in Bussy et al. [2018] (corresponding to Chapter 3).

This study benefitted from the tool we developed to analyze longitudinal data stemming from the reuse of routine care data. Such data have become increasingly available thanks to the development of electronic records coupled with CDW, bbut are difficult to handle as they measured at different times from one patient to another. This is why former studies either focused on one or two parameters, with longitudinal measures [Bargoma et al., 2005, Ballas and Smith, 1992], or considered many parameters but then settled for measures at admission [Rogovik et al., 2009] or at steady state [Garadah et al., 2016].

Instead, our method allows for a broader automated graphical description of any time-dependent covariates repeatedly measured during the timeframe of interest. When used on routine care data, such design could help automatically detect time-dependent covariates without the necessity for a prior covariates selection. Therefore, it is an efficient data-mining tool for studying large amounts of longitudinal processes. We would also argue that such advantages would only gain in relevance in time, with the quick deployment of EHR and CDW in various hospitals throughout developed countries.

Nevertheless, our study presents several limits.

2.4.2 Limits

1. Because of a lack of a gold standard criterion, the diagnosis of non-complicated VOC was inferred from several criteria to exclude both complicated VOC episodes and non-VOC related pain episodes. By restricting the diagnostic criteria too much, we might have induced a diagnostic bias.
2. As it is performed on a monocentric cohort, our study's results should be considered exploratory until reproductibility is confirmed.
3. With the GPUH internal medicine department being an SCD expertise center, it could potentially induce a selection bias compared with usual VOC hospitalization and early care practices ; similarly, since biological and vital monitoring are prescribed by experts, there is possibly a measurement bias compared with usual VOC monitoring practices.

Thus, there is a need for further investigations to properly interpret our results. We would argue that a similar study performed on a multicentric cohort from HER and CDW equipped centers, with different level of expertise, would strengthen and complete our results as well as confirm their reproductibility.

2.5 Concluding remarks

The chapter presents a new approach to study and visualize non-specific time-dependent variables. More specifically, is describes an original method to virtually automatize non-parametric trajectory visualization for any type of repeatedly measured covariates. In our case, it allowed us to quickly detect potentially relevant biomarkers and vital parameters to monitor during a VOC episode, namely hemoglobin, leucocytes and more specifically eosinophils, CRP and temperature. Interesting variables are quickly identified by the presence of specific trends when looking at averaged trajectories, sometimes after specific stratification. We also performed above/under threshold proportion analysis on visually selected covariates : over 95% of non-complicated VOC stays showed a baseline CRP value of under 100 mg/L within the first day following admission, and displayed no fever throughout the entire stay. Nevertheless, no isolated biomarker was sufficient to completely disprove the diagnosis, nor to prove the presence of a complication. Therefore, although individual biomarkers may help physicians question the diagnosis or suspect a complication, global assessment through clinical expertise remains essential in the management and surveillance of VOC episodes. A similar study performed on a multicentric cohort from HER and CDW equipped centers, with different levels of expertise, could potentially strengthen and complement our results as well as confirm their reproducibility.

Software

All the methodology discussed in this chapter is implemented in `Python`. The code is available from https://github.com/SimonBussy/redcvo in the form of annotated programs, together with a notebook tutorial. All generated figures (for all variables) are also available.

Appendices

2.A CMA scaling system

The CMA ("Complications ou Morbidité Associée") scaling system is a score used to rate hospital stays' severity in the French healthcare system. It is part of the PMSI ("Programme de Médicalisation des Systèmes d'Information"), a program that facilitates healthcare facility financing. The CMA scale ranges from 1 to 4 : level 1 corresponds to "sans CMA", meaning the stay isn't severe ; level 4 is the most severe level.

2.B Filtered-out ICD10 codes

The stays associated with the ICD10 codes given in Table 2.B.1 were excluded of the study.

TABLE 2.B.1 ICD10 codes used for patients exclusion.

ICD10 codes	About
A15 - A19	Tuberculosis
A30 - A49	Other bacterial diseases
A80 - A89	Viral infections of the central nervous system
B50 - B64	Protozoal diseases
B95 - B97	Bacterial, viral and other infectious agents
G00 - G09	Inflammatory diseases of the central nervous system
I70 - I79	Diseases of arteries, arterioles and capillaries
I80.0 - I80.2	Phlebitis and thrombophlebitis
I82	Other venous embolism and venous thrombosis
J01	Sinusitis
J02	Pharyngitis
J14 - J18	Bacterial pneumonia
J20	Acute bronchitis
J32	Chronic sinusitis
J69	Pneumonitis due to solids and liquids
K04	Diseases of pulp and periapical tissues
K12.2	Cellulitis
K81	Cholecystitis
K83	Other diseases of biliary tracts
K85	Acute pancreatitis
M10	Gout
M11	Other crystal arthropathies
M86	Osteomyelitis
N10	Acute tubulo-interstitial nephritis
N41.0	Acute prostatitis
R57.2	Septic shock
R65.0 - R65.1	Systemic Inflammatory Response Syndrome [SIRS] of infectious origin
T80 - T88	Infectious complications of surgical and medical care

2.C Mean trajectory and confidence interval

For each laboratory and vital parameter, we plotted a mean trajectory and confidence interval using the following method.

1. We created a uniform t_k time grid from the overall first measure to the last one. Zero time mark was set at the timestamp of the stay's admission. For specific covariates with known day/night cycles (*e.g.* temperature and blood pressure), zero time mark was arbitrarily set at 6pm on the day of the admission to take into account day/night cycle.

2. We fitted a first order smoothing spline f_i on each i stay's individual trajectory.

3. From the fitted individual stays trajectories we computed a mean trajectory with its confidence interval. For this final step, we followed a specific procedure to enhance precision and performance :
   (a) We calculated for each stay the variable's $f_i(t_k)$ value at every t_k time mark on the grid. If an individual trajectory's time span was shorter than the overall time grid, we considered the missing time mark's $f_i(t_k)$ values as NA. That means that individual trajectories were not imputed with spline extrapolations neither before the first available value nor after the last.
   (b) We then obtained a matrix were each row is a stay, each column is a time mark, and each cell is a measured or spline imputed value.
   (c) We assumed that for each t_k time mark, the $f_i(t_k)$ values are drawn from a Gaussian distribution. Thus, for each t_k time mark, we calculate the mean value with its 95% Gaussian confidence interval.
   (d) Additionally, we stratified this procedure for subgroups defined by readmission delay (with below and above 30 days after discharge).

4. Finally, for some selected variables whose averaged trajectories suggested that they could potentially discriminate between non-complicated and complicated VOC episodes, we calculated the proportion of patients all along the hospital stay whose values were above or below specific thresholds. The thresholds were chosen based on both clinical relevance and graphical description of the averaged trajectory.

Chapitre 3

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

Sommaire

Abstract. Choosing the most performing method in terms of outcome prediction or variables selection is a recurring problem in prognosis studies, leading to many publications on methods comparison. But some aspects have received little attention. First, most comparison studies treat prediction performance and variable selection aspects separately. Second, methods are either compared within a binary outcome setting (based on an arbitrarily chosen delay) or within a survival setting, but not both. In this chapter, we propose a comparison methodology to weight up those different settings both in terms of prediction and variables selection, while incorporating advanced machine learning strategies. Using a high-dimensional case study on a sickle-cell disease (SCD) cohort, we compare 8 statistical methods. In the binary outcome setting, we consider logistic regression (LR), support vector machine (SVM), random forest (RF), gradient boosting (GB) and neural network (NN) ; while on the survival analysis setting, we consider the Cox Proportional Hazards (PH), the CURE and the C-mix models. We then compare performances of all methods both in terms of risk prediction and variable selection, with a focus on the use of elastic net regularization technique. Among all assessed statistical methods, the C-mix model yields the better performances in both the two considered settings, as well as interesting interpretation aspects. There is some consistency in selected covariates across methods within a setting, but not much across the two settings. It appears that learning withing the survival setting first, and then going back to a binary prediction using the survival estimates significantly enhance binary predictions.

Résumé. La question du choix du modèle relativement à ses performances prédictives et sa capacité à sélectionner les covariables explicatives est récurrente dans les études pronostiques, donnant lieu à de nombreuses publications visant à comparer des modèles. Mais certains aspects ont reçus peu d'attention. D'abord, la plupart des études comparatives traitent séparément la question des performances prédictives et celle de la sélection de variables. Ensuite, les modèles sont soit comparés dans un cadre de prédiction binaire (basé sur le choix arbitraire d'un délai), soit dans un cadre d'analyse de survie, mais jamais dans les deux cadres à la fois. Dans ce chapitre, nous proposons une méthodologie pour comparer des modèles dans ces deux cadres, en terme de prédiction et de sélection de variables, en mettant l'accent sur l'utilisation de la régularisation elastic net. Nous considérons 8 méthodes, appliquées sur une étude de cas de grande dimension avec une cohorte de malades atteints de drépanocytose. Dans le cadre de prédiction binaire, nous considérons les modèles de régression logistique, machine à vecteurs de support, forêts aléatoires, gradient boosting et réseaux de neurones. Dans le cadre d'analyse de survie, nous considérons le modèle à risques proportionnels de Cox, le CURE et le C-mix. Parmi toutes ces méthodes, le C-mix donne les meilleurs résultats dans les deux cadres et procure de plus des aspects intéressant d'interprétabilité. Une certaine consistance est ob-

servée dans la sélection des variables influentes par les différents modèles au sein d'un même cadre, elle est moins forte inter-cadre. Il apparaît qu'apprendre d'abord dans le cadre plus général d'analyse de survie pour ensuite produire des prédictions binaires à l'aide des fonctions de survie estimées améliore significativement les prédictions.

3.1 Introduction

Recently, many statistical developments have been performed to tackle prognostic studies analysis. Beyond accurate risk estimation, interpretation of the results in terms of covariates importance is required to assess risk factors, with the ultimate aim of developing better diagnostic and therapeutic strategies [Pittman et al., 2004].

In most studies, covariate selection ability and model prediction performance are regarded separately. On the one hand, a considerable amount of studies report on covariates relevancy in multivariate models, mostly in the form of ajusted odds ratio [Little et al., 2009] (for instance using logistic regression (LR) model [Bender and Grouven, 1996, Mikolajczyk et al., 2008]) without reporting on the method's prediction performance (goodness-of-fit and overfitting aspects are neglected) ; namely disregarding the question : *is the model prediction still accurate on new data, unseen during the training phase ?* While on the other hand, most studies focusing on a method's predictive performance do not mention its variable selection ability [Guyon and Elisseeff, 2003], thus making it not well suited for the high-dimensional setting. Such settings are becoming increasingly common in a context where the number of available covariates to consider as potential risk factors is tremendous, especially with the development of electronic health record (EHR).

In this chapter, we discuss both aspects (prediction performance and covariates selection) for all considered methods, with a particular emphasis on the *elastic net* regularization method [Zou and Hastie, 2005]. Regularization has emerged as a dominant theme in machine learning and statistics. It provides an intuitive and principled tool for learning from high-dimensional data.

Then, a lot of studies consider prognosis as a binary outcome, namely whether the event-of-interest (death, relapse or hospital readmission for instance) occurs whithin a pre-specified period of time we denote ϵ [Tong et al., 2016, Rich et al., 1995, Vinson et al., 1990, Boulding et al., 2011]. In the following, we refer to this framework as the *binary outcome setting*, and we denote $T \geq 0$ the time elapsed before the event-of-interest and $X \in \mathbb{R}^d$ the vector of d covariates recorded at the hospital during a stay. In this setting, we are interested in predicting $T \leq \epsilon$. Such an *a priori* choice for ϵ is questionable, since any conclusion regarding both prediction and covariates relevancy is completely conditioned on the threshold value ϵ [Chen et al., 2012]. Hence, it is hazardous to make general inference on the probability distribution of the time-to-event outcome given the covariates from such a restrictive binary prediction setting.

An alternative setting to model prognosis is the survival analysis one, that takes the quantitative censored times as outcomes. The time T is right censored since in practice, some patients have not been readmitted before the end of follow-up. In the following, we refer to this setting as the *survival analysis setting* [Kleinbaum and Klein, 2010] and we denote Y the right-censored duration, that is $Y = \min(T, C)$

with C the time when the patient is lost to follow-up. Few studies compare the survival analysis and binary outcome settings and none of them considers simultaneously the prediction and the variable selection aspects in a high dimensional setting. For instance in Chen et al. [2012], only the Cox Proportional Hazards (PH) model [Cox, 1972] is considered in the survival analysis setting and a dimentionality reduction phase (or screening) is performed prior to the models comparison, as it is often the case [Dai et al., 2006, Boulesteix and Strobl, 2009].

Our case study focuses on hospital readmission following vaso-occlusive crisis (VOC) for patients with sickle-cell disease (SCD). SCD is the most frequent monogenic disorder worldwide. It is responsible for repeated VOC, which are acute painful episodes, utlimately resulting in increased morbidity and mortality [Bunn, 1997, Platt et al., 1991]. Although there are some studies regarding risk factors of early complications, only a few of them specifically addressed the question of early-readmission prediction after a VOC episode [Brousseau et al., 2010, Rees et al., 2003].

For a few decades, hospital readmissions have been known to be responsible for huge costs [Friedman and Basu, 2004, Kocher and Adashi, 2011] ; they are also a measure of health care quality. Today, hospitals have limited ressources they can allocate to each patient. Therefore, identifying patients at high risk of readmissions is a paramount question and predictive models are often used to tackle it.

The purpose of this chapter is to compare different statistical methods to analyse readmission. To make such comparisons, we consider both the predictive performance and the covariates selection aspect of each model, on the same high-dimensional set of covariates.

In the binary outcome setting, we consider LR [Hosmer Jr et al., 2013] and support vector machine (SVM) [Schölkopf and Smola, 2002] with linear kernel, being both penalized with the elastic net regularization [Zou and Hastie, 2005] to deal with the high dimensional setting and avoid overfitting [Hawkins, 2004]. We also consider random forest (RF) [Breiman, 2001], gradient boosting (GB) [Friedman, 2002] and artificial neural networks (NN) [Yegnanarayana, 2009].

We then abstain from the *a priori* threshold choice and consider the survival analysis setting. We apply first the Cox PH model [Cox, 1972]. We also apply the CURE model [Farewell, 1982, Kuk and Chen, 1992], that considers one fraction of the population as cured or not subject to any risk of readmimssion. Finally, we consider the recently developped high dimensional C-mix mixture model [Bussy et al., 2018]. The three considered models in this setting are also penalized with the elastic net regularization.

3.2 Methods

3.2.1 Motivating case study

We consider a monocentric retrospective cohort study of $n = 286$ patients. George Pompidou University Hospital (GPUH) is an expertise center for SCD adult patients [bnd]. Data is extracted from the GPUH Clinical Data Warehouse (CDW) using the i2b2 star-shaped standard [Zapletal et al., 2010]. It contains routine care data divided into several categories among them demographics, vital signs, diagnoses (ICD-10 [Organization, 2004]), procedures (French CCAM classification [Trombert-Paviot et al., 2003]), EHR clinical data from structured questionnaires, free text reports, Logical Observation Identifiers Names and Codes (LOINC), biological test results, and Computerized Provider Order Entry (CPOE) drug prescriptions. The sample included all stays from patients admitted to the internal medicine department for VOC (ICD-10 57.0 or 57.2) between January 1st 2010 and December 31st 2015.

Over half of the patients has only one stay during the follow-up period. We hence randomly sample one stay per patient and focus on the early-readmission risk afterwards. This enables us, in addition, to work on the *independent and identically distributed* standard statistical framework.

3.2.2 Covariates

We extracted demographic data (*e.g.* sex, date of birth, last known vital status), as well as both qualitative (*e.g.* the admission at any point during the stay to an ICU, the type of opioid drug received) and quantitative time-dependent variables (*e.g.* biological results, vital sign values, intraveinous opiod syringes parameters) regarding each stay.

We also extracted all the free text reports from the patients' EHR regardless of the source department and the stay. In order to facilitate variable extraction from such textual reports, we used a locally developed browser-accessible tool called FASTVISU [Escudié et al., 2015]. This software is linked with the CDW, and allowed us to quickly check throughout these textual reports for highlighted information and to vote for variable status (*e.g.* SCD genotype) or value (*e.g.* baseline hemoglobinemia).

Keywords using regular expressions are used to focus on specific mentions within the text. Variables extracted using this tool were the following : SCD genotype, baseline hemoglobinemia, medical history (with a focus on previous VOC complications and SCD-related chronic organ damages), and lifestyle related information. For time-dependent variables, status was determined per stay, including the ones that were not related to a VOC episode (*e.g.* annual check-ups).

We extracted for the included patients all stays encoded as VOC to derive time length from and until the respectively previous and consecutive stays. Regarding demographic data, we derived the patient's age at admission for each stay. For each time-dependent covariate, all patient relative time series have different number of points and different length. We then propose a method to extract several covariates from each time series, to make the use of usual machine learning algorithms possible :

— Regarding all vital parameters and oxygen use, we derived them by calculating the average value and the linear regression's slope for the last 48 hours of the stay, as well as the delay between the end of oxygen support and the hospital discharge.

— Regarding biological variables, we only kept the ones that were measured at least once for more than 50% of the stays. We considered the last measured value for each of them before discharge. Additionally, for covariates with at least 2 distinct measurements per stay, we considered the linear regression's slope for the last 48 hours of the stay. In order to maximize the amount of biological data, we also retrieved the biological values measured in the emergency department, prior to the administrative admission of the patient.

— For each time-dependent covariate and for each stay, we fit a distinct Gaussian process on the last 48 hours of the stay for all patient with at least 3 distinct measurements during this period, and extract the corresponding hyper-parameters as covariates for our problem.

Indeed, Gaussian processes are known to fit EHR data well ; see for instance Pimentel et al. [2013], where a distinct Gaussian process is also fitted for each patient and each time-dependent covariate, in order to cluster patients into groups in the hyper-parameter space. In our study, we instead use the hyper-parameters as covariates in a supervised learning way. We use Gaussian process with linear average function and a sum-kernel composed by a constant kernel which modifies the mean of the Gaussian process, a radial-basis function kernel, and a white kernel to explain the noise-component of the signal.

After a binary encoding of the categorical covariates, the final dimension of the working space (number of considered covariates) is $d = 174$. Therefore, the number of patients is less than 2 times as many as the number of covariates, making it difficult to use standard regression techniques. More details on data extraction, missing data imputation, as well as a precise list of all considered covariates, are given in Sections 3.A.1, 3.A.2 and 3.A.3 respectively.

3.2.3 Statistical methods and analytical strategies

a) Binary outcome setting

In this setting, we consider as early-readmission any readmission occuring within 30 days of hospital discharge after a previous hospital stay for VOC, the 30 days threshold being a standard choice in SCD studies [Brousseau et al., 2010, Frei-Jones et al., 2009]. A first drawback of this setting (which is rarely mentionned) is that patients having both a censored time and $c_i \leq \epsilon$ have to be excluded from the procedure, since we do not know if $t_i \leq \epsilon$ or not. Figure 3.1 gives an illustration of this last point. In our case, 7 patients have to be excluded because of this issue.

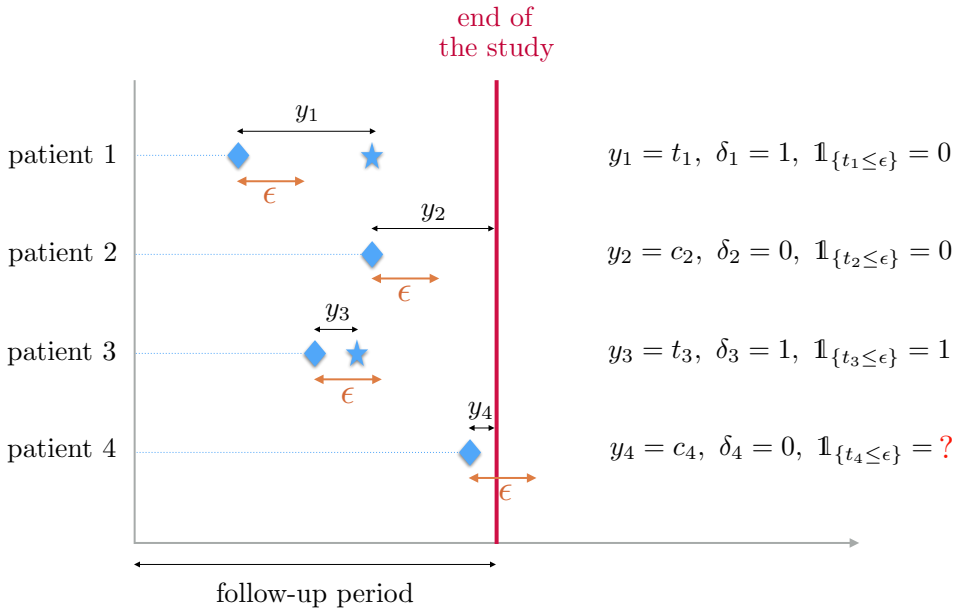


FIGURE 3.1 Illustration of the problem of censored data that cannot be labeled when using a threshold ϵ. $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$ is the censoring indicator which is equal to 1 if Y_i is censored and 0 otherwise. In the binary outcome setting, patient 4 would be excluded.

We first consider LR [Hosmer Jr et al., 2013] and linear kernel SVM [Schölkopf and Smola, 2002], both penalized with the elastic net regularization [Zou and Hastie, 2005]. For a given model, using this penalization means adding the following term

$$\gamma\Big((1-\eta)\|\beta\|_1 + (\eta/2)\|\beta\|_2^2\Big)$$

to the cost function (the negative likelihood for instance) in order to minimize it in $\beta \in \mathbb{R}^d$, a vector of coefficients that quantifies the impact of each biomedical

covariates on the associated prediction task. This means that the elastic net regularization term is a linear combination of the lasso (ℓ_1) and ridge (squared ℓ_2) penalties for a fixed $\eta \in (0, 1)$, tuning parameter γ, and where we denote

$$\|\beta\|_p = \Big(\sum_{i=1}^{d} |\beta_i|^p\Big)^{1/p}$$

the ℓ_p-norm of β. One advantage of this regularization method is its ability to perform model selection (for the lasso part) and to pinpoint the most important covariates relatively to the prediction objective. On the other hand, the ridge part allows to handle potential correlation between covariates [Zou and Hastie, 2005]. The penalization parameter γ is carefully chosen using the same cross-validation procedure [Kohavi et al., 1995] for all competing models. Note that in practice, the intercept is not regularized.

We also consider other machine learning algorithms in the ensemble methods class such as RF [Breiman, 2001] and GB [Friedman, 2002]. For both algorithms, all hyper-parameters are tuned using a randomized search cross-validation procedure [Bergstra and Bengio, 2012]. For instance for RF : the number of trees in the forest, the maximum depth of the tree or the minimum number of samples required to split an internal node.

Note also that regarding the covariates importance for RF and GB, we use the Gini importance [Menze et al., 2009], defined as the total decrease in node impurity weighted by the probability of reaching that node (which is approximated by the proportion of samples reaching that node) averaged over all trees of the ensemble. That is why the corresponding coefficients are all positive for those two models, which is to be kept in mind.

Finally, we consider NN [Yegnanarayana, 2009] in the form of a multilayer perceptron neural network with one hidden layer. We use stochastic gradient-based optimizer for NN and rectified linear units activation function to get sparse activation and be able to compare covariate importance [Glorot et al., 2011]. The regularization term as well as the number of neurons in the hidden layer are also cross-validated though a random search optimization.

Note that many studies in the literature choose hyper-parameters of the models, without mentioning any statistical procedure to determine them without *a priori* [Puddu and Menotti, 2012].

For all considered models in this setting, we use the reference implementations from the `scikit-learn` library [Pedregosa et al., 2011b].

b) Survival analysis setting

The Cox PH model is by far the most widely used in the survival analysis setting ; see Cox [1972] and Simon et al. [2011] for the penalized version. It is a regression model that describes the relation between intensity of events and covariates, given by

$$\lambda(t|X = x) = \lambda_0(t)\exp(x^\top \beta)$$

where λ_0 is a baseline intensity describing how the event hazard changes over time at baseline levels of covariates, and β is a vector quantifying the multiplicative impact on the hazard ratio of each covariate. We use the `R` packages `survival` and `glmnet` to train this model.

An alternative to the Cox PH model is the CURE model [Farewell, 1982] that considers one fraction of the population as not subject to any risk of readmission, with a logistic function for the incidence part and a parametric survival model. We add an elastic net regularization term and use the appropriate implementation of the QNEM algorithm detailed in Section 4.4.1.

Finally, we apply the C-mix model [Bussy et al., 2018] that is designed to learn risk groups in a high dimensional survival setting. For a given patient i, it provides a marker $\pi_{\hat{\beta}}(x_i)$ estimating the probability that the patient is at high risk of early-readmission. Note that $\hat{\beta}$ denotes the estimate vector after the training phase for any model.

We randomly split data into a training set and a test set (30% for testing, cross-validation is done on the training). In both binary outcome and survival analysis settings, all the prediction performances are evaluated on the test set after the training phase, using the relevant metrics detailed hereafter. Note also that for all considered models (except RF and GB), continuous covariates are standardized through a pre-processing step, which allows proper comparability between the covariates' effects whithin each model.

3.2.4 Metrics used for analysis

In the binary outcome setting, the natural metric used to evaluate performances is the AUC [Bradley, 1997]. In the survival analysis setting, the natural equivalent is the C-index (implemented in the `python` package `lifelines`), that is

$$\mathbb{P}[M_i > M_j | Y_i < Y_j, Y_i < \tau]$$

with $i \neq j$ two independent patients, τ corresponding to the follow-up period duration [Heagerty and Zheng, 2005], and M_i the natural risk marker of the model for patient i : $\exp(x_i^\top \hat{\beta})$ for the Cox PH model, the probability of being uncured for the CURE model and $\pi_{\hat{\beta}}(x_i)$ for the C-mix.

To compare the two settings, one can predict the survival function $\hat{S}_i$ for each model and for patients i in the test set. Then, for a given threshold ϵ, one can now use

$$\hat{S}_i(\epsilon|X_i = x_i)$$

for each model to predict whether or not $T_i \leq \epsilon$ on the test set – relaying to the binary outcome setting – thus assessing performances using the classical AUC score. Then, with $\epsilon = 30$ days, one can directly compare prediction performances with those obtained in the binary outcome setting.
Details on the survival function estimation procedure for each model are given in Section 3.B.1.

Finally, we compute the pairwise Pearson correlation between the absolute (because of the positive vectors for RF and GB) covariates importance vectors of each method to obtain a similarity measure in terms of covariates selection [Kalousis et al., 2007].

3.3 Results

Table 3.1 compares the prediction performances of the different methods in both considered settings using appropriate metrics. Corresponding hyper-parameters obtained by cross-validation are detailed in Section 3.B.2.

TABLE 3.1 Comparison of prediction performances in the two considered settings, with best results in bold.

Setting	Metric	Model	Score
Survival analysis	C-index	CURE	0.718
		Cox PH	0.725
		C-mix	**0.754**
Binary outcome	AUC	SVM	0.524
		GB	0.561
		LR	0.616
		NN	0.707
		RF	0.738
		CURE ($\epsilon = 30$)	0.831
		Cox PH ($\epsilon = 30$)	0.855
		C-mix ($\epsilon = 30$)	**0.940**

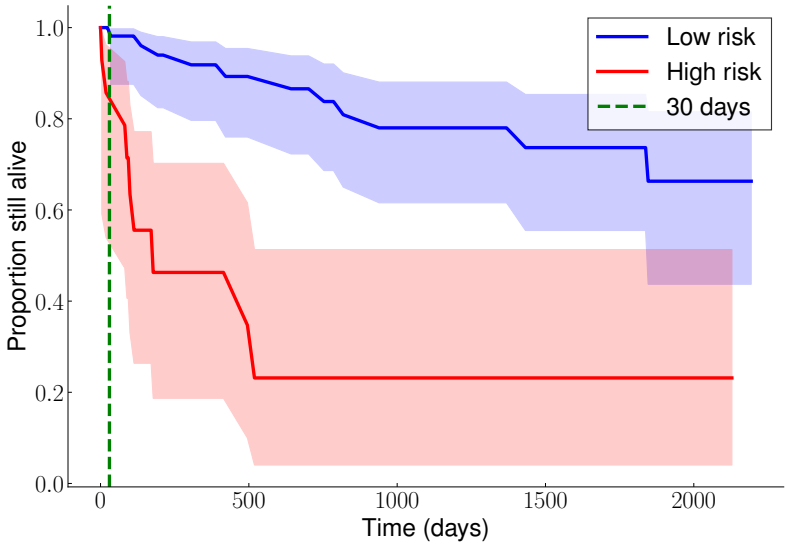


FIGURE 3.1 Estimated survival curves per subgroups (blue for low risk and red for high risk) with the corresponding 95 % confidence bands

Thus, making binary predictions from survival analysis models using estimated survival function highly improves performances. The C-mix yields the best results. Figure 3.1 displays the estimated survival curves for the low and high risk of early-readmission subgroups learned by this model. Note the clear separation between the two subgroups.

Based on those early-readmission risk learned subgroups, we test for significant differences between them using Fisher-exact test [Upton, 1992] for binary covariate, and Wilcoxon rank-sum test [Wilcoxon, 1945] for quantitative covariate.

Then, we similarly test for significant difference, on each covariate, between naively created groups obtained with the binary outcome setting ($\epsilon = 30$ days). We also use the log-rank test [Harrington and Fleming, 1982] on each covariate, directly involving quantitative readmission delays.

Finally, we compared the obtained significance (the p-value) for each test, on each covariate. The tests induced by the C-mix model are the most significant ones for almost all covariates. The top-6 p-values of the tests are compared in Figure 3.2.

Taking the most significant C-mix groups highlighted in Figure 3.2, Figure 3.5 shows either boxplot (for quantitative covariates) or repartition (for qualitative covariates) comparison between those groups. One can now easily visualize and interpret early-readmission risk data-driven grouping, and focus on specific covariate.

For instance, it appears that patients among the high risk group tend to have

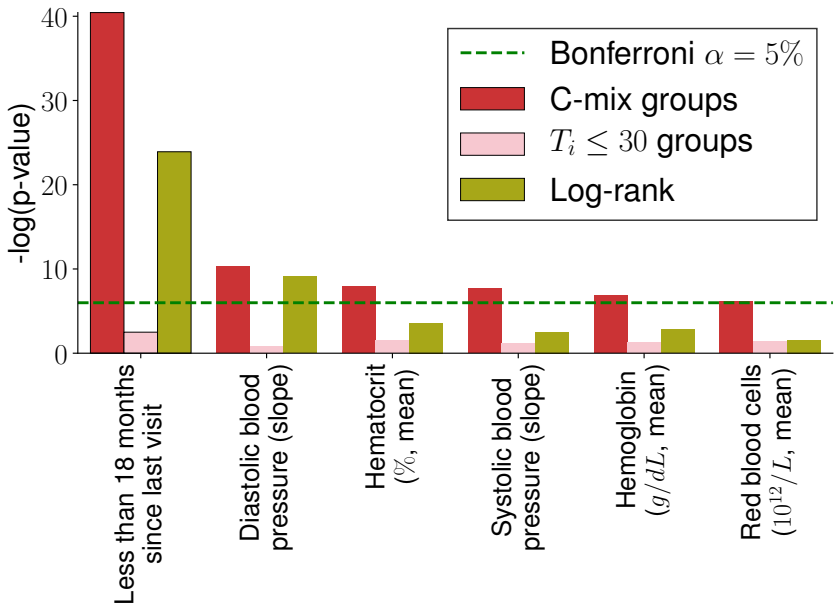


FIGURE 3.2 Comparison of the tests based on the C-mix groups, on the $\epsilon = 30$ days relative groups and on survival times. We arbitrarily shows only the tests with corresponding p-values below the level $\alpha = 5\%$, with the classical Bonferroni multitests correction [Bonferroni, 1935].

a lower hemoglobin level, as well as a slightly lowering diastolic blood pressure in the last 48 hours of the stay (while slightly uppering for the low risk group). It also appears that less patients among the low risk group have visited the emergency department in the last 18 months.

Let us now focus on the covariates selection aspect for each method. Figure 3.3 gives an insight on the covariates importance relatively to each model for 20 arbitrarily chosen covariates (selected on decreasing importance order for the C-mix model). The result with all covariates can be found in Section 3.B.3. One can observe some consistency between methods. Figure 3.4 gives a global similarity comparison measure in terms of covariates selection. We observe higher similarities between methods within a single setting.

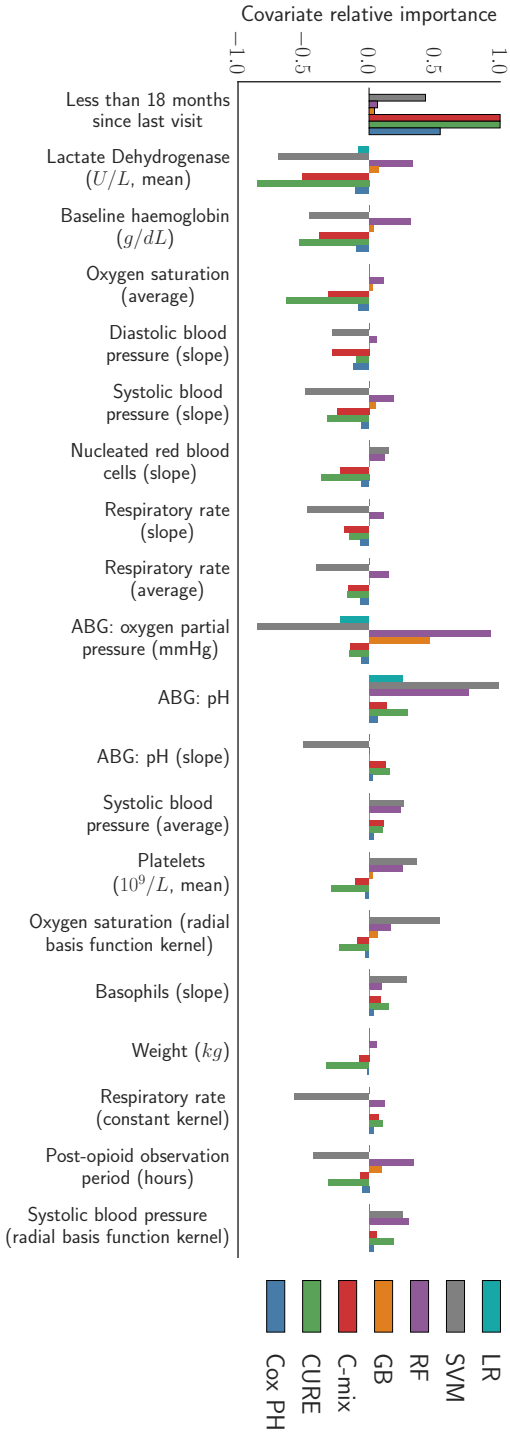


FIGURE 3.3 Comparison of the top-20 covariates importance ordered on the C-mix estimates. Note that some time-dependent covariates, such as average cinetic during the last 48 hours of the stay (slope) or Gaussian Processes kernels parameters, appear to have significant importances.

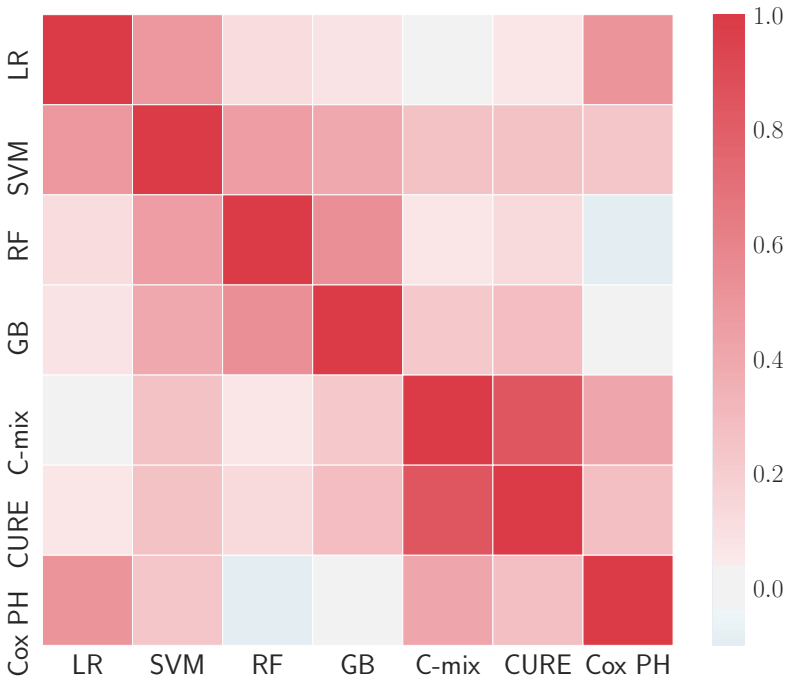


FIGURE 3.4 Pearson correlation matrix for comparing covariates selection similarities between methods. Red means high correlations.

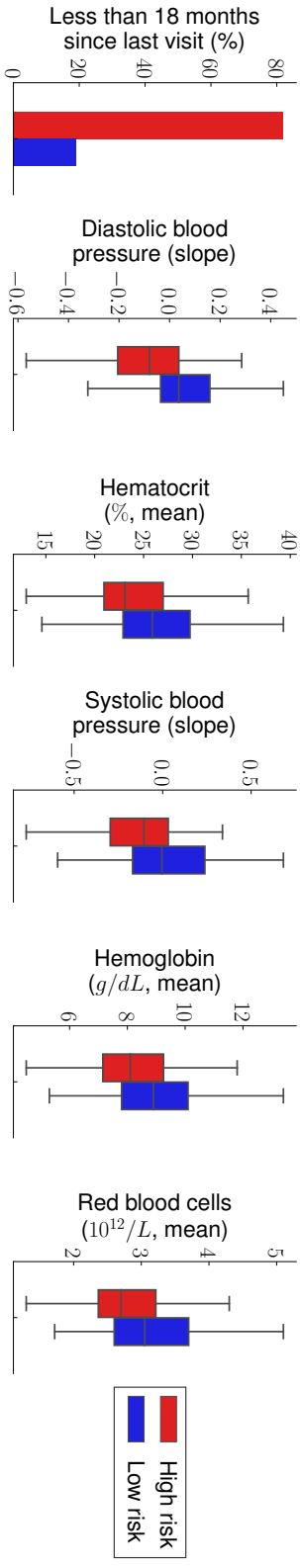


FIGURE 3.5 Covariates boxplot comparison between the most significant C-mix groups.

3.4 Discussion

In this chapter, rather than trying to be exhaustive in terms of considered methods, we choose, accordingly with the aim of this chapter, to offer a methodology for fairly comparing methods in the two considered settings. Also, we do not try different ϵ values, as it is done in Chen et al. [2012] (where emphasis is on performance metrics), since our focus is to propose a general comparison and interpretation methodology, with an analysis that remains valid for any choice of ϵ value.

In the binary outcome setting, classifiers highly depend on how the risk groups are defined : a slight change of the survival threshold ϵ for assignment of classes can lead to different prediction results [Chen et al., 2012]. In our dataset, only 5.2% of the visits lead to a readmission within 30 days. We are then in a classical setup where the adverse event appears rarely in the data at our disposal. In such setting, a vast amount of temporal information is lost since the model only knows if a readmission occurs before the threshold delay or not. It appears that taking all the information through the survival analysis setting first, and then going back to a binary prediction using the survival estimate, significantly enhances any binary prediction, which intuitively makes sense.

Among all methods, the C-mix holds the best results. Its good performances compared to other methods is already shown in Bussy et al. [2018], both in synthetic and real data. While the Cox PH regression model is widely used to analyze time-to-event data, it relies on the proportional hazard ratio assumption. But in the case of VOC for instance, it is plausible that these early-readmissions are the consequences of the same ongoing crisis (hospital discharge before the VOC recovery), whereas late-readmissions are genuine new unrelated crisis (recurrence). This would suggest that the proportional hazard ratio assumption for Cox PH model (or its related models like the competing risks model, the marginal model or the frailty model ; for this reason not considered in this study) is not respected in this situation. The CURE model main hypothesis being that a proportion of the patient is cured is questionable too. Those reasons partly explain the good performances of the C-mix model that does not rely on any restrictive hypothesis.

In this study, data extraction was performed with no *a priori* on the relevance of each variable. For instance, we extracted all biological covariates that have been measured during a patient's stay, without presuming of their importance on readmission risk. Selected variables in our use case are relevant from a clinical point of view, highlighting the capacity of regularization methods to pinpoint clinically relevant covariates.

The most important covariates in the survival setting are linked to the severity of the underlying SCD (*e.g.* crisis frequency, baseline hemoglobin), while selected

covariates in the binary outcome setting are more related to the crisis biological parameters (*e.g.* arterial blood gas parameters). Some covariates appear to be selected in both settings (*e.g.* mean lactate deshydrogenase). All selected covariates make sens from a clinical point of view, and the difference between the two settings seems to be related to the underlying hypotheses of each setting : as binary setting only takes information on early readmission, crisis related parameters are favored ; meanwhile in the survival setting, parameters related to the severity of the underlying SCD are favored. This underlines why it is crucial, when working on prognosis analysis, to use several methods to get an insight of the most important covariates.

3.5 Concluding remarks

In this chapter, we compare methods in terms of prediction performances and covariates selection for different statistical and machine learning methods on a readmission framework with high dimensional EHR data. We particularly focus on comparing survival and binary outcome settings. Methods from both settings must be considered when working on a prognosis study. Indeed, important covariates are possibly different depending on the setting : for instance in our case study, we highlight important covariates linked either to the severity of the underlying SCD or to the severity of the crisis.

Not only do frequent readmissions affect SCD patients' quality of life, they also impact hospitals' organization and induce unnecessary costs. Our study lays the groundwork for the development of powerful methods which could help provide personalized care. Indeed, such early-readmission risk-predicting tools could help physicians decide whether or not a specific patient should be discharged of the hospital. Nevertheless, most selected covariates were derived from raw or unstructured extracted data, making it difficult to implement the proposed predictive models into routine clinical practice.

All results in the binary outcome setting rely on a critical readmission delay choice, which is a questionable - if not counterproductive - bias in readmission risk studies. Additionally, we point out the idea that learning in the survival setting, rather than directly from the binary outcome setting, and then making binary predictions through the estimated survival function for a given delay threshold can dramatically enhance performances.

Finally, the C-mix model yields the better performances and can be an interesting alternative to more classical methods found in the medical literature to deal with prognosis studies in a high dimensional framework. Moreover, it provides powerful

interpretations aspects that could be useful in both clinical research and daily practice (see Figure 3.5). It would be interesting to generalize our conclusions to external datasets, which is the purpose of further investigations.

Software

All the methodology discussed in this chapter is implemented in `Python`. The code that generates all figures is available from https://github.com/SimonBussy/early-readmission-prediction in the form of annotated programs, together with notebook tutorials.

Appendices

3.A Details on covariates

3.A.1 Covariates creation

Since SCD patients are frequently treated with opioids to control the pain induced from VOCs, some may develop, over time, an addiction to these products. Such addiction may cause readmission and often interferes with hospitalization timeline. In order to limit confusion bias, we excluded patients encoded as opioid addicts (ICD-10 F11) as well as those who were treated with substitute products such as Methadone or Buprenorphine, both determined from hospitalization reports and drug prescriptions.

Regarding opioid treatment related information from the CDW, based on doctors and nurses inputs, variables extracted were the following :

— the specific molecule of each prescription,

— the specific dosage form of each prescription,

— the initiation and ending timestamps of each prescription.

From these variables, we also derived the following :

— the delay between the end of the last syringe received and the hospital discharge,

— the number of syringes used per day on average,

— the slope from the linear regression of the delay between syringes throughout the stay.

Regarding intravenous opioid treatments, we also extracted bolus dosage, maximum dosage, and refractory period. In order to capture both the average level and the general trend of these covariates, we derived them by calculating the slope and intercept from the linear regression of each of these parameters throughout the stay.

3.A.2 Missing data

We substitute missing medical history related data as follows : if a specific medical condition or VOC complication is mentioned in a report, this item is considered as part of the patient' medical history for every chronologically following stays ; if a specific medical condition or VOC complication is explicitly stated as absent from the medical history in a report, this item is considered absent in all the previous stays.

For other specific covariates, we proceed that way :

— for the patients' baseline hemoglobin value, we use the last hemoglobin value measured during the first included stay,

— for the dichotomous variables regarding the patient's entourage and professional activity, we use the most represented value amongst all stays (of all patients),

— we consider non-mentioned medical history or VOC complications as absent,

— we consider that all patients received both opioid treatments and oxygen therapy at admission in the emergency room. Therefore, we consider the post-opioid observation period, as well as the post-oxygen observation period, to be the same time length as the entire stay.

For all remaining covariates, we impute as follows (after the random sampling of one stay per patient) :

— numerical variables are imputed with their median values,

— categorical variables are imputed with their most represented values.

3.A.3 List of covariates

Table 3.A.1 summarizes the concepts used and their basic properties.

TABLE 3.A.1 List of the considered concepts. For each one, we display the name (with unit), the category, the sub-category if relevant, and the type ("Q" for Qualitative, "B" for Binary and "C" for Categorical). For practical purposes, we only display basic concepts without describing the list of covariates induced from it and used in practice, since the process of covariates extraction is thoroughly described in the chapter. For instance, the temperature concept gives rise to 5 covariates, relatively to its average and slope in the last 48 hours as well as the corresponding Gaussian Process kernel hyper-parameters.

Name (unit)	Category	Sub-category	Type	Name (unit)	Category	Type
Red blood cells ($10^{12}/L$)	Biological data	Complete blood count	Q	Respiratory rate (mvt/min)	Clinical data	Q
Hemoglobin (g/dL)	Biological data	Complete blood count	Q	Heart rate (bpm)	Clinical data	Q
Haemoglobin gap to baseline (g/dL)	Biological data	Complete blood count	Q	Oxygen saturation (%)	Clinical data	Q
Hematocrit (%)	Biological data	Complete blood count	Q	Temperature ($°C$)	Clinical data	Q
Mean cell volume (fl)	Biological data	Complete blood count	Q	Post-oxygen observation period (hours)	Clinical data	Q
Mean corpuscular hemoglobin (pg)	Biological data	Complete blood count	Q	Systolic blood pressure ($mmHg$)	Clinical data	Q
Mean corpuscular hemoglobin concentration (%)	Biological data	Complete blood count	Q	Diastolic blood pressure ($mmHg$)	Clinical data	Q
Reticulocytes ($10^9/L$)	Biological data	Complete blood count	Q	Gender	General features	B
Nucleated red blood cells ($10^9/L$)	Biological data	Complete blood count	Q	Baseline haemoglobin (g/dL)	General features	Q
White blood cells ($10^9/L$)	Biological data	Complete blood count	Q	Genotype	General features	B
Neutrophils ($10^9/L$)	Biological data	Complete blood count	Q	Distance between home and GPUH (km)	General features	Q
Neutrophils (%)	Biological data	Complete blood count	Q	Driving time from home to GPUH (minutes)	General features	Q
Basophils ($10^9/L$)	Biological data	Complete blood count	Q	Age at hospital admission	General features	Q
Basophils (%)	Biological data	Complete blood count	Q	French DRG code (GHM)	General features	C
Eosinophils ($10^9/L$)	Biological data	Complete blood count	Q	Severity level of the stay	General features	C
Eosinophils (%)	Biological data	Complete blood count	Q	Length of hospital stay (hours)	General features	Q
Monocytes ($10^9/L$)	Biological data	Complete blood count	Q	Time length since last admission (days)	General features	Q
Monocytes (%)	Biological data	Complete blood count	Q	Less than 18 months since last admission	General features	Q
Lymphocytes ($10^9/L$)	Biological data	Complete blood count	Q	Time length to next admission (days)	General features	Q
Lymphocytes (%)	Biological data	Complete blood count	Q	Stayed in ICU	General features	B
Platelets ($10^9/L$)	Biological data	Complete blood count	Q	Number of RBC transfusions	General features	Q
Mean platelet volume (fl)	Biological data	Complete blood count	Q	Professional activity	Lifestyle	B
Hemoglobin S (%)	Biological data	Hemoglobin electrophoresis	Q	Household situation	Lifestyle	B
Hemoglobin F (%)	Biological data	Hemoglobin electrophoresis	Q	Acute chest syndrom	Medical history	B
Asparate transaminase (U/L)	Biological data	Liver function test	Q	Avascular bone necrosis	Medical history	B
Alanine transaminase (U/L)	Biological data	Liver function test	Q	Priapism (only for males)	Medical history	B
Alkaline phosphatase (U/L)	Biological data	Liver function test	Q	Ischemic stroke	Medical history	B
Gamma glutamyl-tranferase (U/L)	Biological data	Liver function test	Q	Leg skin ulceration	Medical history	B
Direct bilirubin (mol/L)	Biological data	Liver function test	Q	Heart failure	Medical history	B
Total bilirubin (mol/L)	Biological data	Liver function test	Q	Pulmonary hypertension	Medical history	B
Urea ($mmol/L$)	Biological data	Renal function test	Q	Known nephropathy	Medical history	B
Creatinine (mol/L)	Biological data	Renal function test	Q	Known retinopathy	Medical history	B
Renal function by MDRD ($mL/min/1,73m^2$)	Biological data	Renal function test	Q	Dialysis	Medical history	B
Sodium ($mmol/L$)	Biological data	Serum electrolytes	Q	Received Morphine	Opioid use	B
Potassium ($mmol/L$)	Biological data	Serum electrolytes	Q	Received Oxycodone	Opioid use	B
Chloride ($mmol/L$)	Biological data	Serum electrolytes	Q	Received orally administered opioids	Opioid use	B
Bicarbonate ($mmol/L$)	Biological data	Serum electrolytes	Q	Number of syringes received per day	Opioid use	Q
Total calcium ($mmol/L$)	Biological data	Serum electrolytes	Q	Delay between syringes (slope)	Opioid use	Q
Proteins (g/L)	Biological data	Serum electrolytes	Q	Post-opioid observation period (hours)	Opioid use	Q
Glucose ($mmol/L$)	Biological data	Serum electrolytes	Q	Bolus dosage	Opioid use	Q
C-reactive protein (mg/L)	Biological data	Other blood markers	Q	Maximum dosage	Opioid use	Q
Lactate Dehydrogenase (U/L)	Biological data	Other blood markers	Q	Refractory period	Opioid use	Q
Weight (kg)	Clinical data	Body dimensions	Q			
Size (cm)	Clinical data	Body dimensions	Q			
Body mass index (kg/m^2)	Clinical data	Body dimensions	Q			

3.B Details on experiments

3.B.1 Survival function estimation

For the Cox PH model, the survival

$$\mathbb{P}[T_i > t | X_i = x_i]$$

for patient i in the test set is estimated by

$$\hat{S}_i(t|X_i = x_i) = [\hat{S}_0^{\text{cox}}(t)]^{\exp(x_i^\top \hat{\beta})},$$

where $\hat{S}_0^{\text{cox}}$ is the estimated survival function of baseline population ($x = 0$) obtained using the Breslow estimate of λ_0 [Breslow, 1972].

For the CURE or the C-mix models, it is naturally estimated by

$$\hat{S}_i(t|X_i = x_i) = \pi_{\hat{\beta}}(x_i)\hat{S}_1(t) + \Big(1 - \pi_{\hat{\beta}}(x_i)\Big)\hat{S}_0(t),$$

where $\hat{S}_0$ and $\hat{S}_1$ are the Kaplan-Meier estimators [Kaplan and Meier, 1958] of the low and high risk of early-readmission subgroups respectively learned by the C-mix model : patients with

$$\pi_{\hat{\beta}}(x_i) > 0.5$$

are clustered in the high risk subgroup, others in the low risk one ; or cured and uncured subgroups respectively learned by the CURE model.

3.B.2 Hyper-parameters tuning

Let us summarize the hyper-parameters obtained after the cross-validation procedure for each method. First, we take $\eta = 0.1$ for all method using elastic net regularization to ensure covariates selection.

The strengh of the penalty is tuned to 42.81 for LR, 0.05 for SVM, 0.03 for C-mix, 0.008 for CURE and 0.014 for Cox PH. For RF, the maximum depth is 7, the minium sample's split is 3, the minimum sample's leaf is 2, the criterion is the entropy and the number of estimators is tuned to 200. For GB, the maximum depth is 7, the minimum sample's split is 3, the minimum sample's leaf is 4 and the number of estimators is 200. Finally for NN, the hidden layer's sizes is 3, the regularization term is tuned to 0.13.

3.B.3 Covariates importance comparison

Figure 3.B.1 gives the covariates importance estimates for all covariates and all considered methods.

FIGURE 3.B.1 Comparison of covariates importance, ordered on the C-mix estimates. Note that for RF and GB models, coefficients are, by construction, always positive.

3.C Competing interests

The authors declare that they have no competing interests. This study received approval from the institutional review board from Georges Pompidou University Hospital (IRB 00001072 - project n° CDW_2014_0008) and the French data protection authority (CNIL - n° 1922081).

Chapitre 4

C-mix : a high dimensional mixture model for censored durations

Sommaire

Abstract. We introduce a supervised learning mixture model for censored durations (C-mix) to simultaneously detect subgroups of patients with different prognosis and order them based on their risk. Our method is applicable in a high-dimensional setting, i.e. with a large number of biomedical covariates. Indeed, we penalize the negative log-likelihood by the elastic net, which leads to a sparse parameterization of the model and automatically pinpoints the relevant covariates for the survival prediction. Inference is achieved using an efficient Quasi-Newton Expectation Maximization (QNEM) algorithm, for which we provide convergence properties. The statistical performance of the method is examined on an extensive Monte Carlo simulation study, and finally illustrated on three publicly available genetic cancer datasets with high-dimensional covariates. We show that our approach outperforms the state-of-the-art survival models in this context, namely both the CURE and Cox proportional hazards models penalized by the elastic net, in terms of C-index, AUC(t) and survival prediction. Thus, we propose a powerfull tool for personalized medicine in cancerology.

Résumé. Nous proposons un modèle de mélange de durées avec censure, le C-mix, qui apprend de façon supervisée à ordonner des patients suivant leur risque qu'un événement d'intérêt se produise rapidement, tout en déterminant des sous-groupes de pronostics différents au sein de la population. Notre méthode s'applique dans un contexte de grande dimension, où le nombre de covariables biomédicales à disposition est grand. Nous pénalisons alors la log-vraisemblance négative par la regularisation elastic net, ce qui impose une paramétrisation parcimonieuse du modèle et permet d'identifier les covariables qui influencent la prédiction du risque. Un algorithme Quasi-Newton EM (QNEM) est proposé pour l'inférence, ainsi qu'une preuve de sa convergence. Les performances du modèle sont évaluées lors d'une étude par simulation de Monte Carlo, puis illustrées sur trois jeux de données publiques de grande dimension en cancérologie. Notre approche obtient de meilleurs résultats en terme de C-index, d'AUC(t) et de prédiction de survie, comparé aux méthodes usuelles d'analyse de survie : à savoir le modèle de CURE et le modèle à risques proportionnels de Cox, tous deux pénalisés par l'elastic net. Le C-mix constitue ainsi un nouvel outil prometteur pour la médecine personnalisée en cancérologie.

4.1 Introduction

Predicting subgroups of patients with different prognosis is a key challenge for personalized medicine, see for instance Alizadeh et al. [2000] and Rosenwald et al. [2002] where subgroups of patients with different survival rates are identified based on gene expression data. A substantial number of techniques can be found in the literature to predict the subgroup of a given patient in a classification setting, namely when subgroups are known in advance [Golub et al., 1999, Hastie et al., 2001a, Tibshirani et al., 2002]. We consider in the present chapter the much more difficult case where subgroups are unknown.

In this situation, a first widespread approach consists in first using unsupervised learning techniques applied on the covariates – for instance on the gene expression data [Bhattacharjee et al., 2001, Beer et al., 2002, Sørlie et al., 2001] – to define subsets of patients and then estimating the risks in each of them. The problem of such techniques is that there is no guarantee that the identified subgroups will have different risks. Another approach to subgroups identification is conversely based exclusively on the survival times : patients are then assigned to a "low-risk" or a "high-risk" group based on whether they were still alive [Shipp et al., 2002, Van't Veer et al., 2002]. The problem here is that the resulting subgroups may not be biologically meaningful since the method do not use the covariates, and prognosis prediction based on covariates is not possible.

The method we propose uses both the survival information of the patients and its covariates in a supervised learning way. Moreover, it relies on the idea that exploiting the subgroups structure of the data, namely the fact that a portion of the population have a higher risk of early death, could improve the survival prediction of future patients (unseen during the learning phase).

We propose to consider a mixture of event times distributions in which the probabilities of belonging to each subgroups are driven by the covariates (*e.g.* gene expression data, patients characteristics, therapeutic strategy or omics covariates). Our C-mix model is hence part of the class of model-based clustering algorithms, as introduced in Banfield and Raftery [1993].

More precisely, to model the heterogeneity within the patient population, we introduce a latent variable

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

and our focus is on the conditional distribution of Z given the values of the covariates $X = x$. Now, conditionally on the latent variable Z, the distribution of duration time T is different, leading to a mixture in the event times distribution.

For a patient with covariates x, the conditional probabilities

$$\pi_k(x) = \mathbb{P}[Z = k|X = x]$$

of belonging to the k-th risk group can be seen as scores, that can help decision-making for physicians. As a byproduct, it can also shed light on the effect of the covariates (which combination of biomedical markers are relevant to a given event of interest).

Our methodology differs from the standard survival analysis approaches in various ways, that we describe in this paragraph. First, the Cox proportional hazards (PH) model (Cox [1972]) (by far the most widely used in such a setting) is a regression model that describes the relation between intensity of events and covariates x via

$$\lambda(t|X = x) = \lambda_0(t)\exp(x^\top \beta^{\text{cox}}), \tag{4.1}$$

where λ_0 is a baseline intensity, and β^{cox} is a vector quantifying the multiplicative impact on the hazard ratio of each covariate. As in our model, high-dimensional covariates can be handled, via *e.g.* penalization , see Simon et al. [2011]. But it does not permit the stratification of the population in groups of homogeneous risks, hence does not deliver a simple tool for clinical practice. Moreover, we show in the numerical sections that the C-mix model can be trained very efficiently in high dimension, and outperforms the standard Cox PH model by far in the analysed datasets.

Other models condiser mixtures of event times distributions. In the CURE model (see Farewell [1982] and Kuk and Chen [1992]), one fraction of the population is considered as cured (hence not subject to any risk). This can be very limitating, as for a large number of applications (*e.g.* rehospitalization for patients with chronic diseases or relapse for patients with metastatic cancer), all patients are at risk. We consider, in our model, that there is always an event risk, no matter how small. Other mixture models have been considered in survival analysis : see Kuo and Peng [2000] for a general study about mixture model for survival data or De Angelis et al. [1999] in a cancer survival analysis setting, to name but a few. Unlike our algorithm, none of these algorithms considers the high dimensional setting.

A precise description of the model is given in Section 4.2. Section 4.3 focuses on the regularized version of the model with an elastic net penalization to exploit dimension reduction and prevent overfitting. Inference is presented under this framework, as well as the convergence properties of the developed algorithm. Section 4.4 highlights the simulation procedure used to evaluate the performances and compares it with state-of-the-art models. In Section 4.5, we apply our method to genetic datasets. Finally, we discuss the obtained results in Section 4.6.

4.2 A censored mixture model

Let us present the survival analysis framework. We assume that, the conditional density of the duration T given $X = x$ is a mixture

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

of $K \geq 1$ densities f_k, for $t \geq 0$ and $\alpha_k \in \mathbb{R}^{d_k}$ some parameters to estimate. The weights combining these distributions depend on the patient biomedical covariates x and are such that

$$\sum_{k=0}^{K-1} \pi_k(x) = 1.$$

This is equivalent to saying that conditionally on a latent variable $Z = k \in \{0, \ldots, K-1\}$, the density of T at time $t \geq 0$ is $f_k(t\,; \alpha_k)$, and we have

$$\mathbb{P}[Z = k|X = x] = \pi_k(x) = \pi_{\beta_k}(x)$$

where

$$\beta_k = (\beta_{k,1}, \ldots, \beta_{k,d}) \in \mathbb{R}^d$$

denotes a vector of coefficients that quantifies the impact of each biomedical covariates on the probability that a patient belongs to the k-th group. Consider a logistic link function for these weights given by

$$\pi_{\beta_k}(x) = \frac{e^{x^\top \beta_k}}{\sum_{k=0}^{K-1} e^{x^\top \beta_k}}. \tag{4.2}$$

The hidden status Z has therefore a multinomial distribution

$$\mathcal{M}\Big(\pi_{\beta_0}(x), \ldots, \pi_{\beta_{K-1}}(x)\Big).$$

The intercept term is here omitted without loss of generality. The graphical model representation of the C-mix is given in Figure 4.1.

In practice, information loss occurs of right censoring type. This is taken into acount in our model by introducing the following : a time $C \geq 0$ when the individual "leaves" the target cohort, a right-censored duration Y and a censoring indicator Δ, defined by

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

where $\min(a, b)$ denotes the minimum between two numbers a and b, and $\mathbb{1}$ denotes the indicator function.

In order to write a likelihood and draw inference, we make the two following hypothesis.

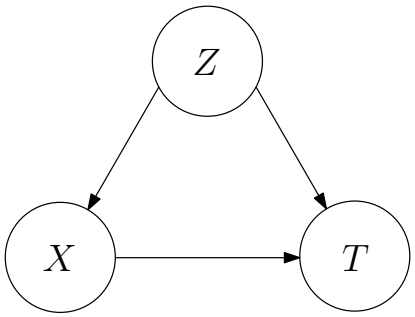


FIGURE 4.1 Graphical model representation of the C-mix.

Hypothesis 1 *T and C are conditionally independent given Z and X.*

Hypothesis 2 *C is independent of Z.*

Hypothesis 1 is classical in survival analysis [Klein and Moeschberger, 2005], while Hypothesis 2 is classical in survival mixture models [Kuo and Peng, 2000, De Angelis et al., 1999]. Under this hypothesis, denoting g the density of the censoring C, F the cumulative distribution function corresponding to a given density f, $\bar{F} = 1 - F$ and

$$F(y^-) = \lim_{\substack{u \to y \\ u \leq y}} F(u),$$

we have

$$\mathbb{P}[Y \leq y, \Delta = 1] = \mathbb{P}[T \leq y, T \leq C] = \int_0^y f(u)\bar{G}(u)\mathrm{d}u \quad \text{and}$$
$$\mathbb{P}[Y \leq y, \Delta = 0] = \mathbb{P}[C \leq y, C < T] = \int_0^y g(u)\bar{F}(u)\mathrm{d}u.$$

Then, denoting

$$\theta = (\alpha_0, \ldots, \alpha_{K-1}, \beta_0, \ldots, \beta_{K-1})^\top$$

the parameters to infer and considering an independent and identically distributed (i.i.d.) cohort of n patients given by

$$(x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n) \in \mathbb{R}^d \times \mathbb{R}_+ \times \{0, 1\},$$

the log-likelihood of the C-mix model can be written

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

where we use the notations $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^\top$. Note that from now on, all computations are done conditionally on the covariates $(x_i)_{i=1,\dots,n}$. An important fact is that we *do not need to know or parametrize* $\bar{G}$ nor g, namely the distribution of the censoring, for inference in this model (since all $\bar{G}$ and g terms vanish in Equation (4.5)).

4.3 Inference of C-mix

In this section, we describe the procedure for estimating the parameters of the C-mix model. We begin by presenting the Quasi-Newton Expectation Maximization (QNEM) algorithm we use for inference. We then focus our study on the convergence properties of the algorithm.

4.3.1 QNEM algorithm

In order to avoid overfitting and to improve the prediction power of our model, we use elastic net regularization [Zou and Hastie, 2005] by minimizing the penalized objective

$$\ell_n^{\text{pen}}(\theta) = -\ell_n(\theta) + \sum_{k=0}^{K-1} \gamma_k \Big((1-\eta) \|\beta_k\|_1 + \frac{\eta}{2} \|\beta_k\|_2^2 \Big), \tag{4.3}$$

where we add a linear combination of the lasso (ℓ_1) and ridge (squared ℓ_2) penalties for a fixed $\eta \in [0,1]$, tuning parameter γ_k, and where we denote

$$\|\beta_k\|_p = \Big(\sum_{i=1}^{d} |\beta_{k,i}|^p \Big)^{1/p}$$

the ℓ_p-norm of β_k. One advantage of this regularization method is its ability to perform model selection (the lasso part) and pinpoint the most important covariates relatively to the prediction objective. On the other hand, the ridge part allows to handle potential correlation between covariates [Zou and Hastie, 2005]. Note that in practice, the intercept is not regularized.

In order to derive an algorithm for this objective, we introduce a so-called Quasi-Newton Expectation Maximization (QNEM), being a combination between an EM algorithm [Dempster et al., 1977] and a L-BFGS-B algorithm [Zhu et al., 1997]. For the EM part, we need to compute the negative completed log-likelihood (here scaled by n^{-1}), namely the negative joint distribution of $\boldsymbol{y}$, $\boldsymbol{\delta}$ and $\boldsymbol{z} = (z_1, \dots, z_n)^\top$. It can

be written

$$\begin{aligned}\ell_n^{\text{comp}}(\theta) &= \ell_n^{\text{comp}}(\theta; \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{z}) \\ &= -n^{-1} \sum_{i=1}^{n} \Big\{ \delta_i \Big[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_i=k\}} \Big(\log \pi_{\beta_k}(x_i) + \log f_k(y_i; \alpha_k) \Big) + \log \bar{G}(y_i^-) \Big] \\ &\quad + (1-\delta_i) \Big[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_i=k\}} \Big(\log \pi_{\beta_k}(x_i) + \log \bar{F}_k(y_i^-; \alpha_k) \Big) + \log g(y_i) \Big] \Big\}.\end{aligned} \tag{4.4}$$

Suppose that we are at step $l+1$ of the algorithm, with current iterate denoted

$$\theta^{(l)} = (\alpha_0^{(l)}, \ldots, \alpha_{K-1}^{(l)}, \beta_0^{(l)}, \ldots, \beta_{K-1}^{(l)})^\top.$$

For the E-step, we need to compute the expected log-likelihood given by

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

We note that

$$q_{i,k}^{(l)} = \mathbb{E}_{\theta^{(l)}}[\mathbb{1}_{\{z_i=k\}} | y_i, \delta_i] = \mathbb{P}_{\theta^{(l)}}[z_i = k | y_i, \delta_i] = \frac{\Lambda_{k,i}^{(l)}}{\sum_{r=0}^{K-1} \Lambda_{r,i}^{(l)}} \tag{4.5}$$

with

$$\Lambda_{k,i}^{(l)} = \Big[f_k(y_i; \alpha_k^{(l)}) \bar{G}(y_i^-) \Big]^{\delta_i} \Big[g(y_i) \bar{F}_k(y_i^-; \alpha_k^{(l)}) \Big]^{1-\delta_i} \pi_{\beta_k^{(l)}}(x_i) \tag{4.6}$$

so that $Q_n(\theta, \theta^{(l)})$ is obtained from (4.4) by replacing the two $\mathbb{1}_{\{z_i=k\}}$ occurrences with $q_{i,k}^{(l)}$. Depending on the chosen distributions f_k, the M-step can either be explicit for the updates of α_k (see Section 4.3.3 below for the geometric distributions case), or obtained using a minimization algorithm otherwise.

Let us focus now on the update of β_k in the M-step of the algorithm. By denoting

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

the quantities involved in Q_n that depend on β_k, the update for β_k therefore requires to minimize

$$R_{n,k}^{(l)}(\beta_k) + \gamma_k \Big((1-\eta) \|\beta_k\|_1 + \frac{\eta}{2} \|\beta_k\|_2^2 \Big). \tag{4.7}$$

The minimization Problem (4.7) is a convex problem. It looks like the logistic regression objective, where labels are not fixed but softly encoded by the expectation step (computation of $q_{i,k}^{(l)}$ above, see Equation (4.5)).

We minimize (4.7) using the well-known L-BFGS-B algorithm [Zhu et al., 1997]. This algorithm belongs to the class of quasi-Newton optimization routines, which

solve the given minimization problem by computing approximations of the inverse Hessian matrix of the objective function. It can deal with differentiable convex objectives with box constraints. In order to use it with ℓ_1 penalization, which is not differentiable, we use the trick borrowed from Andrew and Gao [2007] : for $a \in \mathbb{R}$, write $|a| = a_+ + a_-$, where a_+ and a_- are respectively the positive and negative part of a, and add the constraints $a_+ \geq 0$ and $a_- \geq 0$. Namely, we rewrite the minimization problem (4.7) as the following differentiable problem with box constraints

$$\begin{aligned} &\text{minimize} && R_{n,k}^{(l)}(\beta_k^+ - \beta_k^-) + \gamma_k(1-\eta)\sum_{j=1}^{d}(\beta_{k,j}^+ + \beta_{k,j}^-) + \gamma_k \frac{\eta}{2}\|\beta_k^+ - \beta_k^-\|_2^2 \\ &\text{subject to} && \beta_{k,j}^+ \geq 0 \text{ and } \beta_{k,j}^- \geq 0 \text{ for all } j \in \{1,\dots,d\}, \end{aligned} \tag{4.8}$$

where

$$\beta_k^{\pm} = (\beta_{k,1}^{\pm}, \dots, \beta_{k,d}^{\pm})^\top.$$

The L-BFGS-B solver requires the exact value of the gradient, which is easily given by

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

In Algorithm 1, we describe the main steps of the QNEM algorithm to minimize the function given in Equation (4.3).

Algorithm 1: QNEM Algorithm for inference of the C-mix model

Input: Training data $(x_i, y_i, \delta_i)_{i\in\{1,\dots,n\}}$; starting parameters $(\alpha_k^{(0)}, \beta_k^{(0)})_{k\in\{0,\dots,K-1\}}$; tuning parameters $\gamma_k \geq 0$

Output: Last parameters $(\alpha_k^{(l)}, \beta_k^{(l)})_{k\in\{0,\dots,K-1\}}$

1. **for** $l = 0, \dots,$ *until convergence* **do**
2. Compute $(q_{i,k}^{(l)})_{k\in\{0,\dots,K-1\}}$ using Equation (4.5)
3. Compute $(\alpha_k^{(l+1)})_{k\in\{0,\dots,K-1\}}$
4. Compute $(\beta_k^{(l+1)})_{k\in\{0,\dots,K-1\}}$ by solving (4.8) with the L-BFGS-B algorithm
5. **Return :** $(\alpha_k^{(l)}, \beta_k^{(l)})_{k\in\{0,\dots,K-1\}}$

The penalization parameters γ_k are chosen using cross-validation, see Section 4.A for precise statements about this procedure and about other numerical details.

4.3.2 Convergence to a stationary point

We are addressing here convergence properties of the QNEM algorithm described in Section 4.3.1 for the minimization of the objective function defined in

Equation (4.3). Let us denote

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

Convergence properties of the EM algorithm in a general setting are well known, see Wu [1983]. In the QNEM algorithm, since we only improve $Q_n^{\text{pen}}(\theta, \theta^{(l)})$ instead of a minimization of $Q_n(\theta, \theta^{(l)})$, we are not in the EM algorithm setting but in a so called generalized EM (GEM) algorithm setting [Dempster et al., 1977]. For such an algorithm, we do have the descent property, in the sense that the criterion function given in Equation (4.3) is reduced at each iteration, namely

$$\ell_n^{\text{pen}}(\theta^{(l+1)}) \leq \ell_n^{\text{pen}}(\theta^{(l)}).$$

Let us make two hypothesis.

Hypothesis 3 *The duration densities f_k are such that ℓ_n^{pen} is bounded for all θ.*

Hypothesis 4 *$Q_n^{pen}(\theta, \theta^{(l)})$ is continuous in θ and $\theta^{(l)}$, and for any fixed $\theta^{(l)}$, $Q_n^{pen}(\theta, \theta^{(l)})$ is a convex function in θ and is strictly convex in each coordinate of θ.*

Under Hypothesis 3, $l \mapsto \ell_n^{\text{pen}}(\theta^{(l)})$ decreases monotically to some finite limit. By adding Hypothesis 4, convergence of the QNEM algorithm to a stationary point can be shown. In particular, the stationary point is here a local minimum.

Theorem 4.3.1 *Under Hypothesis 3 and 4, and considering the QNEM algorithm for the criterion function defined in Equation (4.3), every cluster point $\bar{\theta}$ of the sequence $\{\theta^{(l)}; l = 0, 1, 2, \ldots\}$ generated by the QNEM algorithm is a stationary point of the criterion function defined in Equation (4.3).*

A proof is given in Section 4.B.

4.3.3 Parameterization

Let us discuss here the parametrization choices we made in the experimental part. First, in many applications - including the one addressed in Section 4.5 - we are interested in identifying one subgroup of the population with a high risk of adverse event compared to the others. Then, in the following, we consider $Z \in \{0, 1\}$ where $Z = 1$ means high-risk of early death and $Z = 0$ means low risk. Moreover, in such a setting where $K = 2$, one can compare the learned groups by the C-mix and the ones learned by the CURE model in terms of survival curves (see Figure 4.3).

To simplify notations and given the constraint formulated in Equation 4.2, we set $\beta_0 = 0$ and we denote $\beta = \beta_1$ and $\pi_\beta(x)$ the conditional probability that a patient belongs to the group with high risk of death, given its covariates x.

In practice, we deal with discrete times in days. It turns out that the times of the data used for applications in Section 4.5 is well fitted by Weibull distributions. This choice of distribution is very popular in survival analysis, see for instance Klein and Moeschberger [2005]. We then first derive the QNEM algorithm with

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

with here $\alpha_k = (\phi_k, \mu_k) \in (0,1)\times\mathbb{R}_+$, ϕ_k being the scale parameter and μ_k the shape parameter of the distribution.

As explained in the following Section 4.4, we select the best model using a cross-validation procedure based on the C-index metric, and the performances are evaluated according to both C-index and AUC(t) metrics (see Sections 4.4.3 for details). Those two metrics have the following property : if we apply any mapping on the marker vector (predicted on a test set) such that the order between all vector coefficient values is conserved, then both C-index and AUC(t) estimates remain unchanged. In other words, by denoting

$$(M_i)_{i\in\{1,\dots,n_{\text{test}}\}}$$

the vector of markers predicted on a test set of n_{test} individuals, if ψ is a function such that for all $(i,j) \in \{1,\dots,n_{\text{test}}\}^2$, one has

$$M_i < M_j \Rightarrow \psi(M_i) < \psi(M_j),$$

then both C-index and AUC(t) estimates induced by

$$(M_i)_{i\in\{1,\dots,n_{\text{test}}\}}$$

or by

$$\big(\psi(M_i)\big)_{i\in\{1,\dots,n_{\text{test}}\}}$$

are the same.

The order in the marker coefficients is actually paramount when the performances are evaluated according to the mentioned metrics. Furthermore, it turns out that empirically, if we add a constraint on the mixture of Weibull that enforces an *order like* relation between the two distributions f_0 and f_1, the performances are improved. To be more precise, the constraint to impose is that the two density curves do not intersect. We then choose to impose the following : the two scale parameters are equal, *i.e.* $\phi_0 = \phi_1 = \phi$. Indeed under this hypothesis, we do have that

$$\mu_0 < \mu_1 \Rightarrow \forall t \in \mathbb{R}_+, f_0(t;\alpha_0) > f_1(t;\alpha_1)$$

for all $\phi \in (0,1)$.

With this Weibull parameterization, updates for α_k are not explicit in the QNEM algorithm, and consequently require some iterations of a minimization algorithm.

Seeking to have explicit updates for α_k, we then derive the algorithm with geometric distributions instead of Weibull (geometric being a particular case of Weibull with $\mu_k = 1$), namely

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

with $\alpha_k \in (0,1)$.

With this parameterization, we obtain from Equation (4.6)

$$\Lambda_{1,i}^{(l)} = \left[\alpha_1^{(l)}(1-\alpha_1^{(l)})^{y_i-1}\right]^{\delta_i}\left[(1-\alpha_1^{(l)})^{y_i}\right]^{1-\delta_i}\pi_{\beta^{(l)}}(x_i) \quad \text{and}$$
$$\Lambda_{0,i}^{(l)} = \left[\alpha_0^{(l)}(1-\alpha_0^{(l)})^{y_i-1}\right]^{\delta_i}\left[(1-\alpha_0^{(l)})^{y_i}\right]^{1-\delta_i}\left(1-\pi_{\beta^{(l)}}(x_i)\right),$$

which leads to the following explicit M-step

$$\alpha_0^{(l+1)} = \frac{\sum_{i=1}^n \delta_i(1-q_i^{(l)})}{\sum_{i=1}^n (1-q_i^{(l)})y_i} \quad \text{and} \quad \alpha_1^{(l+1)} = \frac{\sum_{i=1}^n \delta_i q_i^{(l)}}{\sum_{i=1}^n q_i^{(l)} y_i}.$$

In this setting, implementation is hence straightforward. Note that Hypothesis 3 and 4 are immediately satisfied with this geometric parameterization.

In Section 4.5, we note that performances are similar for the C-mix model with Weibull or geometric distributions on all considered biomedical datasets. The geometric parameterization leading to more straightforward computations, it is the one used to parameterize the C-mix model in what follows, if not otherwise stated. Let us focus now on the performance evaluation of the C-mix model and its comparison with the Cox PH and CURE models, both regularized with the elastic net.

4.4 Performance evaluation

In this section, we first briefly introduce the models we consider for performance comparisons. Then, we provide details regarding the simulation study and data generation. The chosen metrics for evaluating performances are then presented, followed by the results.

4.4.1 Competing models

The first model we consider is the Cox PH model penalized by the elastic net, denoted Cox PH in the following. In this model introduced in Cox [1972], the partial log-likelihood is given by

$$\ell_n^{\text{cox}}(\beta) = n^{-1}\sum_{i=1}^n \delta_i\Big(x_i^\top\beta - \log \sum_{i':y_{i'}\geq y_i} \exp(x_{i'}^\top\beta)\Big).$$

We use respectively the `R` packages `survival` and `glmnet` [Simon et al., 2011] for the partial log-likelihood and the minimization of the following quantity

$$-\ell_n^{\text{cox}}(\beta) + \gamma\Big((1-\eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2\Big),$$

where γ is chosen by the same cross-validation procedure than the C-mix model, for a given η (see Section 4.A. Ties are handled via the Breslow approximation of the partial likelihood [Breslow, 1972].

We remark that the model introduced in this chapter cannot be reduced to a Cox model. Indeed, the C-mix model intensity can be written (in the geometric case)

$$\lambda(t) = \frac{\alpha_1(1-\alpha_1)^{t-1} + \alpha_0(1-\alpha_0)^{t-1}\ \exp(x^\top\beta)}{(1-\alpha_1)^t + (1-\alpha_0)^t\ \exp(x^\top\beta)},$$

while it is given by Equation (4.1) in the Cox model.

Finally, we consider the CURE [Farewell, 1982] model penalized by the elastic net and denoted CURE in the following, with a logistic function for the incidence part and a parametric survival model for $S(t|Z=1)$, where $Z=0$ means that patient is cured, $Z=1$ means that patient is not cured, and

$$S(t) = \exp(-\int_0^t \lambda(s)\mathrm{d}s)$$

denotes the survival function. In this model, we then have $S(t|Z=0)$ constant and equal to 1. We add an elastic net regularization term, and since we were not able to find any open source package where CURE models were implemented with a regularized objective, we used the QNEM algorithm in the particular case of CURE model. We just add the constraint that the geometric distribution $\mathcal{G}(\alpha_0)$ corresponding to the cured group of patients ($Z=0$) has a parameter $\alpha_0 = 0$, which does not change over the algorithm iterations. The QNEM algorithm can be used in this particular case, were some terms have disapeared from the completed log-likelihood, since in the CURE model case we have

$$\Big\{i \in \{1,\dots,n\} : z_i = 0,\ \delta_i = 1\Big\} = \varnothing.$$

Note that in the original introduction of the CURE model in Farewell [1982], the density of uncured patients directly depends on individual patient covariates, which is not the case here.

We also give additional simulation settings in Section 4.C. First, the case where $d \gg n$, including a comparison of the screening strategy we use in Section 4.5 with the iterative sure independence screening [Fan et al., 2010] (ISIS) method. We also add simulations where data is generated according to the C-mix model with gamma distributions instead of geometric ones, and include the accelerated failure time model [Wei, 1992] (AFT) in the performances comparison study.

4.4.2 Simulation design

In order to assess the proposed method, we perform an extensive Monte Carlo simulation study. Since we want to compare the performances of the 3 models mentioned above, we consider 3 simulation cases for the time distribution : one for each competing model. We first choose a coefficient vector

$$\beta = (\underbrace{\nu, \dots, \nu}_{s}, 0, \dots, 0) \in \mathbb{R}^d,$$

with $\nu \in \mathbb{R}$ being the value of the active coefficients and $s \in \{1, \dots, d\}$ a sparsity parameter. For a desired low-risk patients proportion $\pi_0 \in [0, 1]$, the high-risk patients index set is given by

$$\mathcal{H} = \Big\{ \lfloor (1 - \pi_0) \times n \rfloor \text{ random samples without replacement} \Big\} \subset \{1, \dots, n\},$$

where $\lfloor a \rfloor$ denotes the largest integer less than or equal to $a \in \mathbb{R}$. For the generation of the covariates matrix, we first take

$$[x_{ij}] \in \mathbb{R}^{n \times d} \sim \mathcal{N}(0, \Sigma(\rho)),$$

with $\Sigma(\rho)$ a $(d \times d)$ Toeplitz covariance matrix [Mukherjee and Maiti, 1988] with correlation $\rho \in (0, 1)$. We then add a gap $\in \mathbb{R}^+$ value for patients $i \in \mathcal{H}$ and subtract it for patients $i \notin \mathcal{H}$, only on active covariates plus a proportion $r_{cf} \in [0, 1]$ of the non-active covariates considered as confusion factors, that is

$$x_{ij} \leftarrow x_{ij} \pm \text{gap for } j \in \Big\{ 1, \dots, s, \dots, \lfloor (d - s) r_{cf} \rfloor \Big\}.$$

Note that this is equivalent to generate the covariates according to a Gaussian mixture.

Then we generate

$$Z_i \sim \mathcal{B}\Big(\pi_\beta(x_i)\Big)$$

in the C-mix or CURE simulation case, where $\pi_\beta(x_i)$ is computed given Equation (4.2), with geometric distributions for the durations (see Section 4.3.3). We obtain

$$T_i \sim \mathcal{G}(\alpha_{Z_i})$$

in the C-mix case, and

$$T_i \sim \infty \mathbb{1}_{\{Z_i = 0\}} + \mathcal{G}(\alpha_1) \mathbb{1}_{\{Z_i = 1\}}$$

in the CURE case. For the Cox PH model, we take

$$T_i \sim -\log(U_i) \exp(-x_i^\top \beta),$$

with $U_i \sim \mathcal{U}([0,1])$ and where $\mathcal{U}([a,b])$ stands for the uniform distribution on a segment $[a,b]$.

The distribution of the censoring variable C_i is geometric $\mathcal{G}(\alpha_c)$, with $\alpha_c \in (0,1)$. The parameter α_c is tuned to maintain a desired censoring rate $r_c \in [0,1]$, using a formula given in Section 4.D. The values of the chosen hyper parameters are sumarized in Table 6.1.

TABLE 4.1 Hyper-parameters choice for simulation

η	n	d	s	r_{cf}	ν	ρ	π_0	gap	r_c	α_0	α_1
0.1	100, 200, 500	30, 100	10	0.3	1	0.5	0.75	0.1, 0.3, 1	0.2, 0.5	0.01	0.5

Note that when simulating under the CURE model, the proportion of censored time events is at least equal to π_0 : we then choose $\pi_0 = 0.2$ for the CURE simulations only.

Finally, we want to assess the stability of the C-mix model in terms of variable selection and compare it to the CURE and Cox PH models. To this end, we follow the same simulation procedure explained in the previous lines. For each simulation case, we make vary the two hyper-parameters that impact the most the stability of the variable selection, that is the gap varying in $[0,2]$ and the confusion rate r_{cf} varying in $[0,1]$. All other hyper-parameters are the same than in Table 6.1, except $s = 150$ and with the choice $(n,d) = (200,300)$. For a given hyper-parameters configuration (gap, r_{cf}), we use the following approach to evaluate the variable selection power of the models. Denoting

$$\tilde{\beta}_i = |\hat{\beta}_i| / \max\Big\{|\hat{\beta}_i|, i \in \{1,\dots,d\}\Big\},$$

if we consider that $\tilde{\beta}_i$ is the predicted probability that the true β_i equals ν, then we are in a binary prediction setting and we use the resulting AUC of this problem. Explicit examples of such AUC computations are given in Section 4.E.

4.4.3 Metrics

We detail in this section the metrics considered to evaluate risk prediction performances. Let us denote by M the marker under study. Note that $M = \pi_{\hat{\beta}}(X)$ in the C-mix and the CURE model cases, and $M = \exp(X^\top \hat{\beta}^{\text{cox}})$ in the Cox PH model case. We denote by h the probability density function of marker M, and assume that the marker is measured once at $t = 0$.

For any threshold ξ, cumulative true positive rates and dynamic false positive rates are two functions of time respectively defined as

$$\text{TPR}^{\mathbb{C}}(\xi,t) = \mathbb{P}[M > \xi | T \leq t]$$

and

$$\mathrm{FPR}^{\mathbb{D}}(\xi, t) = \mathbb{P}[M > \xi | T > t].$$

Then, as introduced in Heagerty et al. [2000], the cumulative dynamic time-dependent AUC is defined as follows

$$\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t) = \int_{-\infty}^{\infty} \mathrm{TPR}^{\mathbb{C}}(\xi, t) \left| \frac{\partial \mathrm{FPR}^{\mathbb{D}}(\xi, t)}{\partial \xi} \right| \mathrm{d}\xi,$$

that we simply denote $\mathrm{AUC}(t)$ in the following. We use the Inverse Probability of Censoring Weighting (IPCW) estimate of this quantity with a Kaplan-Meier estimator of the conditional survival function $\mathbb{P}[T > t | M = m]$, as proposed in Blanche et al. [2013] and already implemented in the `R` package `timeROC`.

A common concordance measure that does not depend on time is the C-index [Harrell et al., 1996] defined by

$$\mathcal{C} = \mathbb{P}[M_i > M_j | T_i < T_j],$$

with $i \neq j$ two independent patients (which does not depend on i, j under the i.i.d. sample hypothesis). In our case, T is subject to right censoring, so one would typically consider the modified $\mathcal{C}_\tau$ defined by

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

with τ corresponding to the fixed and prespecified follow-up period duration [Heagerty and Zheng, 2005]. A Kaplan-Meier estimator for the censoring distribution leads to a nonparametric and consistent estimator of $\mathcal{C}_\tau$ [Uno et al., 2011], already implemented in the `R` package `survival`.

Hence in the following, we consider both $\mathrm{AUC}(t)$ and C-index metrics to assess performances.

4.4.4 Results of simulation

We present now the simulation results concerning the C-index metric in the case $(d, r_c) = (30, 0.5)$ in Table 4.2. See Section 4.F for results on other configurations for (d, r_c). Each value is obtained by computing the C-index average and standard deviation (in parenthesis) over 100 simulations. The $\mathrm{AUC}(t)$ average (bold line) and standard deviation (bands) over the same 100 simulations are then given in Figure 4.1, where $n = 100$. Note that the value of the gap can be viewed as a difficulty level of the problem, since the higher the value of the gap, the clearer the separation between the two populations (low risk and high risk patients).

The results measured both by $\mathrm{AUC}(t)$ and C-index lead to the same conclusion : the C-mix model almost always leads to the best results, even under model

misspecification, *i.e.* when data is generated according to the CURE or Cox PH model. Namely, under CURE simulations, C-mix and CURE give very close results, with a strong improvement over Cox PH. Under Cox PH and C-mix simulations, C-mix outperforms both Cox PH and CURE. Surprisingly enough, this exhibits a strong generalization property of the C-mix model, over both Cox PH and CURE. Note that this phenomenon is particularly strong for small gap values, while with an increasing gap (or an increasing sample size n), all procedures barely exhibit the same performance. It can be first explained by the non parametric baseline function in the Cox PH model, and second by the fact that unlike the Cox PH model, the C-mix and CURE models exploit directly the mixture aspect.

Finally, Figure 4.2 gives the results concerning the stability of the variable selection aspect of the competing models. The C-mix model appears to be the best method as well considering the variable selection aspect, even under model misspecification. We notice a general behaviour of our method that we describe in the following, which is also shared by the CURE model only when the data is simulated according to itself, and which justifies the log scale for the gap to clearly distinguish the three following phases.

For very small gap values (less than 0.2), the confusion rate r_{cf} does not impact the variable selection performances, since adding very small gap values to the covariates is almost imperceptible. This means that the resulting AUC is the same when there is no confusion factors and when $r_{cf} = 1$ (that is when there are half active covariates and half confusion ones).

For medium gap values (saying between 0.2 and 1), the confusion factors are more difficult to identify by the model as there number goes up (that is when r_{cf} increases), which is precisely the confusion factors effect we expect to observe.

Then, for large gap values (more than 1), the model succeeds in vanishing properly all confusion factors since the two subpopulations are more clearly separated regarding the covariates, and the problem becomes naturally easier as the gap increases.

TABLE 4.2 Average C-index on 100 simulated data and standard deviation in parenthesis, with $d = 30$ and $r_c = 0.5$. For each configuration, the best result appears in bold.

		Estimation								
		$n = 100$			$n = 200$			$n = 500$		
Simulation	gap	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
C-mix	0.1	**0.786 (0.057)**	0.745 (0.076)	0.701 (0.075)	**0.792 (0.040)**	0.770 (0.048)	0.739 (0.055)	**0.806 (0.021)**	0.798 (0.023)	0.790 (0.024)
	0.3	**0.796 (0.055)**	0.739 (0.094)	0.714 (0.088)	**0.794 (0.036)**	0.760 (0.058)	0.744 (0.055)	**0.801 (0.021)**	0.784 (0.027)	0.783 (0.026)
	1	**0.768 (0.062)**	0.734 (0.084)	0.756 (0.066)	**0.766 (0.043)**	0.736 (0.054)	0.764 (0.042)	0.772 (0.026)	0.761 (0.027)	**0.772 (0.025)**
CURE	0.1	0.770 (0.064)	**0.772 (0.062)**	0.722 (0.073)	**0.790 (0.038)**	**0.790 (0.038)**	0.758 (0.049)	0.798 (0.025)	**0.799 (0.024)**	0.787 (0.025)
	0.3	**0.733 (0.073)**	0.732 (0.072)	0.686 (0.072)	0.740 (0.053)	**0.741 (0.053)**	0.714 (0.060)	**0.751 (0.029)**	**0.751 (0.029)**	0.738 (0.030)
	1	**0.659 (0.078)**	0.658 (0.078)	0.635 (0.070)	**0.658 (0.053)**	**0.658 (0.053)**	0.647 (0.047)	**0.657 (0.031)**	**0.657 (0.031)**	0.656 (0.032)
Cox PH	0.1	**0.940 (0.041)**	0.937 (0.044)	0.850 (0.097)	**0.959 (0.021)**	0.958 (0.020)	0.915 (0.042)	**0.964 (0.012)**	**0.964 (0.012)**	0.950 (0.016)
	0.3	**0.956 (0.030)**	0.955 (0.029)	0.864 (0.090)	**0.966 (0.020)**	0.965 (0.020)	0.926 (0.043)	0.968 (0.013)	**0.969 (0.012)**	0.959 (0.016)
	1	0.983 (0.016)	**0.985 (0.015)**	0.981 (0.019)	0.984 (0.012)	0.985 (0.011)	**0.988 (0.010)**	0.984 (0.007)	0.985 (0.006)	**0.990 (0.005)**

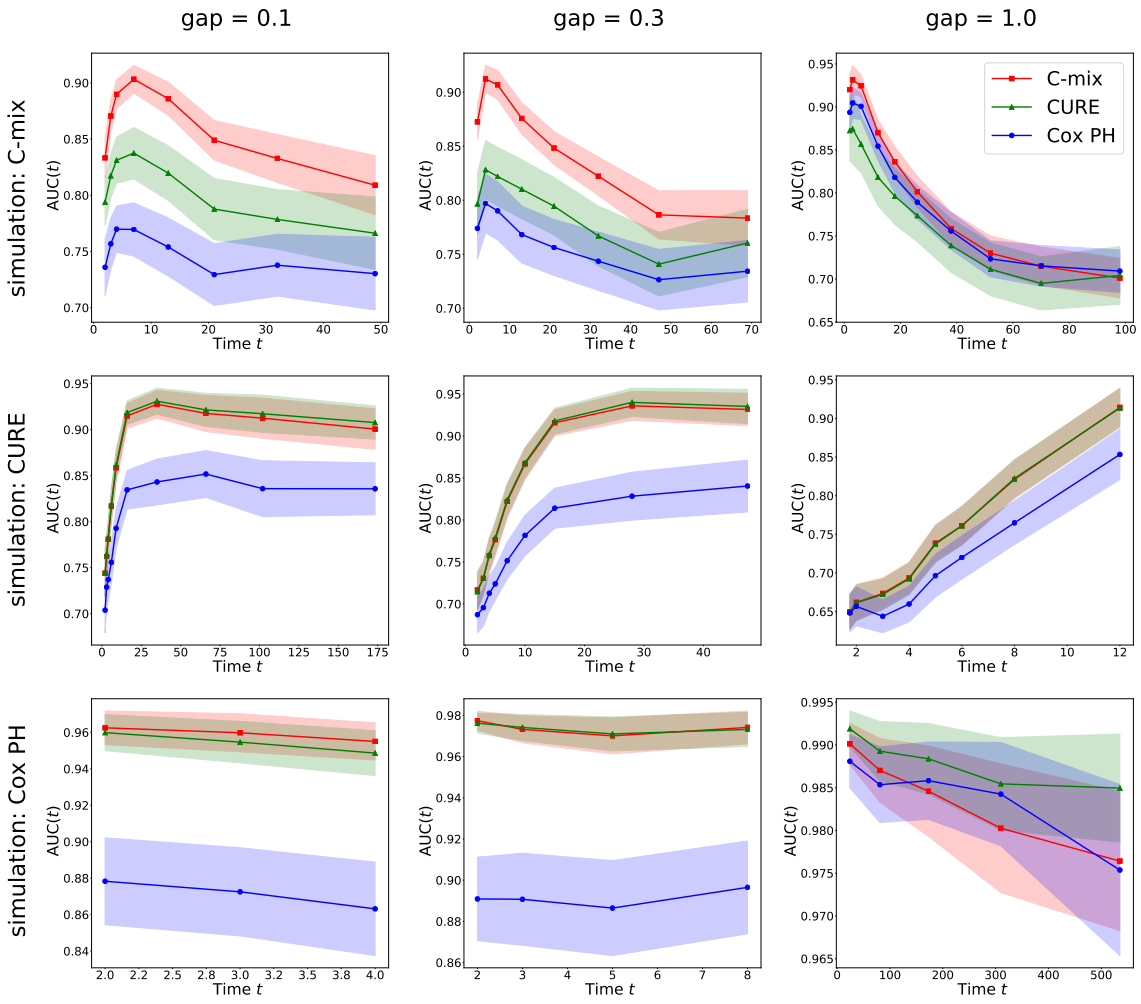


FIGURE 4.1 Average (bold lines) and standard deviation (bands) for AUC(t) on 100 simulated data with $n = 100$, $d = 30$ and $r_c = 0.5$. Rows correspond to the model simulated (cf. Section 4.4.2) while columns correspond to different gap values (the problem becomes more difficult as the gap value decreases). Surprisingly, our method gives almost always the best results, even under model misspecification (see Cox PH and CURE simulation cases on the second and third rows).

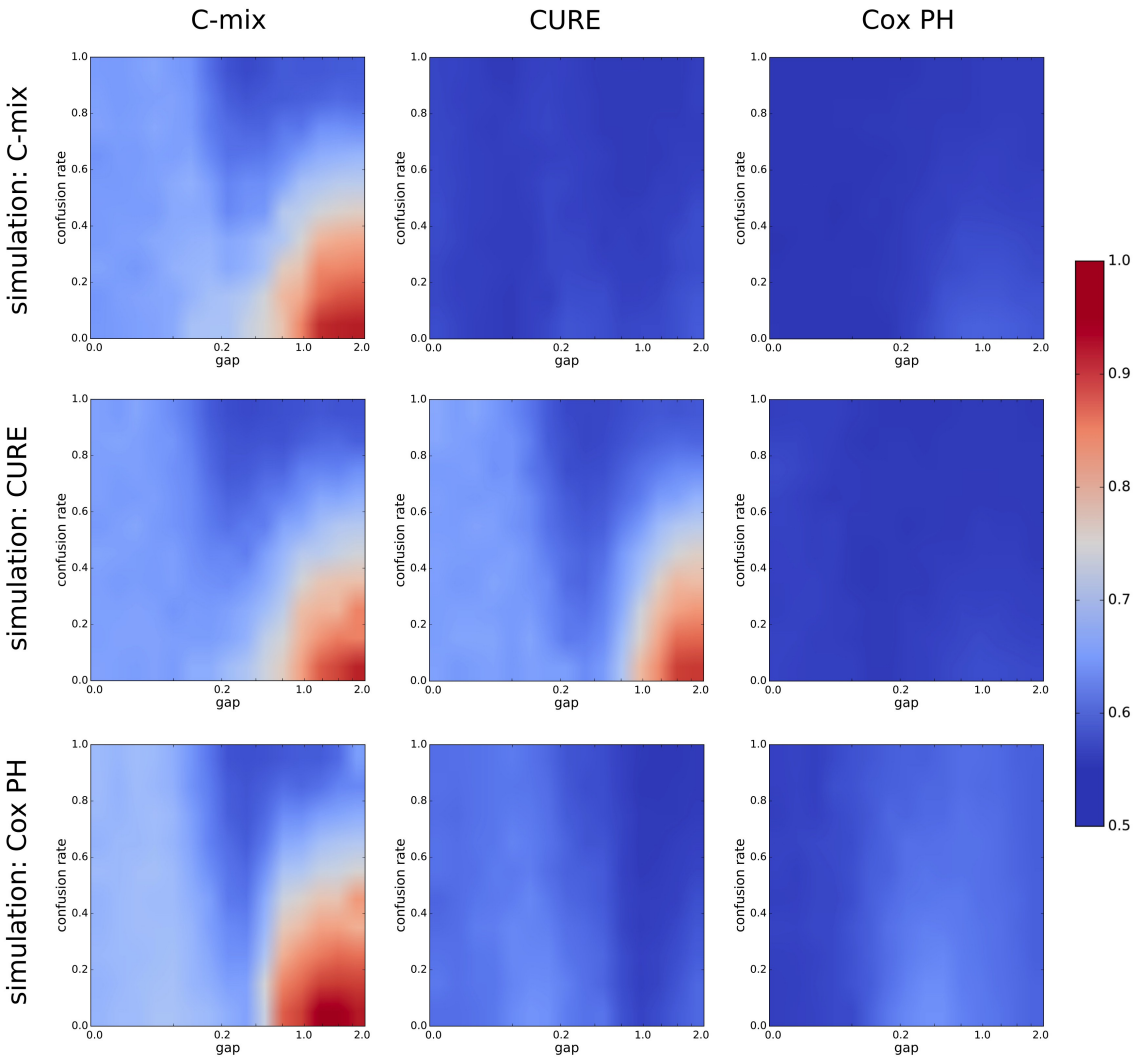


FIGURE 4.2 Average AUC calculated according to Section 4.4.2 and obtained after 100 simulated data for each (gap, r_{cf}) configuration (a grid of 20x20 different configurations is considered). A Gaussian interpolation is then performed to obtain smooth figures. Note that the gap values are *log*-scaled. Rows correspond to the model simulated while columns correspond to the model under consideration for the variable selection evaluation procedure. Our method gives the best results in terms of variable selection, even under model misspecification.

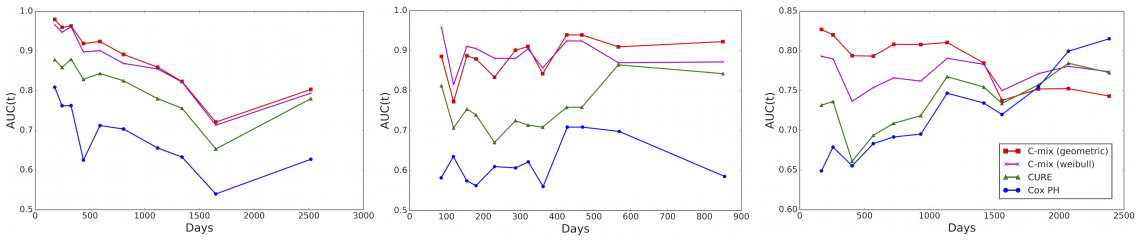
4.5 Application to genetic data

In this section, we apply our method on three genetic datasets and compare its performance to the Cox PH and CURE models. We extracted normalized expression data and survival times Y in days from breast invasive carcinoma (BRCA, $n = 1211$), glioblastoma multiforme (GBM, $n = 168$) and kidney renal clear cell carcinoma (KIRC, $n = 605$).

These datasets are available on The Cancer Genome Atlas (TCGA) platform, which aims at accelerating the understanding of the molecular basis of cancer through the application of genomic technologies, including large-scale genome sequencing. A more precise description of these datasets is given in Section A.2.2. For each patient, 20531 covariates corresponding to the normalized gene expressions are available. We randomly split all datasets into a training set and a test set (30% for testing, cross-validation is done on the training).

We compare the three models both in terms of C-index and AUC(t) on the test sets. Inference of the Cox PH model fails in very high dimension on the considered data with the `glmnet` package. We therefore make a first variable selection (screening) among the 20531 covariates. To do so, we compute the C-index obtained by univariate Cox PH models (not to confer advantage to our method), namely Cox PH models fitted on each covariate separately. We then ordered the obtained 20531 C-indexes by decreasing order and extracted the top $d = 100$, $d = 300$ and $d = 1000$ covariates. We then apply the three methods on the obtained covariates.

The results in terms of AUC(t) curves are given in Figure 4.1 for $d = 300$, where we distinguish the C-mix model with geometric or Weibull distributions.



(a) BRCA (b) GBM (c) KIRC

FIGURE 4.1 AUC(t) comparison on the three TCGA data sets considered, for $d = 300$. We observe that C-mix model leads to the best results (higher is better) and outperforms both Cox PH and CURE in all cases. Results are similar in terms of performances for the C-mix model with geometric or Weibull distributions.

Then it appears that the performances are very close in terms of AUC(t) between the C-mix model with geometric or Weibull distributions, which is also validated if we compare the corresponding C-index for these two parameterizations in Table 4.1.

TABLE 4.1 C-index comparison between geometric or Weibull parameterizations for the C-mix model on the three TCGA data sets considered (with $d = 300$). In all cases, results are very similar for the two distribution choices.

Parameterization		Geometric	Weibull
Cancer	BRCA	0.782	0.780
	GBM	0.755	0.754
	KIRC	0.849	0.835

Similar conclusions in terms of C-index, AUC(t) and computing time can be made on all considered datasets and for any choice of d. Hence, as already mentionned in Section 4.3.3, we only concentrate on the geometric parameterization for the C-mix model. The results in terms of C-index are then given in Table 4.2.

TABLE 4.2 C-index comparison on the three TCGA data sets considered. In all cases, C-mix gives the best results (in bold).

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

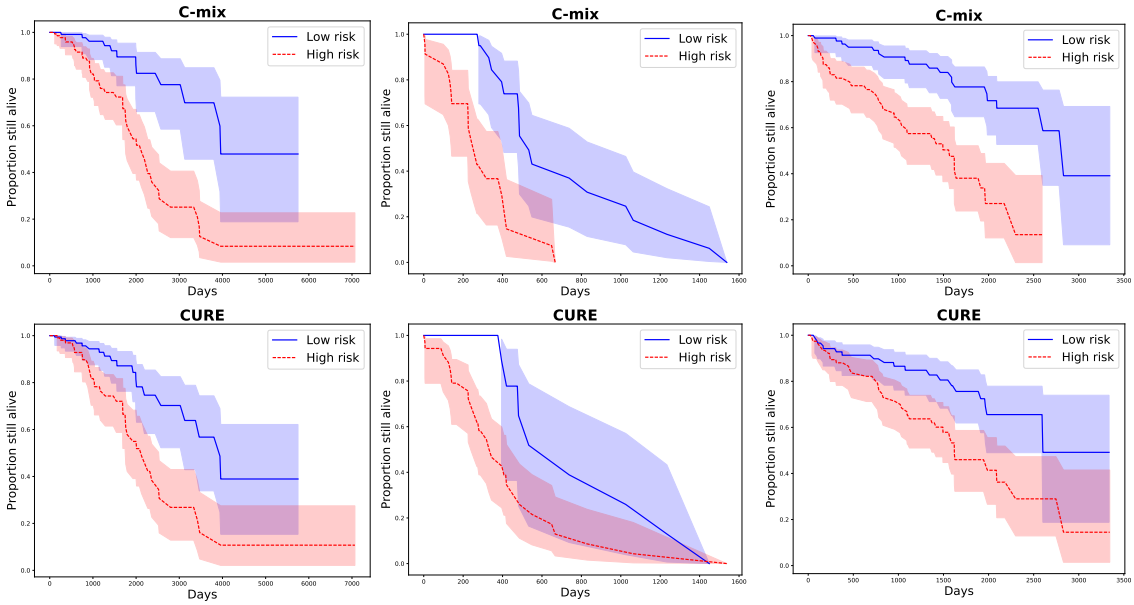
A more direct approach to compare performances between models, rather than only focus on the marker order aspect, is to predict the survival of patients in the test set within a specified short time. For the Cox PH model, the survival $\mathbb{P}[T_i > t|X_i = x_i]$ for patient i in the test set is estimated by

$$\hat{S}_i(t|X_i = x_i) = [\hat{S}_0^{\text{cox}}(t)]^{\exp(x_i^\top \hat{\beta}^{\text{cox}})},$$

where $\hat{S}_0^{\text{cox}}$ is the estimated survival function of baseline population ($x = 0$) obtained using the Breslow estimate of λ_0 [Breslow, 1972]. For the CURE or the C-mix models, it is naturally estimated by

$$\hat{S}_i(t|X_i = x_i) = \pi_{\hat{\beta}}(x_i)\hat{S}_1(t) + \Big(1 - \pi_{\hat{\beta}}(x_i)\Big)\hat{S}_0(t),$$

where $\hat{S}_0$ and $\hat{S}_1$ are the Kaplan-Meier estimators [Kaplan and Meier, 1958] of the low and high risk subgroups respectively, learned by the C-mix or CURE models (patients with $\pi_{\hat{\beta}}(x_i) > 0.5$ are clustered in the high risk subgroup, others in the low risk one). The corresponding estimated survival curves are given in Figure 4.2. We observe that the subgroups obtained by the C-mix are more clearly separated

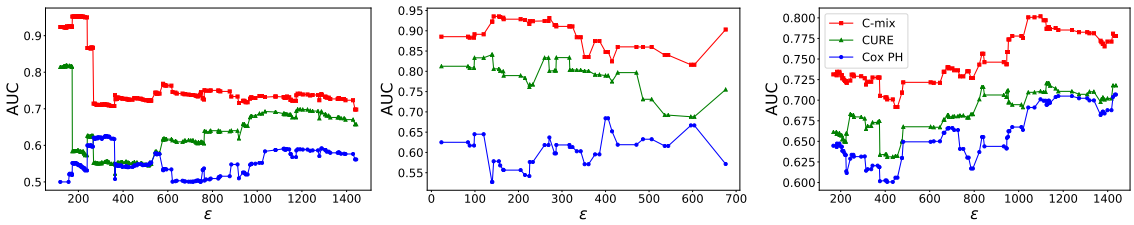


(a) BRCA (b) GBM (c) KIRC

FIGURE 4.2 Estimated survival curves per subgroups (blue for low risk and red for high risk) with the corresponding 95 % confidence bands for the C-mix and CURE models : BRCA in column (a), GBM in column (b) and KIRC in column (c).

in terms of survival than those obtained by the CURE model.

For a given time ϵ, one can now use $\hat{S}_i(\epsilon|X_i = x_i)$ for each model to predict whether or not $T_i > \epsilon$ on the test set, resulting on a binary classification problem that we assess using the classical AUC score. By moving ϵ within the first years of follow-up, since it is the more interesting for physicians in practice, one obtains the curves given in Figure 4.3.



(a) BRCA (b) GBM (c) KIRC

FIGURE 4.3 Comparison of the survival prediction performances between models on the three TCGA data sets considered (still with $d = 300$). Performances are, onces again, much better for the C-mix over the two other standard methods.

Let us now focus on the runtime comparison between the models in Table 4.3. We choose the BRCA dataset to illustrate this point, since it is the larger one ($n = 1211$) and consequently provides more clearer time-consuming differences.

TABLE 4.3 Computing time comparison in second on the BRCA dataset ($n = 1211$), with corresponding C-index in parenthesis and best result in bold in each case. This times concern the learning task for each model with the best hyper parameter selected after the cross validation procedure. It turns out that our method is by far the fastest in addition to providing the best performances. In particular, the QNEM algorithm is faster than the `R` implementation `glmnet`.

Model		C-mix	CURE	Cox PH
d	100	**0.025** (**0.792**)	1.992 (0.764)	0.446 (0.705)
	300	**0.027** (**0.782**)	2.343 (0.753)	0.810 (0.723)
	1000	**0.139** (**0.817**)	12.067 (0.613)	2.145 (0.577)

We also notice that despite using the same QNEM algorithm steps, our CURE model implementation is slower since convergence takes more time to be reached, as shows Figure 4.4. Figure 4.5 provides sample code for the use of our implementation of the C-mix and the CURE models.

In Section 4.G, the top 20 selected genes for each cancer type and for all models are presented (for $d = 300$). Literature on those genes is mined to estimate two simple scores that provide information about how related they are to cancer in general first, and second to cancer plus the survival aspect, according to scientific publications. It turns out that some genes have been widely studied in the literature (*e.g.* FLT3 for the GBM cancer), while for others, very few publications were retrieved (*e.g.* TRMT2B still for the GBM cancer).

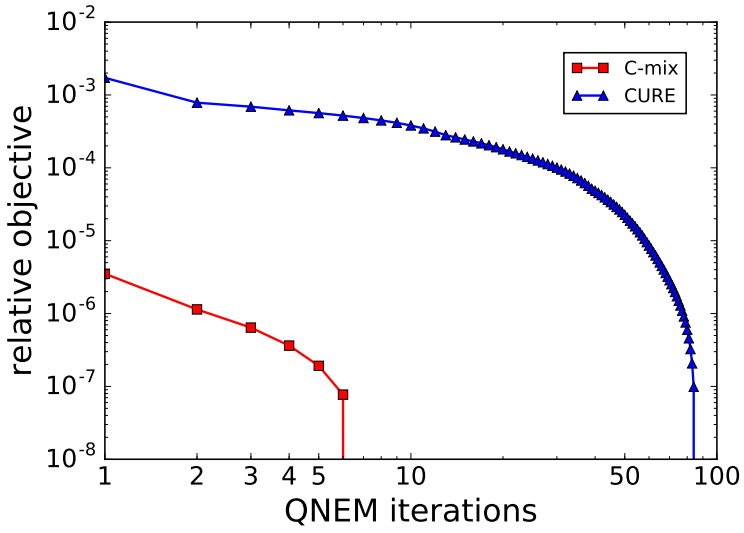


FIGURE 4.4 Convergence comparison between C-mix and CURE models through the QNEM algorithm. The relative objective is here defined at iteration l as $\left(\ell_n^{\text{pen}}(\theta^{(l)}) - \ell_n^{\text{pen}}(\hat{\theta})\right)/\ell_n^{\text{pen}}(\hat{\theta})$, where $\hat{\theta}$ is naturally the parameter vector returned at the end of the QNEM algorithm, that is once convergence is reached. Note that both iteration and relative objective axis are *log*-scaled for clarity. We observe that convergence for the C-mix model is dramaticaly faster than the CURE one.

```python
from QNEM.inference import QNEM
from QNEM.simulation import CensoredGeomMixtureRegression

# Generate data
simu = CensoredGeomMixtureRegression(n_samples=1000, n_features=100,
                                     n_active_features=30)
X, Y, delta = simu.simulate()

# Choose between C-mix or CURE model
model = 'C-mix'

# Fit the model with a penalty strength equal to 10
learner = QNEM(model=model, l_elastic_net=10, eta=.1, max_iter=100,
               tol=1e-6, warm_start=True, fit_intercept=True)
learner.fit(X, Y, delta)

# Obtain the estimated marker
coeffs = learner.coeffs
marker = QNEM.predict_proba(X, fit_intercept, coeffs)
```

FIGURE 4.5 Sample python code for the use of the C-mix.

4.6 Concluding remarks

In this chapter, a mixture model for censored durations (C-mix) has been introduced, and a new efficient estimation algorithm (QNEM) has been derived, that considers a penalization of the likelihood in order to perform covariate selection and to prevent overfitting.

A strong improvement is provided over the CURE and Cox PH approches (both penalized by the elastic net), which are, by far, the most widely used for biomedical data analysis. But more importantly, our method detects relevant subgroups of patients regarding their risk in a supervised learning procedure, and takes advantage of this identification to improve survival prediction over more standard methods. An extensive Monte Carlo simulation study has been carried out to evaluate the performance of the developed estimation procedure. It showed that our approach is robust to model misspecification.

The proposed methodology has then been applied on three high dimensional datasets. On these datasets, C-mix outperforms both Cox PH and CURE, in terms of AUC(t), C-index or survival prediction. Moreover, many gene expressions pinpointed by the feature selection aspect of our regularized method are relevant for medical interpretations (*e.g.* NFKBIA, LEF1, SUSD3 or FAIM3 for the BRCA cancer, see Zhou et al. [2007] or Oskarsson et al. [2011]), whilst others must involve further investigations in the genetic research community. Finally, our analysis provides, as a by-product, a new robust implementation of CURE models in high dimension.

Software

All the methodology discussed in this chapter is implemented in `Python`. The code is available from `https://github.com/SimonBussy/C-mix` in the form of annotated programs, together with a notebook tutorial.

Acknowledgements

The results shown in this chapter are based upon data generated by the TCGA Research Network and freely available from `http://cancergenome.nih.gov/`. *Conflict of Interest* : None declared.

Appendices

4.A Numerical details

Let us first give some details about the starting point of Algorithm 1. For all $k \in \{0, \dots, K-1\}$, we simply use $\beta_k^{(0)}$ as the zero vector, and for $\alpha_k^{(0)}$ we fit a censored parametric mixture model on $(y_i)_{i=1,\dots,n}$ with an EM algorithm.

Concerning the V-fold cross validation procedure for tuning γ_k, we use $V = 5$ and the cross-validation metric is the C-index. Let us precise that we choose γ_k as the largest value such that error is within one standard error of the minimum, and that a grid-search is made during the cross-validation on an interval $[\gamma_k^{\max} \times 10^{-4}, \gamma_k^{\max}]$, with $\gamma_k^{\max}$ the interval upper bound computed in the following.

Let us consider the following convex minimization problem resulting from Equation (8), at a given step l :

$$\hat{\beta}_k \in \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \; R_{n,k}^{(l)}(\beta) + \gamma_k \Big((1-\eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2 \Big).$$

Regarding the grid of candidate values for γ_k, we consider

$$\gamma_k^1 \leq \gamma_k^2 \leq \cdots \leq \gamma_k^{\max}.$$

At $\gamma_k^{\max}$, all coefficients $\hat{\beta}_{k,j}$ for $j \in \{1, \dots, d\}$ are exactly zero. The KKT conditions [Boyd and Vandenberghe, 2004] claim that

$$\begin{cases} \dfrac{\partial R_{n,k}^{(l)}(\hat{\beta}_k)}{\partial \beta_j} = \gamma_k(1-\eta)\ \operatorname{sgn}(\hat{\beta}_{k,j}) + \eta \hat{\beta}_{k,j} & \forall j \in \hat{\mathcal{A}}_k \\ \left| \dfrac{\partial R_{n,k}^{(l)}(\hat{\beta}_k)}{\partial \beta_j} \right| < \gamma_k(1-\eta) & \forall j \notin \hat{\mathcal{A}}_k \end{cases},$$

where

$$\hat{\mathcal{A}}_k = \Big\{ j \in \{1, \dots, d\} : \hat{\beta}_{k,j} \neq 0 \Big\}$$

is the active set of the $\hat{\beta}_k$ estimator, and

$$\operatorname{sgn}(x) = \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x<0\}}$$

for all $x \in \mathbb{R} \setminus \{0\}$.

Then, using (10), one obtains

$$\forall j \in \{1, \dots, d\},\ \hat{\beta}_{k,j} = 0 \Rightarrow \forall j \in \{1, \dots, d\},\ \left| n^{-1} \sum_{i=1}^{n} q_{i,k}^{(l)}\ \frac{1}{2}\ x_{ij} \right| < \gamma_k(1-\eta)$$

Hence, we choose the following upper bound for the grid search interval during the cross-validation procedure

$$\gamma_k^{\max} = \frac{1}{2n(1-\eta)} \max_{j \in \{1,\dots,d\}} \sum_{i=1}^{n} |x_{ij}|.$$

4.B Proof of Theorem 1

Let us denote

$$D = \sum_{k=0}^{K-1} d_k + Kd$$

the number of coordinates of θ so that one can write

$$\theta = (\theta_1, \ldots, \theta_D) = (\alpha_0, \ldots, \alpha_{K-1}, \beta_0, \ldots, \beta_{K-1})^\top \in \Theta \subset \mathbb{R}^D.$$

We denote $\bar{\theta}$ a cluster point of the sequence

$$S = \{\theta^{(l)}; l = 0, 1, 2, \ldots\}$$

generated by the QNEM algorithm, that is

$$\forall \varepsilon > 0, V_\varepsilon(\bar{\theta}) \cap S \setminus \{\bar{\theta}\} \neq \varnothing,$$

with $V_\varepsilon(\bar{\theta})$ the epsilon-neighbourhood of $\bar{\theta}$. We want to prove that $\bar{\theta}$ is a stationary point of the non-differentiable function $\theta \mapsto \ell_n^{\text{pen}}(\theta)$, which means [Tseng, 2001] :

$$\forall r \in \mathbb{R}^D, {\nu_n^{pen}}'(\bar{\theta}; r) = \lim_{\zeta \to 0} \frac{\ell_n^{\text{pen}}(\bar{\theta} + r\zeta) - \ell_n^{\text{pen}}(\bar{\theta})}{\zeta} \geq 0. \tag{4.9}$$

The proof is inspired by Bertsekas [1995]. The conditional density of the complete data given the observed data can be written

$$k(\theta) = \frac{\exp\left(\ell_n^{\text{comp}}(\theta)\right)}{\exp\left(\ell_n(\theta)\right)}.$$

Then, one has

$$\ell_n^{\text{pen}}(\theta) = Q_n^{\text{pen}}(\theta, \theta^{(l)}) - H(\theta, \theta^{(l)}), \tag{4.10}$$

where we introduced

$$H(\theta, \theta^{(l)}) = \mathbb{E}_{\theta^{(l)}}[\log\left(k(\theta)\right)].$$

The key argument relies on the following facts that hold under Hypothesis 3 and 4 :

— $Q_n^{\text{pen}}(\theta, \theta^{(l)})$ is continuous in θ and $\theta^{(l)}$,

— for any fixed $\theta^{(l)}$ (at the $(l+1)$-th M step of the algorithm), $Q_{n,\theta^{(l)}}^{\text{pen}}(\theta)$ is convex in θ and strictly convex in each coordinate of θ.

Let $r \in \mathbb{R}^D$ be an arbitrary direction, then Equations (4.9) and (4.10) yield

$${\ell_n^{\text{pen}}}'(\bar{\theta}; r) = {Q_{n,\bar{\theta}}^{\text{pen}}}'(\bar{\theta}; r) - \langle \bigtriangledown H_{\bar{\theta}}(\bar{\theta}), r \rangle.$$

Hence, by Jensen's inequality we get

$$\forall \theta \in \Theta, H(\theta^{(l)}, \theta^{(l)}) \leq H(\theta, \theta^{(l)}), \tag{4.11}$$

and so $\theta \mapsto H_{\bar{\theta}}(\theta)$ is minimized for $\theta = \theta^{(l)}$, then we have $\bigtriangledown H_{\bar{\theta}}(\bar{\theta}) = 0$. It remains to prove that

$$Q^{\text{pen}}_{n,\bar{\theta}}{}'(\bar{\theta}; r) \geq 0.$$

Let us focus on the proof of the following expression

$$\forall x_1, Q^{\text{pen}}_{n,\bar{\theta}}(\bar{\theta}) \leq Q^{\text{pen}}_{n,\bar{\theta}}(x_1, \bar{\theta}_2, \ldots, \bar{\theta}_D). \tag{4.12}$$

Denoting

$$w_i^{(l)} = (\theta_1^{(l+1)}, \ldots, \theta_i^{(l+1)}, \theta_{i+1}^{(l)}, \ldots, \theta_D^{(l)})$$

and from the definition of the QNEM algorithm, we first have

$$Q^{\text{pen}}_{n,\theta^{(l)}}(\theta^{(l)}) \geq Q^{\text{pen}}_{n,\theta^{(l)}}(w_1^{(l)}) \geq \cdots \geq Q^{\text{pen}}_{n,\theta^{(l)}}(w_{D-1}^{(l)}) \geq Q^{\text{pen}}_{n,\theta^{(l)}}(\theta^{(l+1)}), \tag{4.13}$$

and second for all x_1,

$$Q^{\text{pen}}_{n,\theta^{(l)}}(w_1^{(l)}) \leq Q^{\text{pen}}_{n,\theta^{(l)}}(x_1, \theta_2^{(l)}, \ldots, \theta_D^{(l)}).$$

Consequently, if $(w_1^{(l)})_{l\in\mathbb{N}}$ converges to $\bar{\theta}$, one obtains (4.12) by continuity taking the limit $l \to \infty$. Let us now suppose that $(w_1^{(l)})_{l\in\mathbb{N}}$ does not converge to $\bar{\theta}$, so that $(w_1^{(l)} - \theta^{(l)})_{l\in\mathbb{N}}$ does not converge to 0. Or equivalently : there exists a subsequence $(w_1^{(l_j)} - \theta^{(l_j)})_{j\in\mathbb{N}}$ not converging to 0.

Then, denoting

$$\psi^{(l_j)} = \|w_1^{(l_j)} - \theta^{(l_j)}\|_2,$$

we may assume that there exists $\bar{\psi} > 0$ such that $\forall j \in \mathbb{N}, \psi^{(l_j)} > \bar{\psi}$ by removing from the subsequence $(w_1^{(l_j)} - \theta^{(l_j)})_{j\in\mathbb{N}}$ any terms for which $\psi^{(l_j)} = 0$.

Let

$$s_1^{(l_j)} = \frac{w_1^{(l_j)} - \theta^{(l_j)}}{\psi^{(l_j)}},$$

so that $(s_1^{(l_j)})_{j\in\mathbb{N}}$ belongs to a compact set ($\|s_1^{(l_j)}\| = 1$) and then converges to $\bar{s_1} \neq 0$. Let us fix some $\epsilon \in [0, 1]$, then $0 \leq \epsilon\bar{\psi} \leq \psi^{(l_j)}$. Moreover, $\theta^{(l_j)} + \epsilon\bar{\psi}s_1^{(l_j)}$ lies on the segment joining $\theta^{(l_j)}$ and $w_1^{(l_j)}$, and consequently belongs to Θ since Θ is convex. As $Q^{\text{pen}}_{n,\theta^{(l_j)}}(.)$ is convex and $w_1^{(l_j)}$ minimizes this function over all values that differ from $\theta^{(l_j)}$ along the first coordinate, one has

$$\begin{aligned} Q^{\text{pen}}_{n,\theta^{(l_j)}}(w_1^{(l_j)}) &= Q^{\text{pen}}_{n,\theta^{(l_j)}}(\theta^{(l_j)} + \psi^{(l_j)}s_1^{(l_j)}) \\ &\leq Q^{\text{pen}}_{n,\theta^{(l_j)}}(\theta^{(l_j)} + \epsilon\bar{\psi}s_1^{(l_j)}) \\ &\leq Q^{\text{pen}}_{n,\theta^{(l_j)}}(\theta^{(l_j)}). \end{aligned} \tag{4.14}$$

We finally obtain

$$\begin{aligned}
0 &\leq Q^{\text{pen}}_{n,\theta^{(l_j)}}(\theta^{(l_j)}) - Q^{\text{pen}}_{n,\theta^{(l_j)}}(\theta^{(l_j)} + \epsilon\bar{\psi}s_1^{(l_j)}) \\
&\underset{(4.14)}{\leq} Q^{\text{pen}}_{n,\theta^{(l_j)}}(\theta^{(l_j)}) - Q^{\text{pen}}_{n,\theta^{(l_j)}}(w_1^{(l_j)}) \\
&\underset{(4.13)}{\leq} Q^{\text{pen}}_{n,\theta^{(l_j)}}(\theta^{(l_j)}) - Q^{\text{pen}}_{n,\theta^{(l_j)}}(\theta^{(l_j+1)}) \\
&\underset{(4.10)}{\leq} \ell_n^{\text{pen}}(\theta^{(l_j)}) - \ell_n^{\text{pen}}(\theta^{(l_j+1)}) + \underbrace{H_{\theta^{(l_j)}}(\theta^{(l_j)}) - H_{\theta^{(l_j)}}(\theta^{(l_j+1)})}_{\underset{(4.11)}{\leq} 0} \\
&\leq \ell_n^{\text{pen}}(\theta^{(l_j)}) - \ell_n^{\text{pen}}(\theta^{(l_j+1)}) \underset{j\to+\infty}{\longrightarrow} \ell_n^{\text{pen}}(\bar{\theta}) - \ell_n^{\text{pen}}(\bar{\theta}) = 0
\end{aligned}$$

By continuity of the function $Q_n^{\text{pen}}(x, y)$ in both x and y and taking the limit $j \to \infty$, we conclude that

$$Q^{\text{pen}}_{n,\bar{\theta}}(\bar{\theta} + \epsilon\bar{\psi}\bar{s_1}) = Q^{\text{pen}}_{n,\bar{\theta}}(\bar{\theta})$$

for all $\epsilon \in [0, 1]$.

Since $\bar{\psi}\bar{s_1} \neq 0$, this contradicts the strict convexity of the function

$$x_1 \mapsto Q^{\text{pen}}_{n,\theta^{(l)}}(x_1, \theta_2^{(l)}, \ldots, \theta_D^{(l)})$$

and establishes that $(w_1^{(l)})_{l\in\mathbb{N}}$ converges to $\bar{\theta}$.

Hence (4.12) is proved. Repeating the argument to each coordinate, we deduce that $\bar{\theta}$ is a coordinate-wise minimum, and finally conclude that $\ell_n^{\text{pen}\,\prime}(\bar{\theta}; r) \geq 0$ [Tseng, 2001]. Thus, $\bar{\theta}$ is a stationary point of the criterion function defined in Equation (4).

□

4.C Additional comparisons

In this section, we consider two extra simulation settings. First, we consider the case $d \gg n$, which is the setting of our application on TCGA datasets. Then, we add another simulation case under the C-mix model using gamma distributions instead of geometric ones. The shared parameters in the two cases are given in Table 6.1.

TABLE 4.C.1 Hyper-parameters choice for simulation.

η	n	s	r_{cf}	ν	ρ	π_0	gap	r_c
0.1	250	50	0.5	1	0.5	0.75	0.1	0.5

4.C.1 Case $d \gg n$

Data is here generated under the C-mix model with

$$(\alpha_0, \alpha_1) = (0.1, 0.5)$$

and

$$d \in \{200, 500, 1000\}.$$

The 3 models are trained on a training set and risk prediction is made on a test set. We also compare the 3 models when a dimension reduction step is performed at first, using two different screening methods. The first is based on univariate Cox PH models, namely the one we used in Section 4.5 of the chapter (in our application to genetic data), where we select here the top 100 variables. This screening method is hence referred as "top 100" in the following.

The second is the iterative sure independence screening (ISIS) method introduced in Fan et al. [2010], using the `R` package `SIS` [Saldana and Feng, 2016]. Prediction performances are compared in terms of C-index, while variable selection performances are compared in terms of AUC using the method detailed in Section 4.E, and we also add two more classical scores [Fan et al., 2010] for comparison : the median ℓ_1 and squared ℓ_2 estimation errors, given by $\|\beta - \hat{\beta}\|_1$ and $\|\beta - \hat{\beta}\|_2$ respectively. Results are given in Table 4.C.2.

The C-mix model obtains constantly the best C-index performances in prediction, for all settings. Moreover, the "top 100" screening method improve the 3 models prediction power, while ISIS method only improve the Cox PH model prediction power. As expected, ISIS method significantly improve the Cox PH model in terms of variable selection and obtains the best results for $d = 500$ and 1000. Conclusions in terms of variable selection are the same relatively to the AUC, ℓ_1 and squared ℓ_2 estimation errors. Then, in the chapter, we only focus on the AUC method detailed in Section 4.E.

Note that the Cox PH model obtains the best results in terms of variable selection with the two screening method, since both screening methods are based on the Cox PH model. Thus, one could improve the C-mix variable selection performances by simply use the "top 100" screening method with univariate C-mix, which was not the purpose of the section. Finally, the results obtained justify the screening strategy we use in Section 4.5 of the chapter.

TABLE 4.C.2 Average performances and standard deviation (in parenthesis) on 100 simulated data for different dimension d and different screening method (including no screening). For each configuration, the best result appears in bold.

d	screening	model	C-index	AUC	$\|\beta - \hat{\beta}\|_1$	$\|\beta - \hat{\beta}\|_2$
200	none	C-mix	**0.716 (0.062)**	**0.653 (0.053)**	**51.540 (0.976)**	**7.254 (0.129)**
		CURE	0.701 (0.067)	0.625 (0.052)	51.615 (1.275)	7.274 (0.122)
		Cox PH	0.672 (0.089)	0.608 (0.063)	199.321 (0.490)	99.679 (0.229)
	top 100	C-mix	**0.737 (0.057)**	**0.682 (0.060)**	**52.297 (1.351)**	**7.381 (0.161)**
		CURE	0.714 (0.060)	0.651 (0.050)	52.366 (1.382)	7.386 (0.134)
		Cox PH	0.692 (0.089)	0.630 (0.070)	52.747 (0.530)	7.946 (0.093)
	ISIS	C-mix	**0.691 (0.049)**	0.570 (0.011)	55.493 (1.624)	8.083 (0.394)
		CURE	0.685 (0.050)	0.571 (0.009)	54.461 (1.112)	7.848 (0.211)
		Cox PH	0.690 (0.049)	**0.573 (0.011)**	**48.186 (0.366)**	**6.840 (0.037)**
500	none	C-mix	**0.710 (0.058)**	**0.642 (0.057)**	**51.627 (0.994)**	7.277 (0.106)
		CURE	0.675 (0.057)	0.610 (0.052)	51.920 (2.411)	**7.252 (0.138)**
		Cox PH	0.624 (0.097)	0.567 (0.057)	499.610 (0.396)	157.997 (0.117)
	top 100	C-mix	**0.735 (0.050)**	**0.694 (0.057)**	53.161 (1.708)	7.433 (0.152)
		CURE	0.703 (0.054)	0.649 (0.042)	53.419 (1.818)	7.387 (0.133)
		Cox PH	0.682 (0.087)	0.633 (0.074)	**49.465 (0.428)**	**6.937 (0.094)**
	ISIS	C-mix	**0.677 (0.051)**	0.559 (0.013)	55.229 (1.831)	7.974 (0.375)
		CURE	0.671 (0.051)	0.559 (0.015)	54.187 (1.244)	7.754 (0.227)
		Cox PH	0.675 (0.051)	**0.560 (0.016)**	**48.574 (0.614)**	**6.870 (0.054)**
1000	none	C-mix	**0.694 (0.063)**	**0.633 (0.066)**	**51.976 (1.921)**	**7.272 (0.141)**
		CURE	0.657 (0.067)	0.598 (0.057)	52.078 (2.414)	7.236 (0.138)
		Cox PH	0.579 (0.092)	0.541 (0.050)	999.768 (0.316)	223.558 (0.067)
	top 100	C-mix	**0.726 (0.050)**	**0.693 (0.040)**	53.813 (1.592)	7.149 (0.115)
		CURE	0.685 (0.061)	0.653 (0.037)	54.146 (1.596)	7.383 (0.090)
		Cox PH	0.688 (0.076)	0.668 (0.064)	**52.838 (0.558)**	**6.909 (0.077)**
	ISIS	C-mix	**0.653 (0.062)**	0.553 (0.017)	53.760 (1.949)	7.269 (0.395)
		CURE	0.652 (0.061)	**0.554 (0.015)**	53.928 (1.288)	7.687 (0.236)
		Cox PH	0.652 (0.063)	0.553 (0.015)	**51.826 (0.606)**	**6.895 (0.054)**

4.C.2 Case of times simulated with a mixture of gammas

We consider here the case where data is simulated under the C-mix model with gamma distributions instead of geometric ones, not to confer to the C-mix prior information on the underlying survival distributions. Hence, one has

$$f_k(t; \iota_k, \zeta_k) = \frac{t^{\iota_k - 1} e^{-\frac{t}{\zeta_k}}}{\zeta_k^{\iota_k} \Gamma(\iota_k)},$$

with ι_k the shape parameter, ζ_k the scale parameter and Γ the gamma function. For the simulations, we choose $(\iota_0, \zeta_0) = (5, 3)$ and $(\iota_1, \zeta_1) = (1.5, 1)$, so that density and survival curves are comparable with those in Section 4.C.1, as illustrates Figure 4.C.1 below.

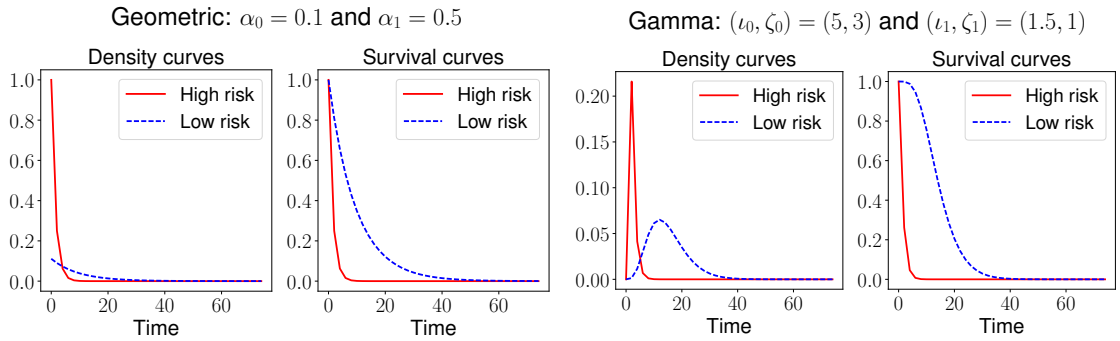


FIGURE 4.C.1 Comparison of the density and survival curves of geometrics laws used in Section 4.C.1 and those used in this section. The supports are then relatively close.

We also add another class of model for comparison in this context : the accelerated failure time model [Wei, 1992] (AFT) ; which can be viewed as a parametric Cox model. Indeed, the semi-parametric property of the Cox PH model could lower its performances compared to completely parametric models such as C-mix and CURE ones, especially in simulations where n is relatively small. We use the `R` package `AdapEnetClass` that implements AFT in a high dimensional setting using two elastic net regularization approaches [Khan and Shaw, 2016] : the adaptive elastic net (denoted AEnet in the following) and the weighted elastic net (denoted WEnet in the following). Results are given in Table 4.C.3 using the same metrics that in Section 4.C.1.

Hence, the C-mix model still gets the best results, both in terms of risk prediction and variable selection. Note that AFT with AEnet and WEnet outperforms the Cox model regularized by the elastic net when $d = 1000$, but is still far behind the C-mix performances.

TABLE 4.C.3 Average performances and standard deviation (in parenthesis) on 100 simulated data for different dimension d with the times simuted with a mixture of gammas. For each configuration, the best result appears in bold.

d	model	C-index	AUC	$\|\beta - \hat{\beta}\|_1$	$\|\beta - \hat{\beta}\|_2$
200	C-mix	**0.701 (0.090)**	**0.659 (0.083)**	**51.339 (2.497)**	**7.186 (0.281)**
	CURE	0.682 (0.058)	0.609 (0.037)	51.563 (1.071)	7.263 (0.097)
	Cox PH	0.664 (0.085)	0.605 (0.065)	199.337 (0.493)	99.686 (0.231)
	AEnet	0.631 (0.062)	0.577 (0.046)	54.651 (2.328)	7.713 (0.426)
	WEnet	0.620 (0.061)	0.544 (0.030)	58.861 (4.298)	8.568 (0.851)
500	C-mix	**0.704 (0.100)**	**0.651 (0.084)**	52.416 (2.311)	7.357 (0.231)
	CURE	0.687 (0.057)	0.609 (0.038)	**52.041 (1.667)**	**7.262 (0.096)**
	Cox PH	0.621 (0.101)	0.559 (0.057)	499.677 (0.381)	158.017 (0.113)
	AEnet	0.604 (0.061)	0.557 (0.030)	55.126 (1.693)	7.616 (0.316)
	WEnet	0.594 (0.065)	0.535 (0.021)	59.736 (2.777)	8.438 (0.626)
1000	C-mix	**0.684 (0.097)**	**0.638 (0.088)**	**52.557 (3.746)**	7.331 (0.277)
	CURE	0.658 (0.057)	0.603 (0.044)	53.120 (3.853)	**7.273 (0.165)**
	Cox PH	0.580 (0.092)	0.538 (0.053)	999.785 (0.334)	223.561 (0.071)
	AEnet	0.586 (0.058)	0.541 (0.024)	54.597 (1.312)	7.495 (0.299)
	WEnet	0.583 (0.054)	0.525 (0.017)	58.746 (2.260)	8.150 (0.551)

4.D Tuning of the censoring level

Suppose that we want to generate data following the procedure detailed in Section 4.4.2, in the C-mix with geometric distributions or CURE case. The question here is to choose α_c for a desired censoring rate r_c, and for some fixed parameters α_0, α_1 and π_0. We write

$$\begin{aligned} 1 - r_c = \mathbb{E}[\Delta] &= \sum_{k=0}^{+\infty}\sum_{j=1}^{+\infty}\Big[\alpha_0(1-\alpha_0)^{j-1}\pi_0 + \alpha_1(1-\alpha_1)^{j-1}(1-\pi_0)\Big]\alpha_c(1-\alpha_c)^{j+k-1} \\ &= \frac{\alpha_0\pi_0\Big[1-(1-\alpha_1)(1-\alpha_c)\Big] + \alpha_1(1-\pi_0)\Big[1-(1-\alpha_0)(1-\alpha_c)\Big]}{\Big[1-(1-\alpha_0)(1-\alpha_c)\Big]\Big[1-(1-\alpha_1)(1-\alpha_c)\Big]}. \end{aligned}$$

See Appendix A.2.4 for details on the calculus derivation. Then, if we denote $\bar{r}_c = 1 - r_c$, $\bar{\alpha}_c = 1 - \alpha_c$, $\bar{\alpha}_0 = 1 - \alpha_0$, $\bar{\alpha}_1 = 1 - \alpha_1$ and $\bar{\pi}_0 = 1 - \pi_0$, we can choose α_c for a fixed r_c by solving the following quadratic equation

$$(\bar{r}_c\ \bar{\alpha}_0\ \bar{\alpha}_1)\bar{\alpha}_c^2 + \Big(\alpha_0\pi_0\bar{\alpha}_1 + \alpha_1\bar{\pi}_0\ \bar{\alpha}_0 - \bar{r}_c(\bar{\alpha}_1 + \bar{\alpha}_0)\Big)\bar{\alpha}_c + (r_c - \alpha_0\pi_0 - \alpha_1\bar{\pi}_0) = 0,$$

for which one can prove that there is always a unique root in $(0, 1)$.

4.E Details on variable selection evaluation

Let us recall that the true underlying β used in the simulations is given by

$$\beta = (\underbrace{\nu, \dots, \nu}_{s}, 0, \dots, 0) \in \mathbb{R}^d,$$

with s the sparsity parameter, being the number of "active" variables. To illustrate how we assess the variable selection ability of the considered models, we give in Figure 4.E.1 an example of β with $d = 100$, $\nu = 1$ and $s = 30$. We simulate data according to this vector (and to the C-mix model) with two different (gap, r_{cf}) values : $(0.2, 0.7)$ and $(1, 0.3)$. Then, we give the two corresponding estimated vectors $\hat{\beta}$ learned by the C-mix on this data.

Denoting

$$\tilde{\beta}_i = |\hat{\beta}_i| / \max\Big\{|\hat{\beta}_i|, i \in \{1, \dots, d\}\Big\}$$

we consider that $\tilde{\beta}_i$ is the predicted probability that the true coefficient β_i corresponding to i-th covariate equals ν. Then, we are in a binary prediction setting where each $\tilde{\beta}_i$ predicts $\beta_i = \nu$ for all $i \in \{1, \dots, d\}$. We use the resulting AUC to assess the variable selection obtained through $\hat{\beta}$.

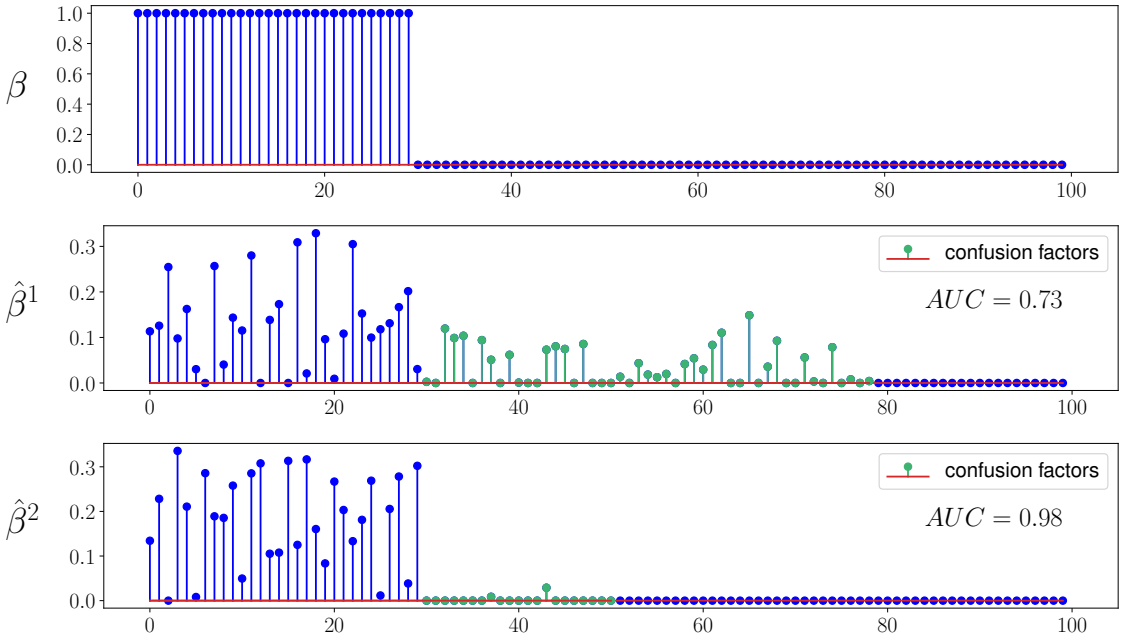


FIGURE 4.E.1 Illustration of the variable selection evaluation procedure. $\hat{\beta}^1$ is learned by the C-mix according to data generated with β and $(\text{gap}, r_{cf}) = (0.2, 0.7)$. We observe that using this gap value to generate data, the model does not succeed to completely vanish the confusion variables (being 70% of the non-active variables, represented in green color), while all other non-active variables are vanished. The corresponding AUC scrore of feature selection is 0.73. $\hat{\beta}^2$ is learned by the C-mix according to data generated with β and $(\text{gap}, r_{cf}) = (1, 0.3)$. The confusion variables are here almost all detected and the corresponding AUC scrore of feature selection is 0.98.

4.F Extended simulation results

Table 4.F.1 bellow presents the results of simulation for the configurations $(d, r_c) = (30, 0.2)$, (100,0.2) and (100,0.5).

TABLE 4.F.1 Average C-index and standard deviation (in parenthesis) on 100 simulated data for different configurations (d, r_c), with geometric distributions for the C-mix model. For each configuration, the best result appears in bold.

$(d, r_c) = (30, 0.2)$

		Estimation								
		$n = 100$			$n = 200$			$n = 500$		
Simulation	gap	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
C-mix	0.1	**0.753 (0.055)**	0.637 (0.069)	0.658 (0.081)	**0.762 (0.034)**	0.664 (0.070)	0.704 (0.051)	**0.767 (0.023)**	0.686 (0.062)	0.749 (0.025)
	0.3	**0.756 (0.050)**	0.599 (0.073)	0.657 (0.075)	**0.761 (0.033)**	0.600 (0.064)	0.713 (0.050)	**0.757 (0.020)**	0.565 (0.049)	0.740 (0.021)
	1	**0.723 (0.059)**	0.710 (0.063)	0.714 (0.062)	**0.723 (0.042)**	0.718 (0.044)	0.721 (0.040)	**0.727 (0.026)**	0.723 (0.028)	0.726 (0.025)
Cox PH	0.1	**0.918 (0.042)**	0.872 (0.070)	0.850 (0.081)	**0.938 (0.022)**	0.911 (0.032)	0.906 (0.034)	**0.949 (0.014)**	0.940 (0.018)	0.938 (0.017)
	0.3	**0.935 (0.034)**	0.906 (0.051)	0.877 (0.066)	**0.947 (0.019)**	0.932 (0.028)	0.915 (0.030)	**0.952 (0.013)**	0.950 (0.015)	0.949 (0.015)
	1	0.956 (0.031)	**0.958 (0.032)**	0.919 (0.065)	0.960 (0.018)	**0.969 (0.016)**	0.951 (0.024)	0.958 (0.011)	**0.968 (0.011)**	0.967 (0.010)

$(d, r_c) = (100, 0.2)$

		Estimation								
		$n = 100$			$n = 200$			$n = 500$		
Simulation	gap	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
C-mix	0.1	**0.736 (0.048)**	0.601 (0.081)	0.656 (0.066)	**0.757 (0.037)**	0.629 (0.079)	0.697 (0.057)	**0.767 (0.020)**	0.659 (0.073)	0.744 (0.024)
	0.3	**0.733 (0.056)**	0.582 (0.063)	0.648 (0.073)	**0.757 (0.035)**	0.572 (0.047)	0.699 (0.057)	**0.758 (0.023)**	0.558 (0.040)	0.736 (0.031)
	1	**0.723 (0.067)**	0.717 (0.073)	0.705 (0.063)	**0.721 (0.041)**	0.716 (0.041)	0.719 (0.046)	0.724 (0.023)	0.720 (0.025)	**0.726 (0.023)**
Cox PH	0.1	**0.892 (0.047)**	0.818 (0.086)	0.830 (0.085)	**0.935 (0.026)**	0.896 (0.048)	0.904 (0.041)	**0.948 (0.013)**	0.935 (0.021)	0.940 (0.015)
	0.3	**0.914 (0.042)**	0.858 (0.076)	0.869 (0.077)	**0.937 (0.025)**	0.909 (0.038)	0.917 (0.030)	**0.957 (0.011)**	0.951 (0.014)	0.951 (0.012)
	1	0.921 (0.040)	**0.937 (0.036)**	0.917 (0.045)	0.918 (0.033)	0.947 (0.035)	**0.951 (0.024)**	0.915 (0.018)	0.959 (0.022)	**0.964 (0.011)**

$(d, r_c) = (100, 0.5)$

		Estimation								
		$n = 100$			$n = 200$			$n = 500$		
Simulation	gap	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
C-mix	0.1	**0.773 (0.064)**	0.710 (0.087)	0.678 (0.078)	**0.798 (0.038)**	0.767 (0.057)	0.744 (0.055)	**0.804 (0.022)**	0.795 (0.024)	0.788 (0.025)
	0.3	**0.781 (0.057)**	0.696 (0.103)	0.697 (0.087)	**0.798 (0.034)**	0.741 (0.064)	0.741 (0.055)	**0.800 (0.021)**	0.778 (0.036)	0.785 (0.023)
	1	**0.772 (0.064)**	0.742 (0.081)	0.760 (0.071)	**0.772 (0.044)**	0.732 (0.074)	0.771 (0.041)	0.770 (0.028)	0.740 (0.059)	**0.771 (0.029)**
CURE	0.1	0.755 (0.070)	**0.759 (0.068)**	0.692 (0.082)	0.780 (0.044)	**0.782 (0.043)**	0.752 (0.052)	**0.795 (0.025)**	**0.795 (0.025)**	0.785 (0.026)
	0.3	0.730 (0.077)	**0.737 (0.076)**	0.674 (0.086)	0.740 (0.042)	**0.740 (0.041)**	0.708 (0.055)	0.753 (0.028)	**0.753 (0.027)**	0.740 (0.031)
	1	**0.663 (0.075)**	0.660 (0.076)	0.659 (0.064)	0.661 (0.053)	**0.661 (0.052)**	0.658 (0.050)	**0.657 (0.032)**	0.657 (0.033)	0.657 (0.034)
Cox PH	0.1	0.916 (0.069)	**0.924 (0.056)**	0.837 (0.097)	**0.950 (0.028)**	0.949 (0.029)	0.911 (0.052)	**0.964 (0.012)**	**0.964 (0.012)**	0.951 (0.016)
	0.3	**0.937 (0.047)**	0.934 (0.050)	0.863 (0.071)	0.955 (0.026)	**0.956 (0.022)**	0.925 (0.037)	**0.968 (0.012)**	**0.968 (0.012)**	0.958 (0.015)
	1	0.963 (0.029)	**0.967 (0.027)**	0.973 (0.024)	0.966 (0.019)	0.970 (0.017)	**0.984 (0.012)**	0.962 (0.012)	0.966 (0.011)	**0.988 (0.006)**

4.G Selected genes per model on the TCGA datasets

In Tables 4.G.1, 4.G.2 and 4.G.3 hereafter, we detail the 20 most significant covariates for each model and for the three considered datasets. For each selected gene, we precise the corresponding effect in percentage, where we define the effect of covariate j as

$$100 \times |\beta_j| \ / \ \|\beta\|_1 \ \%.$$

Then, to explore physiopathological and epidemiological background that could explain the role of the selected genes in cancer prognosis, we search in MEDLINE the number of publications for different requests : (1) selected gene name (*e.g.* UBTF), (2) selected gene name and cancer (*e.g.* UBTF AND cancer[MesH]), (3) selected gene name and cancer survival (*e.g.* UBTF AND cancer[MesH] AND survival).

We then estimate f_1 defined here as the frequency of publication dealing with cancer among all publications for this gene, that is

$$\mathrm{f}_1 = \frac{(2)}{(1)},$$

and f_2 defined as the frequency of publication dealing with survival among publications dealing with cancer, that is

$$\mathrm{f}_2 = \frac{(3)}{(2)}.$$

A f_1 (respectively f_2) close to 1 just informs that the corresponding gene is well known to be highly related to cancer (respectively to cancer survival) by the genetic research community. Note that the CURE and Cox PH models tend to have a smaller support than the C-mix one, since they tend to select less than 20 genes.

Let us precise that our search was performed on the 15th september 2016 at http://www.nlm.nih.gov/bsd/pmresources.html.

TABLE 4.G.1 Top 20 selected genes per model for the BRCA cancer, with the corresponding effects. Dots (·) mean zeros.

Genes	Model effects (%)			MEDLINE data		
	C-mix	CURE	Cox PH	(1)	f_1	f_2
PHKB|5257	9.8	7.2	4.3	1079	0.20	0.37
UBTF|7343	7.8	5.8	21.7	14	0,21	·
LOC100132707	5.7	3.9	18.8	·	·	·
CHTF8|54921	4.4	·	7.2	1	1	·
NFKBIA|4792	4.3	1.9	3.4	247	0.27	0.22
EPB41L4B|54566	3.6	2.6	·	19	0.47	0.22
UGP2|7360	3.6	2.2	·	19	0.15	1
DPY19L2P1|554236	3.3	·	3.3	1	·	·
TRMT2B|79979	3.3	2.2	·	·	·	·
HSD3B7|80270	3.2	1.9	7.6	19	0.05	·
DLAT|1737	3.2	2.9	·	75	0.16	0.16
NIPAL2|79815	2.8	1.9	·	·	·	·
FGD3|89846	2.7	·	5.9	10	0.2	0.5
JRKL|8690	2.7	2.6	·	2	·	·
ZBED1|9189	2.5	2.4	·	6	·	·
KCNJ11|3767	2.3	·	·	647	0.02	·
WAC|51322	2.0	3.2	·	260	0.05	0.25
FLT3|2322	2.0	·	·	4435	0.55	0.42
STK3|6788	1.9	2.3	·	107	0.32	0.15
PAOX|196743	1.9	1.9	·	18	0.11	·
C14orf68|283600	·	3.3	·	·	·	·
LIN7C|55327	·	3.1	·	36	0.06	·
PNRC2|55629	·	2.1	·	15	·	·
SLC39A7|7922	·	1.8	·	22	0.18	·
MAGT1|84061	·	1.7	·	50	0.12	0.17
IRF2|3660	·	·	10.9	310	0.21	0.14
PELO|53918	·	·	7.0	265	0.08	0.04
SUSD3|203328	·	·	5.3	5	0.6	0.67
LEF1|51176	·	·	3.2	940	0.29	0.23
CPA4|51200	·	·	1.4	18	0.22	·

TABLE 4.G.2 Top 20 selected genes per model for the GBM cancer, with the corresponding effects. Dots (·) mean zeros.

Genes	Model effects (%)			MEDLINE data		
	C-mix	CURE	Cox PH	(1)	f_1	f_2
ARMCX6|54470	4.9	·	23.6	1	·	·
FAM35A|54537	4.4	·	21.8	·	·	·
CLEC4GP1|440508	3.9	5.1	2.8	·	·	·
INSL3|3640	3.6	2.7	1.7	404	0.06	0.12
REM1|28954	3.2	·	·	54	0.05	0.66
FAM35B2|439965	3.0	·	·	·	·	·
TSPAN4|7106	2.7	·	·	16	0.31	0.4
AP3M1|26985	2.7	·	·	2	0.5	·
PXN|5829	2.6	·	15.4	891	0.25	0.18
PDE4C|5143	2.5	·	·	67	0.06	0.25
PGBD5|79605	2.5	·	·	5	0.25	·
NRG1|3084	2.4	·	18.5	1207	0.12	0.29
LOC653786	2.2	·	·	·	·	·
FERMT1|55612	2.1	·	·	115	0.19	0.18
PLD3|23646	2.0	·	·	38	0.10	0.25
MIER1|57708	1.9	·	2.1	16	0.31	·
UTP14C|9724	1.8	·	·	5	0.4	·
AZU1|566	1.8	·	·	15	0.2	0.33
KCNC4|3749	1.7	·	·	30	0.1	0.33
FAM35B|414241	1.6	·	·	·	·	·
CRELD1|78987	·	32.2	·	32	0.03	·
HMGN5|79366	·	21.2	·	41	0.54	0.32
PNLDC1|154197	·	12.2	·	3	·	·
LOC493754	·	9.8	·	·	·	·
KIAA0146|23514	·	8.7	·	3	0.67	·
TMCO6⁻55374	·	3.6	·	4	0.25	·
ABLIM1|3983	·	2.1	·	20	0.2	·
OSBPL11|114885	·	1.0	·	·	·	·
TRAPPC1|58485	·	0.9	·	4	0.75	·
TBCEL|219899	·	0.5	·	7	0.28	·
RPL39L|116832	·	·	8.8	10	0.7	0.14
GALE|2582	·	·	3.5	540	0.02	·
BBC3|27113	·	·	0.7	561	0.54	0.38
DUSP6|1848	·	·	0.6	307	0.30	0.22

TABLE 4.G.3 Top 20 selected genes per model for the KIRC cancer, with the corresponding effects. Dots (·) mean zeros.

Genes	Model effects (%)			MEDLINE data		
	C-mix	CURE	Cox PH	(1)	f_1	f_2
BCL2L12|83596	8.6	2.7	·	64	0.72	0.39
MARS|4141	7.5	6.9	7.2	577	0.02	0.1
NUMBL|9253	7.2	28.6	3.3	56	0.14	0.25
CKAP4|10970	6.1	10.6	22.3	825	0.63	0.11
HN1|51155	5.8	3.8	·	13	0.38	0.2
GIPC2|54810	5.7	·	·	15	0.6	0.11
NPR3|4883	5.2	·	·	105	0.05	0.6
GBA3|57733	5.0	·	·	19	0.10	·
SLC47A1|55244	5.0	·	·	70	0.06	·
ALDH3A2|224	4.7	·	2.6	52	0.06	0.33
CCNF|899	4.2	2.8	·	50	0.24	0.08
EHHADH|1962	3.9	·	·	90	0.1	·
SGCB|6443	3.3	·	·	30	·	·
GFPT2|9945	2.7	1.3	·	18	0.22	0.25
PPAP2B|8613	2.3	·	·	29	0.17	0.2
MBOAT7|79143	1.9	13.8	11.1	15	·	·
OSBPL1A|114876	1.5	·	·	7	·	·
C16orf57|79650	1.2	·	·	26	·	·
ATXN7L3|56970	0.9	2.5	·	9	·	·
C16orf59|80178	0.8	·	·	3	0.66	·
STRADA¯92335	·	20.7	53.5	9	·	·
ABCC10|89845	·	3.9	·	80	0.32	0.23
MDK|4192	·	1.2	·	789	0.38	0.23
C16orf59|80178	·	1.1	·	3	0.6	·

Chapitre 5

Binarsity : a penalization for one-hot encoded features in linear supervised learning

Sommaire

Abstract. This chapter deals with the problem of large-scale linear supervised learning in settings where a large number of continuous features are available. We propose to combine the well-known trick of one-hot encoding of continuous features with a new penalization called *binarsity*. In each group of binary features coming from the one-hot encoding of a single raw continuous feature, this penalization uses total-variation regularization together with an extra linear constraint to avoid collinearity within groups. A non-asymptotic oracle inequality for generalized linear models is proposed, and numerical experiments illustrate the good performances of our approach on several datasets. It is also noteworthy that our method has a numerical complexity comparable to standard ℓ_1 penalization.

Résumé. Ce chapitre considère le problème d'apprentissage supervisé linéaire où un grand nombre de covariables continues sont disponibles. Nous proposons de combiner l'encodage "one-hot" des covariables continues avec l'utilisation d'une nouvelle pénalité appelée *binarsity*. Dans chaque groupe de variables binaires générées par l'encodage des covariables continues, cette pénalité impose une régularisation par variation totale ainsi qu'une contrainte linéaire pour traiter le problème de colinéarité dans les groupes. Une inégalité oracle non-asymptotique en prédiction est proposée pour les modèles linéaires généralisés, et la méthode donne de bonnes performances sur les différents jeux de données considérés. La complexité numérique est de plus comparable à celle de la pénalité ℓ_1 classique.

5.1 Introduction

In many applications, datasets used for linear supervised learning contain a large number of continuous features, with a large number of samples. An example is web-marketing, where features are obtained from bag-of-words scaled using tf-idf [Russell, 2013], recorded during the visit of users on websites. A well-known trick [Wu and Coggeshall, 2012, Liu et al., 2002] in this setting is to replace each raw continuous feature by a set of binary features that one-hot encodes the interval containing it, among a list of intervals partitioning the raw feature range. This improves the linear decision function with respect to the raw continuous features space, and can therefore improve prediction. However, this trick is prone to over-fitting, since it increases significantly the number of features.

A new penalization. To overcome this problem, we introduce a new penalization called *binarsity*, that penalizes the model weights learned from such grouped one-hot encodings (one group for each raw continuous feature). Since the binary features within these groups are naturally ordered, the binarsity penalization combines a group total-variation penalization, with an extra linear constraint in each group to avoid collinearity between the one-hot encodings. This penalization forces the weights of the model to be as constant (with respect to the order induced by the original feature) as possible within a group, by selecting a minimal number of relevant cut-points. Moreover, if the model weights are all equal within a group, then the full block of weights is zero, because of the extra linear constraint. This allows to perform raw feature selection.

High-dimensional linear supervised learning. To address the problem of high-dimensionality of features, sparse linear inference is now an ubiquitous technique for dimension reduction and variable selection, see for instance Bühlmann and van de Geer [2011] and Hastie et al. [2001b] among many others. The principle is to induce sparsity (large number of zeros) in the model weights, assuming that only a few features are actually helpful for the label prediction. The most popular way to induce sparsity in model weights is to add a ℓ_1-penalization (lasso) term to the goodness-of-fit [Tibshirani, 1996]. This typically leads to sparse parametrization of models, with a level of sparsity that depends on the strength of the penalization. Statistical properties of ℓ_1-penalization have been extensively investigated, see for instance Knight and Fu [2000], Zhao and Yu [2006], Bunea et al. [2007], Bickel et al. [2009] for linear and generalized linear models and Donoho and Huo [2001], Donoho and Elad [2002], Candès et al. [2008], Candès and Wakin [2008] for compressed sensing, among others.

However, the lasso ignores ordering of features. In Tibshirani et al. [2005], a structured sparse penalization is proposed, known as fused lasso, which provides

superior performance in recovering the true model in such applications where features are ordered in some meaningful way. It introduces a mixed penalization using a linear combination of the ℓ_1-norm and the total-variation penalization, thus enforcing sparsity in both the weights and their successive differences. Fused lasso has achieved great success in some applications such as comparative genomic hybridization [Rapaport et al., 2008], image denoising [Friedman et al., 2007], and prostate cancer analysis [Tibshirani et al., 2005].

Features discretization and cuts. For supervised learning, it is often useful to encode the input features in a new space to let the model focus on the relevant areas [Wu and Coggeshall, 2012]. One of the basic encoding technique is *feature discretization* or *feature quantization* [Liu et al., 2002] that partitions the range of a continuous feature into intervals and relates these intervals with meaningful labels. Recent overviews of discretization techniques can be found in Liu et al. [2002] or Garcia et al. [2013].

Obtaining the optimal discretization is a NP-hard problem [Chlebus and Nguyen, 1998], and an approximation can be easily obtained using a greedy approach, as proposed in decision trees : CART [Breiman et al., 1984] and C4.5 [Quinlan, 1993], among others, that sequentially select pairs of features and cuts that minimize some purity measure (intra-variance, Gini index, information gain are the main examples). These approaches build decision functions that are therefore very simple, by looking only at a single feature at a time, and a single cut at a time. Ensemble methods (boosting [Lugosi and Vayatis, 2004], random forests [Breiman, 2001]) improve this by combining such decisions trees, at the expense of models that are harder to interpret.

Main contribution. This chapter considers the setting of linear supervised learning. The main contribution of this chapter is the idea to use a total-variation penalization, with an extra linear constraint, on the weights of a generalized linear model trained on a binarization of the raw continuous features, leading to a procedure that selects multiple cut-points per feature, looking at all features simultaneously. Our approach therefore increases the capacity of the considered generalized linear model : several weights are used for the binarized features instead of a single one for the raw feature. This leads to a more flexible decision function compared to the linear one : when looking at the decision function as a function of a single raw feature, it is now piecewise constant instead of linear, as illustrated in Figure 5.2 below.

Organization of the chapter. The proposed methodology is described in Section 5.2. Section 5.3 establishes an oracle inequality for generalized linear models. Section 5.4 highlights the results of the method on various datasets and compares

its performances to well known classification algorithms. Finally, we discuss the obtained results in Section 5.5.

Notations. Throughout the chapter, for every $q > 0$, we denote by $\|v\|_q$ the usual ℓ_q-quasi norm of a vector $v \in \mathbb{R}^m$, namely

$$\|v\|_q = (\sum_{k=1}^{m} |v_k|^q)^{1/q},$$

and

$$\|v\|_\infty = \max_{k=1,\dots,m} |v_k|.$$

We also denote

$$\|v\|_0 = |\{k : v_k \neq 0\}|,$$

where $|A|$ stands for the cardinality of a finite set A. For $u, v \in \mathbb{R}^m$, we denote by $u \odot v$ the Hadamard product

$$u \odot v = (u_1 v_1, \dots, u_m v_m)^\top.$$

For any $u \in \mathbb{R}^m$ and any $L \subset \{1, \dots, m\}$, we denote u_L as the vector in $\mathbb{R}^m$ satisfying $(u_L)_k = u_k$ for $k \in L$ and $(u_L)_k = 0$ for $k \in L^\complement = \{1, \dots, m\} \backslash L$. We write, for short, $\mathbf{1}$ (resp. $\mathbf{0}$) for the vector of $\mathbb{R}^m$ having all coordinates equal to one (resp. zero). Finally, we denote by $\text{sign}(x)$ the set of sub-differentials of the function $x \mapsto |x|$, namely

$$\text{sign}(x) = \begin{cases} \{1\} & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \\ \{-1\} & \text{if } x < 0. \end{cases}$$

5.2 The proposed method

Consider a supervised training dataset $(x_i, y_i)_{i=1,\dots,n}$ containing features

$$x_i = (x_{i,1}, \dots, x_{i,p})^\top \in \mathbb{R}^p$$

and labels $y_i \in \mathcal{Y} \subset \mathbb{R}$, that are independent and identically distributed samples of (X, Y) with unknown distribution $\mathbb{P}$. Let us denote

$$\boldsymbol{X} = [x_{i,j}]_{1 \leq i \leq n; 1 \leq j \leq p}$$

the $n \times p$ features matrix vertically stacking the n samples of p raw features. Let $\boldsymbol{X}_{\bullet,j}$ be the j-th feature column of $\boldsymbol{X}$.

Binarization. The binarized matrix $\boldsymbol{X}^B$ is a matrix with an extended number $d > p$ of columns, where the j-th column $\boldsymbol{X}_{\bullet,j}$ is replaced by $d_j \geq 2$ columns $\boldsymbol{X}^B_{\bullet,j,1}, \ldots, \boldsymbol{X}^B_{\bullet,j,d_j}$ containing only zeros and ones. Its i-th row is written

$$x_i^B = (x^B_{i,1,1}, \ldots, x^B_{i,1,d_1}, x^B_{i,2,1}, \ldots, x^B_{i,2,d_2}, \ldots, x^B_{i,p,1}, \ldots, x^B_{i,p,d_p})^\top \in \mathbb{R}^d.$$

In order to simplify presentation of our results, we assume in the chapter that all raw features $\boldsymbol{X}_{\bullet,j}$ are continuous, so that they are transformed using the following one-hot encoding. We consider a full partitioning without overlap, that is

$$\bigcup_{k=1}^{d_j} I_{j,k} = \text{range}(\boldsymbol{X}_{\bullet,j})$$

and $I_{j,k} \cup I_{j,k'} = \varnothing$ for all $k \neq k'$ with $k, k' \in \{1, \ldots, d_j\}$, and define

$$x^B_{i,j,k} = \begin{cases} 1 & \text{if } x_{i,j} \in I_{j,k}, \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, d_j$. A natural choice of intervals is given by quantiles, namely

$$I_{j,1} = \Big[q_j(0), q_j(\frac{1}{d_j})\Big]$$

and

$$I_{j,k} = \Big(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\Big]$$

for $k = 2, \ldots, d_j$, where $q_j(\alpha)$ denotes a quantile of order $\alpha \in [0, 1]$ for $\boldsymbol{X}_{\bullet,j}$. In practice, if there are ties in the estimated quantiles for a given feature, we simply choose the set of ordered unique values to construct the intervals. This principle of binarization is a well-known trick [Garcia et al., 2013], that allows to improve over the linear decision function with respect to the raw feature space : it uses a larger number of model weights, for each interval of values for the feature considered in the binarization. If training data contains also unordered qualitative features, one-hot encoding with ℓ_1-penalization can be used for instance.

Goodness-of-fit. Given a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$, we consider the goodness-of-fit term

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell\Big(y_i, m_\theta(x_i)\Big), \tag{5.1}$$

where

$$m_\theta(x_i) = \theta^\top x_i^B$$

and $\theta \in \mathbb{R}^d$ with $d = \sum_{j=1}^p d_j$.

We then have

$$\theta = (\theta_{1,\bullet}^\top, \dots, \theta_{p,\bullet}^\top)^\top,$$

with $\theta_{j,\bullet}$ corresponding to the group of coefficients weighting the binarized raw j-th feature. We focus on generalized linear models [Green and Silverman, 1994], where the conditional distribution $Y|X = x$ is assumed to be from a one-parameter exponential family distribution with a density of the form

$$y|x \mapsto f^0(y|x) = \exp\left(\frac{ym^0(x) - b\Big(m^0(x)\Big)}{\phi} + c(y,\phi)\right), \tag{5.2}$$

with respect to a reference measure which is either the Lebesgue measure (*e.g.* in the Gaussian case) or the counting measure (*e.g.* in the logistic or Poisson cases), leading to a loss function of the form

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

The density described in (5.2) encompasses several distributions, see Table 5.1. The functions $b(\cdot)$ and $c(\cdot)$ are known, while the natural parameter function $m^0(\cdot)$ is unknown. The dispersion parameter ϕ is assumed to be known in what follows. It is also assumed that $b(\cdot)$ is three times continuously differentiable. It is standard to notice that

$$\mathbb{E}[Y|X = x] = \int y f^0(y|x)\mathrm{d}y = b'\Big(m^0(x)\Big),$$

where b' stands for the derivative of b. This formula explains how b' links the conditional expectation to the unknown m^0. The results given in Section 5.3 rely on the following Assumption.

Assumption 1 *Assume that b is three times continuously differentiable, and that there exist constants $C_n > 0$, and $0 < L_n \leq U_n$ such that $C_n = \max_{i=1,\dots,n} |m^0(x_i)| < \infty$ and $L_n \leq \max_{i=1,\dots,n} b''\Big(m^0(x_i)\Big) \leq U_n$.*

This assumption is satisfied for most standard generalized linear models. In Table 5.1, we list some standard examples that fit in this framework, see also van de Geer [2008] and Rigollet [2012].

Binarsity. Several problems occur when using the binarization trick described above :

(P1) The one-hot-encodings satisfy $\sum_{k=1}^{d_j} \boldsymbol{X}^B_{i,j,k} = 1$ for $j = 1, \dots, p$, meaning that the columns of each block sum to $\mathbf{1}$, making $\boldsymbol{X}^B$ not of full rank by construction.

(P2) Choosing the number of intervals d_j for binarization of each raw feature j is not an easy task, as too many might lead to overfitting : the number of model-weights increases with each d_j, leading to a over-parametrized model.

	ϕ	$b(z)$	$b'(z)$	$b''(z)$	L_n	U_n
Normal	σ^2	$\frac{z^2}{2}$	z	1	1	1
Logistic	1	$\log(1+e^z)$	$\frac{e^z}{1+e^z}$	$\frac{e^z}{(1+e^z)^2}$	$\frac{e^{C_n}}{(1+e^{C_n})^2}$	$\frac{1}{4}$
Poisson	1	e^z	e^z	e^z	e^{-C_n}	e^{C_n}

TABLE 5.1 Examples of standard distributions that fit in the considered setting of generalized linear models, with the corresponding constants in Assumption 1.

(P3) Some of the raw features $\boldsymbol{X}_{\bullet,j}$ might not be relevant for the prediction task, so we want to select raw features from their one-hot encodings, namely induce block-sparsity in θ.

A usual way to deal with (P1) is to impose a linear constraint [Agresti, 2015] in each block. In our penalization term, we impose

$$\sum_{k=1}^{d_j} \theta_{j,k} = 0 \tag{5.3}$$

for all $j = 1, \dots, p$. Now, the trick to tackle (P2) is to remark that within each block, binary features are ordered. We use a within block total-variation penalization

$$\sum_{j=1}^{p} \|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}$$

where

$$\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} = \sum_{k=2}^{d_j} \hat{w}_{j,k} |\theta_{j,k} - \theta_{j,k-1}|, \tag{5.4}$$

with weights $\hat{w}_{j,k} > 0$ to be defined later, to keep the number of different values taken by $\theta_{j,\bullet}$ to a minimal level. Finally, dealing with (P3) is actually a by-product of dealing with (P1) and (P2). Indeed, if the raw feature j is not-relevant, then $\theta_{j,\bullet}$ should have all entries constant because of the penalization (6.6), and in this case all entries are zero, because of (5.3). We therefore introduce the following penalization, called *binarsity*

$$\mathrm{bina}(\theta) = \sum_{j=1}^{p} \Big(\sum_{k=2}^{d_j} \hat{w}_{j,k} |\theta_{j,k} - \theta_{j,k-1}| + \delta_1(\theta_{j,\bullet}) \Big) \tag{5.5}$$

where the weights $\hat{w}_{j,k} > 0$ are defined in Section 5.3 below, and where

$$\delta_1(u) = \begin{cases} 0 & \text{if} \quad \mathbf{1}^\top u = 0, \\ \infty & \text{otherwise.} \end{cases}$$

We consider the goodness-of-fit (5.1) penalized by (5.5), namely

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in \mathbb{R}^d} \Big\{ R_n(\theta) + \operatorname{bina}(\theta) \Big\}. \tag{5.6}$$

An important fact is that this optimization problem is numerically cheap, as explained in the next paragraph. Figure 5.1 illustrates the effect of the binarsity penalization with a varying strength on an example.

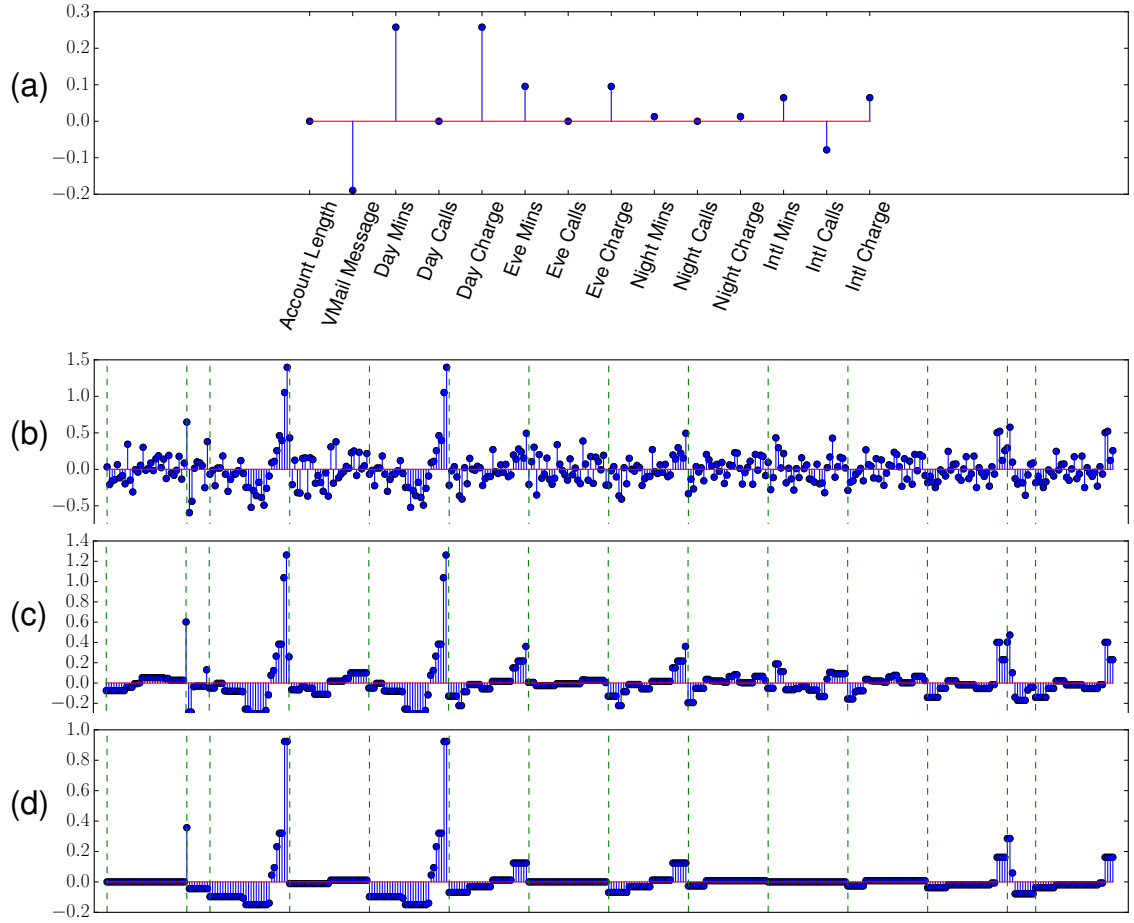


FIGURE 5.1 Illustration of the binarsity penalization on the "Churn" dataset (see Section 5.4 for details) using logistic regression. Figure (a) shows the model weights learned by the lasso method on the continuous raw features. Figure (b) shows the unpenalized weights on the binarized features, where the dotted green lines mark the limits between blocks corresponding to each raw features. Figures (c) and (d) show the weights with medium and strong binarsity penalization respectively. We observe in (c) that some significant cut-points start to be detected, while in (d) some raw features are completely removed from the model, the same features as those removed in (a).

In Figure 5.2, we illustrate on a toy example, when $p = 2$, the decision boundaries obtained for logistic regression (LR) on raw features, LR on binarized features and LR on binarized features with the binarsity penalization.

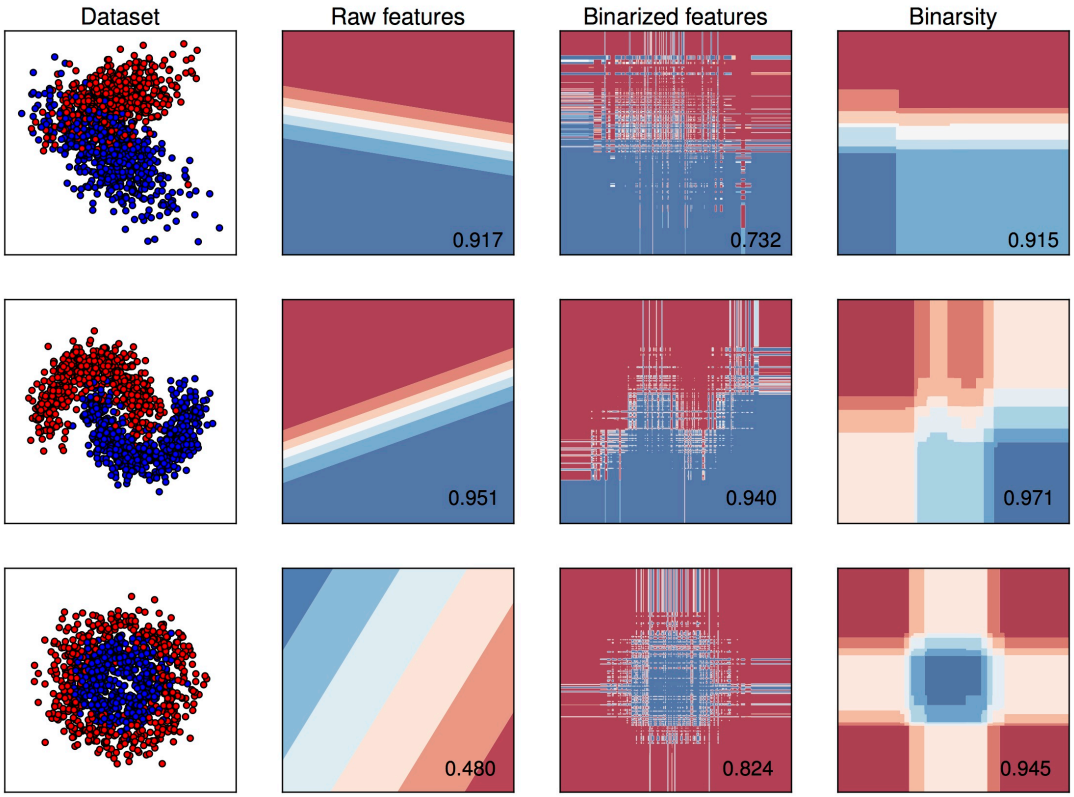


FIGURE 5.2 Illustration of binarsity on 3 simulated toy datasets for binary classification with two classes (blue and red points). We set $n = 1000$, $p = 2$ and $d_1 = d_2 = 100$. In each row, we display the simulated dataset, followed by the decision boundaries for a logistic regression classifier trained on initial raw features, then on binarized features without regularization, and finally on binarized features with binarsity. The corresponding testing AUC score is given on the lower right corner of each figure. Our approach allows to keep an almost linear decision boundary in the first row, while a good decision boundaries are learned on the two other examples, which correspond to non-linearly separable datasets, without apparent overfitting.

Proximal operator of binarsity. The proximal operator and proximal algorithms are important tools for non-smooth convex optimization, with important applications in the field of supervised learning with structured sparsity [Bach et al., 2012]. The proximal operator of a proper lower semi-continuous [Bauschke and Com-

bettes, 2011] convex function $g : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\text{prox}_g(v) \in \text{argmin}_{u \in \mathbb{R}^d} \left\{ \frac{1}{2}\|v - u\|_2^2 + g(u) \right\}.$$

Proximal operators can be interpreted as generalized projections. Namely, if g is the indicator of a convex set $C \subset \mathbb{R}^d$ given by

$$g(u) = \delta_C(u) = \begin{cases} 0 & \text{if } u \in C, \\ \infty & \text{otherwise,} \end{cases}$$

then prox_g is the projection operator onto C. It turns out that the proximal operator of binarsity can be computed very efficiently, using an algorithm [Condat, 2013] that we modify in order to include weights $\hat{w}_{j,k}$. It applies in each group the proximal operator of the total-variation since binarsity penalization is block separable, followed by a centering within each block to satisfy the sum-to-zero constraint, see Algorithm 2 below. We refer to Algorithm 3 in Section 5.B for the weighted total-variation proximal operator.

Proposition 5.2.1 *Algorithm 2 computes the proximal operator of* $\text{bina}(\theta)$ *given by* (5.5).

Algorithm 2: Proximal operator of $\text{bina}(\theta)$, see (5.5)

Input: vector $\theta \in \mathbb{R}^d$ and weights $\hat{w}_{j,k}$ for $j = 1, \dots, p$ and $k = 1, \dots, d_j$
Output: vector $\eta = \text{prox}_{\text{bina}}(\theta)$
for $j = 1$ ***to*** p **do**
  $\beta_{j,\bullet} \leftarrow \text{prox}_{\|\theta_{j,\bullet}\|_{\text{TV},\hat{w}_{j,\bullet}}}(\theta_{j,\bullet})$ (TV-weighted prox in block j, see (6.6))
  $\eta_{j,\bullet} \leftarrow \beta_{j,\bullet} - \frac{1}{d_j} \sum_{k=1}^{d_j} \beta_{j,k}$ (within-block centering)
Return : η

A proof of Proposition 5.2.1 is given in Section 5.A. Algorithm 2 leads to a very fast numerical routine, see Section 5.4. The next section provides a theoretical analysis of our algorithm with an oracle inequality for the prediction error.

5.3 Theoretical guarantees

We now investigate the statistical properties of (5.7) where the weights in the binarsity penalization have the form

$$\hat{w}_{j,k} = \mathcal{O}\Big(\sqrt{\frac{\log d}{n}\ \hat{\pi}_{j,k}}\Big),$$

with

$$\hat{\pi}_{j,k} = \frac{\Big|\Big\{i = 1, \ldots, n : x_{i,j} \in \Big(q_j\Big(\frac{k}{d_j}\Big), q_j(1)\Big]\Big\}\Big|}{n}$$

for all $k \in \{2, \ldots, d_j\}$, see Theorem 5.3.1 for a precise definition of $\hat{w}_{j,k}$. Note that $\hat{\pi}_{j,k}$ corresponds to the proportion of ones in the sub-matrix obtained by deleting the first k columns in the j-th binarized block matrix $\boldsymbol{X}^B_{\bullet,j}$. In particular, we have $\hat{\pi}_{j,k} > 0$ for all j, k. We consider the risk measure defined by

$$R(m_\theta) = \frac{1}{n}\sum_{i=1}^{n}\Big\{ - b'(m^0(x_i))m_\theta(x_i) + b(m_\theta(x_i))\Big\},$$

which is standard with generalized linear models [van de Geer, 2008].

We aim at evaluating how "close" to the minimal possible expected risk our estimated function $m_{\hat{\theta}}$ with $\hat{\theta}$ given by (5.7) is. To measure this closeness, we establish a non-asymptotic oracle inequality with a fast rate of convergence considering the excess risk of $m_{\hat{\theta}}$, namely $R(m_{\hat{\theta}}) - R(m^0)$. To derive this inequality, we consider for technical reasons the following problem instead of (5.6) :

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in B_d(\rho)} \Big\{R_n(\theta) + \operatorname{bina}(\theta)\Big\}, \tag{5.7}$$

where

$$B_d(\rho) = \Big\{\theta \in \mathbb{R}^d : \sum_{j=1}^{p} \|\theta_{j,\bullet}\|_\infty \leq \rho\Big\}.$$

Such a constraint corresponds to the ones usually used in literature for the proof of oracle inequalities for sparse generalized linear models, see for instance van de Geer [2008], a recent contribution for the particular case of Poisson regression being Ivanoff et al. [2016]. Under this assumption, we have

$$\max_{i=1,\ldots,n} \langle x_i^B, \theta\rangle \leq \sum_{j=1}^{p} \|\theta_{j,\bullet}\|_\infty, \tag{5.8}$$

since the entries of $\boldsymbol{X}^B$ are in $\{0, 1\}$, which proves useful for the proof of Theorem 5.3.1 below. The restriction to $B_d(\rho)$ allows to establish a connection, via the notion of self-concordance, see Bach [2010], between the empirical squared ℓ_2-norm and the empirical Kullback divergence (see Lemma 5.C.6 in Appendix 5.C).

We also impose a restricted eigenvalue assumption on $\boldsymbol{X}^B$. For all $\theta \in \mathbb{R}^d$, let

$$J(\theta) = \Big[J_1(\theta), \ldots, J_p(\theta)\Big]$$

be the concatenation of the support sets relative to the total-variation penalization, that is

$$J_j(\theta) = \Big\{k : \theta_{j,k} \neq \theta_{j,k-1}, \text{ for } k = 2, \ldots, d_j\Big\}.$$

Similarly, we denote

$$J^{\complement}(\theta) = \left[J_1^{\complement}(\theta), \dots, J_p^{\complement}(\theta)\right]$$

the complementary of $J(\theta)$. The restricted eigenvalue condition is defined as follow.

Assumption 2 *Let $K = [K_1, \dots, K_p]$ be a concatenation of index sets such that*

$$\sum_{j=1}^{p} |K_j| \leq J^\star, \tag{5.9}$$

where $J^\star$ is a positive integer. Define

$$\kappa(K) \in \inf_{u \in \mathscr{C}_{\mathrm{TV},\hat{w}}(K)\setminus\{\mathbf{0}\}} \left\{ \frac{\|\boldsymbol{X}^B u\|_2}{\sqrt{n}\|u_K\|_2} \right\}$$

with

$$\mathscr{C}_{\mathrm{TV},\hat{w}}(K) = \left\{ u \in \mathbb{R}^d : \sum_{j=1}^{p} \|(u_{j,\bullet})_{K_j{}^{\complement}}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} \leq 2 \sum_{j=1}^{p} \|(u_{j,\bullet})_{K_j}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} \right\}. \tag{5.10}$$

We assume that the following condition holds

$$\kappa(K) > 0$$

for any K satisfying (5.9).

The set $\mathscr{C}_{\mathrm{TV},\hat{w}}(K)$ is a cone composed by all vectors with a support "close" to K. Theorem 5.3.1 gives a risk bound for the estimator $m_{\hat{\theta}}$.

Theorem 5.3.1 *Let Assumptions 1 and 2 be satisfied. Fix $A > 0$ and choose*

$$\hat{w}_{j,k} = \sqrt{\frac{2U_n \phi(A + \log d)}{n}}\, \hat{\pi}_{j,k}. \tag{5.11}$$

Let $\psi(u) = e^u - u - 1$, and consider the following constants

$$C_n(\rho, L_n) = \frac{L_n \psi(-2(C_n + \rho))}{C_n^2(\rho, p)}, \quad \epsilon > \frac{2}{C_n(\rho, L_n)} \quad \textit{and} \quad \zeta = \frac{4}{\epsilon\, C_n(\rho, L_n) - 2}.$$

Then, with probability at least $1 - 2e^{-A}$, any solution $\hat{\theta}$ of problem (5.7) fulfills the following risk bound

$$R(m_{\hat{\theta}}) - R(m^0) \leq (1+\zeta) \inf_{\substack{\theta \in B_d(\rho) \\ \forall j\ \mathbf{1}^\top \theta_{j,\bullet} \\ |J(\theta)| \leq J^*}} \left\{ R(m_\theta) - R(m^0) + \frac{\xi |J(\theta)|}{\kappa^2(J(\theta))} \max_{j=1,\dots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2 \right\}, \tag{5.12}$$

where

$$\xi = \frac{512\, \epsilon^2\, C_n(\rho, L_n)}{\epsilon\, C_n(\rho, L_n) - 2}.$$

A proof of Theorem 5.3.1 is given in Section 5.C. Note that $\hat{w}_{j,k} > 0$, since by construction $\hat{\pi}_{j,k} > 0$ for all j, k. The second term in the right-hand side of (5.12) can be viewed as a variance term, and its dominant term satisfies

$$\frac{|J(\theta)|}{\kappa^2(J(\theta))} \max_{j=1,\dots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2 \leq \frac{\tilde{A} U_n \phi}{\kappa^2(J(\theta))} \frac{|J(\theta)| \log d}{n}, \tag{5.13}$$

for some positive constant $\tilde{A}$. The complexity term in (5.13) depends on both the sparsity and the restricted eigenvalues of the binarized matrix. The value $|J(\theta)|$ characterizes the sparsity of the vector θ, that is the smaller $|J(\theta)|$, the sparser θ. Moreover, for the case of least squares regression, the oracle inequality in Theorem 5.3.1 is sharp, in the sense that $\zeta = 0$ (see Remark 5.C.1 in Section 5.C).

Note that if the model is well-specified, namely if $m^0(x) = \langle x_i^B, \theta^0 \rangle$ for some $\theta^0 \in \mathbb{R}^d$ satisfying the linear constraints $\mathbf{1}^\top \theta^0_{j,\bullet} = 0$ for all $j = 1, \dots, p$, the constant ρ from $B_d(\rho)$ can be bounded by $J^\star \|\theta^0\|_\infty$, which is much smaller than the ambient dimension p. Note also that this constraint is only used for technical reasons, and that we do not use it in the numerical experiments used below.

5.4 Numerical experiments

In this section, we first illustrate the fact that the binarsity penalization is roughly only two times slower than basic ℓ_1-penalization, see the timings in Figure 5.1. We then compare binarsity to a large number of baselines, see Table 5.1, using 9 classical binary classification datasets obtained from the UCI Machine Learning Repository [Lichman, 2013], see Table 5.2.

Name	Description	Reference
Lasso	Logistic regression (LR) with ℓ_1 penalization	Tibshirani [1996]
Group L1	LR with group ℓ_1 penalization	Meier et al. [2008]
Group TV	LR with group total-variation penalization	
SVM	Support vector machine with radial basis kernel	Schölkopf and Smola [2002]
GAM	Generalized additive model	Hastie and Tibshirani [1990]
RF	Random forest classifier	Breiman [2001]
GB	Gradient boosting	Friedman [2002]

TABLE 5.1 Baselines considered in our experiments. Note that Group L1 and Group TV are considered on binarized features.

For each method, we randomly split all datasets into a training and a test set (30% for testing), and all hyper-parameters are tuned on the training set using V-fold cross-validation with $V = 10$. For support vector machine with radial basis

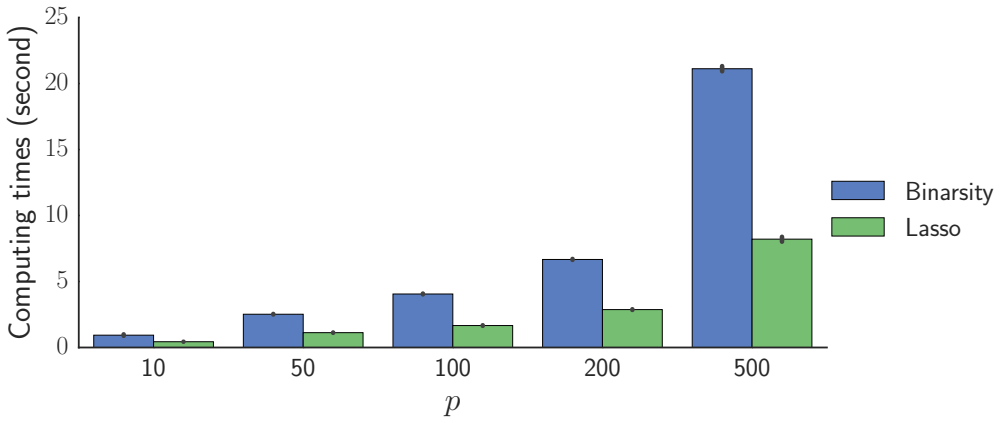


FIGURE 5.1 Average computing time in second (with the black lines representing $\pm$ the standard deviation) obtained on 100 simulated datasets for training a logistic model with binarsity VS lasso penalization, both trained on $\boldsymbol{X}^B$ with $d_j = 10$ for all $j \in 1, \ldots, p$. Features are Gaussian with a Toeplitz covariance matrix with correlation 0.5 and $n = 10000$. Note that the computing time ratio between the two methods stays roughly constant and equal to 2.

Dataset	#Samples	#Features	Reference
Ionosphere	351	34	Sigillito et al. [1989]
Churn	3333	21	Lichman [2013]
Default of credit card	30000	24	Yeh and Lien [2009]
Adult	32561	14	Kohavi [1996]
Bank marketing	45211	17	Moro et al. [2014]
Covertype	550088	10	Blackard and Dean [1999]
SUSY	5000000	18	Baldi et al. [2014]
HEPMASS	10500000	28	Baldi et al. [2016]
HIGGS	11000000	24	Baldi et al. [2014]

TABLE 5.2 Basic informations about the 9 considered datasets.

kernel (SVM), random forests (RF) and gradient boosting (GB), we use the reference implementations from the `scikit-learn` library [Pedregosa et al., 2011a], and we use the `LogisticGAM` procedure from the `pygam` library [1] for the GAM baseline. The binarsity penalization is proposed in the `tick` library [Bacry et al., 2017], we provide sample code for its use in Figure 5.2. Logistic regression with no penalization or ridge penalization gave similar or lower scores for all considered datasets, and are therefore not reported in our experiments.

1. https://github.com/dswah/pyGAM

The binarsity penalization does not require a careful tuning of d_j (number of bins for the one-hot encoding of raw feature j). Indeed, past a large enough value, increasing d_j even further barely changes the results since the cut-points selected by the penalization do not change anymore. This is illustrated in Figure 5.3, where we observe that past 50 bins, increasing d_j even further does not affect the performance, and only leads to an increase of the training time. In all our experiments, we therefore fix $d_j = 50$ for $j = 1, \dots, p$.

The results of all our experiments are reported in Figures 5.5 and 5.4. In Figure 5.5 we compare the performance of binarsity with the baselines on all 9 datasets, using ROC curves and the Area Under the Curve (AUC), while we report computing (training) timings in Figure 5.4. We observe that binarsity consistently outperforms lasso, as well as Group L1 : this highlights the importance of the TV norm within each group. The AUC of Group TV is always slightly below the one of binarsity, and more importantly it involves a much larger training time : convergence is slower for Group TV, since it does not use the linear constraint of binarsity, leading to a ill-conditioned problem (sum of binary features equals 1 in each block). Finally, binarsity outperforms also GAM and its performance is comparable in all considered examples to RF and GB, with computational timings that are orders of magnitude faster, see Figure 5.4. All these experiments illustrate that binarsity achieves an extremely competitive compromise between computational time and performance, compared to all considered baselines.

```python
# input: features X, labels y
from tick.inference import LogisticRegression
from tick.preprocessing import FeaturesBinarizer
from sklearn.model_selection import train_test_split

# binarize data
binarizer = FeaturesBinarizer(n_cuts=50)
X = binarizer.fit_transform(X)

# shuffle and split training and test sets
X, X_test, y, y_test = train_test_split(X, y, stratify=y)

# fit the model
learner = LogisticRegression(penalty='binarsity', C=1,
                             blocks_start=binarizer.blocks_start,
                             blocks_length=binarizer.blocks_length)
learner.fit(X, y)

# predict on test set
y_pred = learner.predict_proba(X_test)[:, 1]
```

FIGURE 5.2 Sample python code for the use of binarsity with logistic regression in the `tick` library, with the use of the `FeaturesBinarizer` transformer for features binarization.

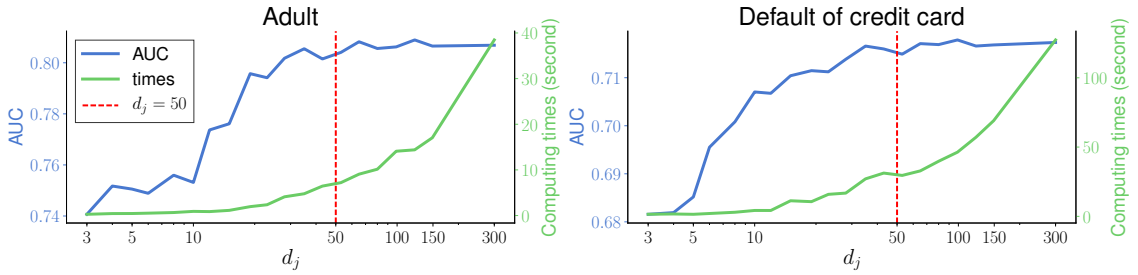


FIGURE 5.3 Impact of the number of bins used in each block (d_j) on the classification performance (measured by AUC) and on the training time using the "Adult" and "Default of credit card" datasets. All d_j are equal for $j = 1, \ldots, p$, and we consider in all cases the best hyper-parameters selected after cross validation. We observe that past $d_j = 50$ bins, performance is roughly constant, while training time strongly increases.

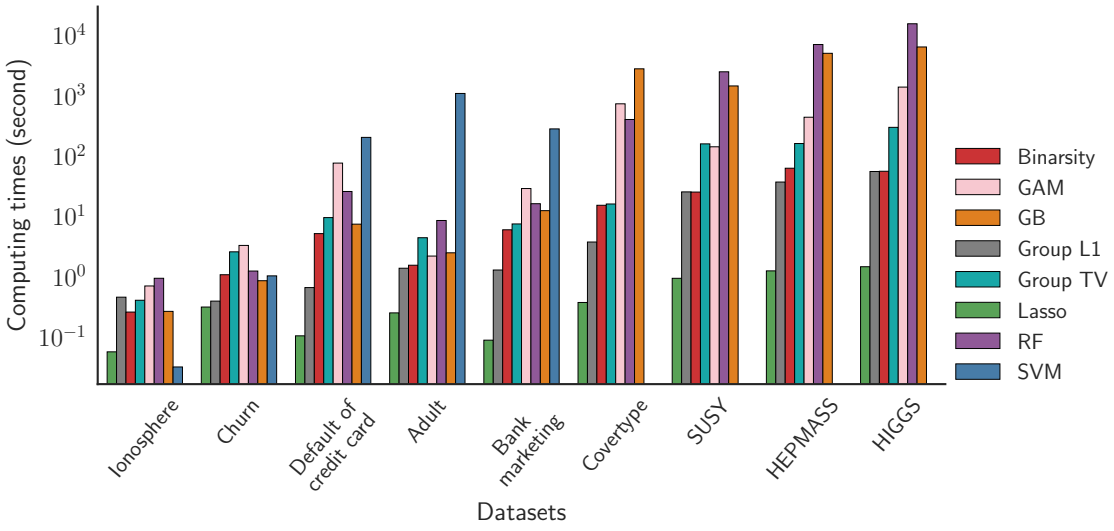


FIGURE 5.4 Computing time comparisons (in seconds) between the methods on the considered datasets. Note that the time values are log-scaled. These timings concern the learning task for each model with the best hyper parameters selected, after the cross validation procedure. The 4 last datasets contain too many examples for the SVM with RBF kernel to be trained in a reasonable time. Roughly, binarsity is between 2 and 5 times slower than ℓ_1 penalization on the considered datasets, but is more than 100 times faster than random forests or gradient boosting algorithms on large datasets, such as HIGGS.

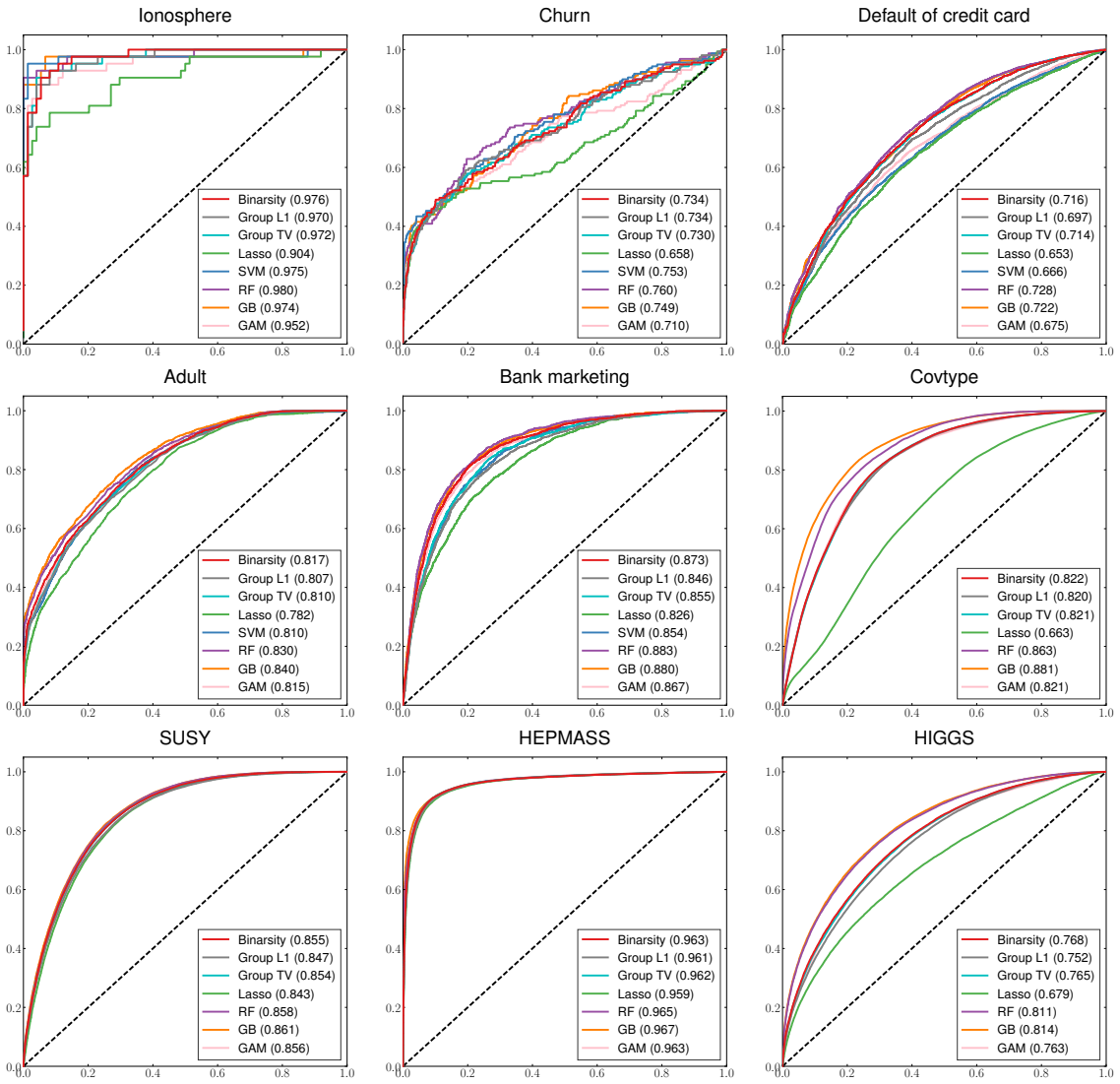


FIGURE 5.5 Performance comparison using ROC curves and AUC scores (given between parenthesis) computed on test sets. The 4 last datasets contain too many examples for SVM (RBF kernel). Binarsity consistently does a better job than lasso, Group L1, Group TV and GAM. Its performance is comparable to SVM, RF and GB but with computational timings that are orders of magnitude faster, see Figure 5.4.

5.5 Concluding remarks

In this chapter, we introduced the binarsity penalization for one-hot encodings of continuous features. We illustrated the good statistical properties of binarsity for generalized linear models by proving a non-asymptotic oracle inequality in prediction. We conducted extensive comparisons of binarsity with state-of-the-art algorithms for binary classification on several standard datasets. Experimental results illustrate that binarsity significantly outperforms lasso, Group L1 and Group TV penalizations and also generalized additive models, while being competitive with random forests and boosting. Moreover, it can be trained orders of magnitude faster than boosting and other ensemble methods. Even more importantly, it provides interpretability. Indeed, in addition to the raw feature selection ability of binarsity, the method pinpoints significant cut-points for all continuous feature. This leads to a much more precise and deeper understanding of the model than the one provided by lasso on raw features. These results illustrate the fact that binarsity achieves an extremely competitive compromise between computational time and performance, compared to all considered baselines.

Software

All the methodology discussed in this chapter is implemented in `Python/C++`. The code that generates all figures is available from `https://github.com/SimonBussy/binarsity` in the form of annotated programs, together with notebook tutorials.

Appendices

5.A Proof : the proximal operator of binarsity

For any fixed $j = 1, \dots, p$, we aim to prove that $\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}_{j,\bullet}}+\delta_1}$ is the composite proximal operators of $\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}_{j,\bullet}}}$ and prox_{δ_1}, namely

$$\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}_{j,\bullet}}+\delta_1}(\theta_{j,\bullet}) = \mathrm{prox}_{\delta_1}\left(\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}_{j,\bullet}}}(\theta_{j,\bullet})\right)$$

for all $\theta_{j,\bullet} \in \mathbb{R}^{d_j}$. Using Theorem 1 in Yu [2013], it is sufficient to show that for all $\theta_{j,\bullet} \in \mathbb{R}^{d_j}$, we have

$$\partial\Big(\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}\Big) \subseteq \partial\Big(\|\,\mathrm{prox}_{\delta_1}(\theta_{j,\bullet})\|_{\mathrm{TV},\hat{w}_{j,\bullet}}\Big). \tag{5.14}$$

Clearly, by the definition of the proximal operator, we have $\mathrm{prox}_{\delta_1}(\theta_{j,\bullet}) = \Pi_{\mathrm{span}\{\mathbf{1}\}^\perp}(\theta_{j,\bullet})$, where $\Pi_{\mathrm{span}\{\mathbf{1}\}^\perp}(\cdot)$ stands for the projection onto the hyperplane $\mathrm{span}\{\mathbf{1}\}^\perp$. Besides, we know that

$$\begin{aligned}
\Pi_{\mathrm{span}\{\mathbf{1}\}^\perp}(\theta_{j,\bullet}) &= \theta_{j,\bullet} - \Pi_{\mathrm{span}\{\mathbf{1}\}}(\theta_{j,\bullet}) \\
&= \theta_{j,\bullet} - \frac{\langle \theta_{j,\bullet}, \mathbf{1}\rangle}{\|\mathbf{1}\|_2^2}\mathbf{1} \\
&= \theta_{j,\bullet} - \bar{\theta}_{j,\bullet}\mathbf{1},
\end{aligned}$$

where $\bar{\theta}_{j,\bullet} = \frac{1}{d_j}\sum_{k=1}^{d_j}\theta_{j,k}$. Now, let us define the $d_j \times d_j$ matrix D_j by

$$D_j = \begin{bmatrix} 1 & 0 & & 0 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{d_j} \times \mathbb{R}^{d_j}. \tag{5.15}$$

We then remark that for all $\theta_{j,\bullet} \in \mathbb{R}^{d_j}$,

$$\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} = \sum_{k=2}^{d_j} \hat{w}_{j,k}|\theta_{j,k} - \theta_{j,k-1}| = \|\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}\|_1.$$

Using subdifferential calculus (see details in the proof of Proposition 5.C.2 below), one has

$$\partial\Big(\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}\Big) = \partial\Big(\|\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}\|_1\Big) = {D_j}^\top \hat{w}_{j,\bullet} \odot \mathrm{sign}(D_j\theta_{j,\bullet}).$$

Then, the linear constraint $\sum_{k=1}^{d_j}\theta_{j,k} = 0$ entails that

$${D_j}^\top \hat{w}_{j,\bullet} \odot \mathrm{sign}(D_j\theta_{j,\bullet}) = {D_j}^\top \hat{w}_{j,\bullet} \odot \mathrm{sign}(D_j(\theta_{j,\bullet} - \bar{\theta}_{j,\bullet}\mathbf{1})),$$

which leads to (5.14). Hence, setting $\beta_{j,\bullet} = \text{prox}_{\|\cdot\|_{\text{TV},\hat{w}_{j,\bullet}}}(\theta_{j,\bullet})$ and $\bar{\beta}_{j,\bullet} = \frac{1}{d_j}\sum_{k=1}^{d_j}\beta_{j,k}$ we get

$$\text{prox}_{\|\cdot\|_{\text{TV},\hat{w}_{j,\bullet}}+\delta_1}(\theta_{j,\bullet}) = \beta_{j,\bullet} - \bar{\beta}_{j,\bullet}\mathbf{1}$$

which gives Algorithm 5.2.1. □

5.B Algorithm of computing proximal operator of weighted TV penalization

We recall here the algorithm given in Alaya et al. [2015] for computing the proximal operator of weighted total-variation penalization. The latter is defined as follows

$$\beta = \text{prox}_{\|\cdot\|_{\text{TV},\hat{w}}}(\theta) \in \text{argmin}_{\theta\in\mathbb{R}^m}\left\{\frac{1}{2}\|\beta-\theta\|_2^2 + \|\theta\|_{\text{TV},\hat{w}}\right\}. \tag{5.16}$$

The proposed algorithm consists in running forwardly through the samples $(\theta_1,\ldots,\theta_m)$. Using the Karush-Kuhn-Tucker (KKT) optimality conditions for a convex optimization [Boyd and Vandenberghe, 2004], at location k, β_k stays constant where $|u_k| < \hat{w}_{k+1}$. Here u_k is a solution to a dual problem associated to the primal problem (5.16). If this is not possible, it goes back to the last location where a jump can be introduced in β, validates the current segment until this location, starts a new segment, and continues. This algorithm is described precisely in Algorithm 3.

Algorithm 3: Proximal operator of weighted TV penalization

Input: vector $\theta = (\theta_1, \dots, \theta_m)^\top \in \mathbb{R}^m$ and weights $\hat{w} = (\hat{w}_1, \dots, \hat{w}_m) \in \mathbb{R}_+^m$.
Output: vector $\beta = \text{prox}_{\|\cdot\|_{\text{TV},\hat{w}}}(\theta)$

1. **Set** $k = k_0 = k_- = k_+ \leftarrow 1$
$\beta_{\min} \leftarrow \theta_1 - \hat{w}_2$; $\beta_{\max} \leftarrow \theta_1 + \hat{w}_2$
$u_{\min} \leftarrow \hat{w}_2$; $u_{\max} \leftarrow -\hat{w}_2$

2. **if** $k = m$ **then**
  $\beta_m \leftarrow \beta_{\min} + u_{\min}$

3. **if** $\theta_{k+1} + u_{\min} < \beta_{\min} - \hat{w}_{k+2}$ **then** `/* negative jump */`
  $\beta_{k_0} = \dots = \beta_{k_-} \leftarrow \beta_{\min}$
  $k = k_0 = k_- = k_+ \leftarrow k_- + 1$
  $\beta_{\min} \leftarrow \theta_k - \hat{w}_{k+1} + \hat{w}_k$; $\beta_{\max} \leftarrow \theta_k + \hat{w}_{k+1} + \hat{w}_k$
  $u_{\min} \leftarrow \hat{w}_{k+1}$; $u_{\max} \leftarrow -\hat{w}_{k+1}$

4. **else if** $\theta_{k+1} + u_{\max} > \beta_{\max} + \hat{w}_{k+2}$ **then** `/* positive jump */`
  $\beta_{k_0} = \dots = \beta_{k_+} \leftarrow \beta_{\max}$
  $k = k_0 = k_- = k_+ \leftarrow k_+ + 1$
  $\beta_{\min} \leftarrow \theta_k - \hat{w}_{k+1} - \hat{w}_k$; $\beta_{\max} \leftarrow \theta_k + \hat{w}_{k+1} - \hat{w}_k$
  $u_{\min} \leftarrow \hat{w}_{k+1}$; $u_{\max} \leftarrow -\hat{w}_{k+1}$

5. **else** `/* no jump */`
  set $k \leftarrow k + 1$
    $u_{\min} \leftarrow \theta_k + \hat{w}_{k+1} - \beta_{\min}$
    $u_{\max} \leftarrow \theta_k - \hat{w}_{k+1} - \beta_{\max}$ **if** $u_{\min} \geq \hat{w}_{k+1}$ **then**
    $\beta_{\min} \leftarrow \beta_{\min} + \frac{u_{\min} - \hat{w}_{k+1}}{k - k_0 + 1}$
    $u_{\min} \leftarrow \hat{w}_{k+1}$
    $k_- \leftarrow k$
  if $u_{\max} \leq -\hat{w}_{k+1}$ **then**
    $\beta_{\max} \leftarrow \beta_{\max} + \frac{u_{\max} + \hat{w}_{k+1}}{k - k_0 + 1}$
    $u_{\max} \leftarrow -\hat{w}_{k+1}$
    $k_+ \leftarrow k$

6. **if** $k < m$ **then**
  go to **3.**

7. **if** $u_{\min} < 0$ **then**
  $\beta_{k_0} = \dots = \beta_{k_-} \leftarrow \beta_{\min}$
  $k = k_0 = k_- \leftarrow k_- + 1$
  $\beta_{\min} \leftarrow \theta_k - \hat{w}_{k+1} + \hat{w}_k$
  $u_{\min} \leftarrow \hat{w}_{k+1}$; $u_{\max} \leftarrow \theta_k + \hat{w}_k - v_{\max}$
  go to **2.**

8. **else if** $u_{\max} > 0$ **then**
  $\beta_{k_0} = \dots = \beta_{k_+} \leftarrow \beta_{\max}$
  $k = k_0 = k_+ \leftarrow k_+ + 1$
  $\beta_{\max} \leftarrow \theta_k + \hat{w}_{k+1} - \hat{w}_k$
  $u_{\max} \leftarrow -\hat{w}_{k+1}$; $u_{\min} \leftarrow \theta_k - \hat{w}_k - u_{\min}$
  go to **2.**

9. **else**
  $\beta_{k_0} = \dots = \beta_m \leftarrow \beta_{\min} + \frac{u_{\min}}{k - k_0 + 1}$

5.C Proof of Theorem 5.3.1 : fast oracle inequality under binarsity

The proof relies on some technical properties given below.

Additional notation. Hereafter, we use the following vector notations

$$\begin{aligned}\boldsymbol{y} &= (y_1, \dots, y_n)^\top,\\ m^0(\boldsymbol{X}) &= \Big(m^0(x_1), \dots, m^0(x_n)\Big)^\top,\\ m_\theta(\boldsymbol{X}) &= \Big(m_\theta(x_1), \dots, m_\theta(x_n)\Big)^\top,\\ \text{and } b'(m_\theta(\boldsymbol{X})) &= \Big(b'(m_\theta(x_1)), \dots, b'(m_\theta(x_n))\Big)^\top,\end{aligned}$$

where we recall that $m_\theta(x_i) = \theta^\top x_i^B$).

5.C.1 Empirical Kullback-Leibler divergence.

Let us now define the Kullback-Leibler divergence between the true probability density funtion f^0 defined in (5.2) and a candidate f_θ within the generalized linear model, that is

$$f_\theta(y|x) = \exp\Big(y m_\theta(x) - b(m_\theta(x))\Big),$$

as follows

$$\begin{aligned}KL_n(f^0, f_\theta) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{P}_{\mathbf{y}|X}}\left[\log\frac{f^0(y_i|x_i)}{f_\theta(y_i|x_i)}\right]\\ &:= KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),\end{aligned}$$

where $\mathbb{P}_{\mathbf{y}|X}$ is the joint distribution of $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ given $\boldsymbol{X} = (x_1, \dots, x_n)^\top$. We then have the following property.

Lemma 5.C.1 *The excess risk verifies* $R(m_\theta) - R(m^0) = \phi KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X}))$.

Proof. Straightforwardly, one has

$$\begin{aligned}&KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X}))\\ &\quad= \phi^{-1}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{P}_{\mathbf{y}|X}}\Big[\Big(-y_i m_\theta(x_i) + b(m_\theta(x_i))\Big) - \Big(-y_i m^0(x_i) + b(m^0(x_i))\Big)\Big]\\ &\quad= \phi^{-1}\Big(R(m_\theta) - R(m^0)\Big).\end{aligned}$$

□

5.C.2 Optimality conditions.

To characterize the solution of the problem (5.7), the following result can be straightforwardly obtained using the Karush-Kuhn-Tucker (KKT) optimality conditions for a convex optimization [Boyd and Vandenberghe, 2004].

Proposition 5.C.2 *A vector* $\hat{\theta} = (\hat{\theta}_{1,\bullet}^\top, \dots, \hat{\theta}_{p,\bullet}^\top)^\top \in \mathbb{R}^d$ *is an optimum of the objective function in* (5.7) *if and only if there exists a sequence of subgradients*

$$\hat{h} = (\hat{h}_{j,\bullet})_{j=1,\dots,p} \in \partial\Big(\|\hat{\theta}\|_{\mathrm{TV},\hat{w}}\Big)$$

and

$$\hat{g} = (\hat{g}_{j,\bullet})_{j=1,\dots,p} \in \partial\Big(\delta_1(\hat{\theta}_{j,\bullet})\Big)_{j=1,\dots,p}$$

such that

$$\nabla R_n(\hat{\theta}_{j,\bullet}) + \hat{h}_{j,\bullet} + \hat{g}_{j,\bullet} = \mathbf{0}_{d_j},$$

where

$$\begin{cases} \hat{h}_{j,\bullet} = D_j^\top\Big(\hat{w}_{j,\bullet} \odot \mathrm{sign}(D_j\hat{\theta}_{j,\bullet})\Big) & \text{if } j \in J(\hat{\theta}), \\ \hat{h}_{j,\bullet} \in D_j^\top\Big(\hat{w}_{j,\bullet} \odot [-1,+1]^{d_j}\Big) & \text{if } j \in J^{\complement}(\hat{\theta}), \end{cases} \tag{5.17}$$

and where $J(\hat{\theta})$ *is the active set of* $\hat{\theta}$*. The subgradient* $\hat{g}_{j,\bullet}$ *belongs to*

$$\partial\Big(\delta_1(\hat{\theta}_{j,\bullet})\Big) = \Big\{\mu_{j,\bullet} \in \mathbb{R}^{d_j} : \langle \mu_{j,\bullet}, \theta_{j,\bullet}\rangle \le \langle \mu_{j,\bullet}, \hat{\theta}_{j,\bullet}\rangle \text{ for all } \theta_{j,\bullet} \text{ such that } \mathbf{1}^\top\theta_{j,\bullet} = 0\Big\}.$$

For the generalized linear model, we have

$$\frac{1}{n}\Big(\boldsymbol{X}_{\bullet,j}^B\Big)^\top\Big(b'(m_{\hat{\theta}}(\boldsymbol{X})) - \boldsymbol{y}\Big) + \hat{h}_{j,\bullet} + \hat{g}_{j,\bullet} + \hat{f}_{j,\bullet} = \mathbf{0}_{d_j}, \tag{5.18}$$

where $\hat{f} = (\hat{f}_{j,\bullet})_{j=1,\dots,p}$ *belongs to the normal cone of the ball* $B_d(\rho)$*.*

Proof. We denote by $\partial(\phi)$ the subdifferential mapping of a convex functional ϕ. The function $\theta \mapsto R_n(\theta)$ is differentiable, so the subdifferential of $R_n(\cdot) + \mathrm{bina}(\cdot)$ at a point $\theta = (\theta_{j,\bullet})_{j=1,\dots,p} \in \mathbb{R}^d$ is given by

$$\partial\Big(R_n(\theta) + \mathrm{bina}(\theta)\Big) = \nabla R_n(\theta) + \partial\Big(\mathrm{bina}(\theta)\Big),$$

where

$$\nabla R_n(\theta) = \Big(\frac{\partial(R_n(\theta))}{\partial(\theta_{1,\bullet})}, \dots, \frac{\partial(R_n(\theta))}{\partial(\theta_{p,\bullet})}\Big)^\top$$

and

$$\partial\Big(\mathrm{bina}(\theta)\Big) = \Big(\partial\Big(\|\theta_{1,\bullet}\|_{\mathrm{TV},\hat{w}_{1,\bullet}}\Big) + \partial\Big(\delta_1(\theta_{1,\bullet})\Big), \dots, \partial\Big(\|\theta_{p,\bullet}\|_{\mathrm{TV},\hat{w}_{p,\bullet}}\Big) + \partial\Big(\delta_1(\theta_{p,\bullet})\Big)\Big)^\top.$$

We have

$$\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} = \|\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}\|_1$$

for all $j = 1, \ldots, p$. Then, by applying some properties of the subdifferential calculus, we get

$$\partial\Big(\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}\Big) = \begin{cases} D_j^\top \operatorname{sign}(\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}) & \text{if } D_j\theta \neq \mathbf{0}_{d_j}, \\ D_j^\top \big(\hat{w}_{j,\bullet} \odot v_j\big) & \text{otherwise }, \end{cases}$$

where $v_j \in [-1, +1]^{d_j}$, for all $j = 1, \ldots, p$. For generalized linear models, we rewrite

$$\hat{\theta} \in \operatorname{argmin}_{\theta\in\mathbb{R}^d} \Big\{R_n(\theta) + \mathrm{bina}(\theta) + \delta_{B_d(\rho)}(\theta)\Big\}, \tag{5.19}$$

where $\delta_{B_d(\rho)}$ is the indicator function for $B_d(\rho)$.

Now, $\hat{\theta} = (\hat{\theta}_{1,\bullet}^\top, \ldots, \hat{\theta}_{p,\bullet}^\top)^\top$ is an optimum of Problem (5.19) if and only if

$$\mathbf{0}_d \in \nabla R_n(m_{\hat{\theta}}) + \partial\Big(\|\hat{\theta}\|_{\mathrm{TV},\hat{w}}\Big) + \partial\Big(\delta_{B_d(\rho)}(\hat{\theta})\Big).$$

Recall that the subdifferential of $\delta_{B_d(\rho)}(\cdot)$ is the normal cone of $B_d(\rho)$, that is

$$\partial\Big(\delta_{B_d(\rho)}(\hat{\theta})\Big) = \Big\{\eta \in \mathbb{R}^d : \langle \eta, \theta\rangle \leq \langle \eta, \hat{\theta}\rangle \text{ for all } \theta \in B_d(\rho)\Big\}. \tag{5.20}$$

Straightforwardly, one obtains

$$\frac{\partial(R_n(\theta))}{\partial(\theta_{j,\bullet})} = \frac{1}{n}(\boldsymbol{X}_{\bullet,j}^B)^\top\Big(b'(m_{\hat{\theta}}(\boldsymbol{X})) - \boldsymbol{y}\Big), \tag{5.21}$$

and equalities (5.21) and (5.20) give equation (5.18), which ends the proof of Proposition 5.C.2. □

5.C.3 Compatibility conditions.

Let us define the block diagonal matrix

$$\mathbf{D} = \mathrm{diag}(\mathbf{D_1}, \ldots, \mathbf{D_p}),$$

with D_j, defined in (5.15), being invertible. We denote its inverse T_j which is defined by the $d_j \times d_j$ lower triangular matrix with entries $(T_j)_{r,s} = 0$ if $r < s$ and $(T_j)_{r,s} = 1$ otherwise. We set

$$\mathbf{T} = \mathrm{diag}(\mathbf{T_1}, \ldots, \mathbf{T_p}).$$

It is clear that $\mathbf{D}^{-1} = \mathbf{T}$. In order to prove Theorem 5.3.1, we need, in addition to Assumption 2, the following results which give a compatibility condition [van de Geer, 2008, van de Geer and Lederer, 2013, Dalalyan et al., 2017] satisfied by the matrix

$\mathbf{T}$ in Lemma 5.C.3 and $\boldsymbol{X}^B\mathbf{T}$ in Lemma 5.C.4. To this end, for any concatenation of subsets $K = [K_1, \dots, K_p]$, we set

$$K_j = \{\tau_j^1, \dots, \tau_j^{b_j}\} \subset \{1, \dots, d_j\} \tag{5.22}$$

for all $j = 1, \dots, p$ and with the convention that $\tau_j^0 = 0$ and $\tau_j^{b_j+1} = d_j + 1$.

Lemma 5.C.3 *Let $\gamma \in \mathbb{R}_+^d$ be a given vector of weights and $K = [K_1, \dots, K_p]$ with K_j given by (5.22) for all $j = 1, \dots, p$. Then for every $u \in \mathbb{R}^d \backslash \{\mathbf{0}_d\}$, we have*

$$\frac{\|\mathbf{T}u\|_2}{\left|\|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1\right|} \geq \kappa_{\mathbf{T},\gamma}(K),$$

where

$$\kappa_{\mathbf{T},\gamma}(K) = \left\{32 \sum_{j=1}^{p} \sum_{k=1}^{d_j} |\gamma_{j,k+1} - \gamma_{j,k}|^2 + 2|K_j| \|\gamma_{j,\bullet}\|_\infty^2 \Delta_{\min,K_j}^{-1}\right\}^{-1/2},$$

and $\Delta_{\min,K_j} = \min_{r=1,\dots b^j} |\tau_j^{r_j} - \tau_j^{r_j-1}|$.

Proof. Using Proposition 3 in Dalalyan et al. [2017], we have

$$\begin{aligned}
&\|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1 \\
&\qquad = \sum_{j=1}^{p} \|u_{K_j} \odot \gamma_{K_j}\|_1 - \|u_{K_j^\complement} \odot \gamma_{K_j^\complement}\|_1 \\
&\qquad \leq \sum_{j=1}^{p} 4\|T_j u_{j,\bullet}\|_2 \left\{2 \sum_{k=1}^{d_j} |\gamma_{j,k+1} - \gamma_{j,k}|^2 + 2(b_j+1)\|\gamma_{j,\bullet}\|_\infty^2 \Delta_{\min,K_j}^{-1}\right\}^{1/2}.
\end{aligned}$$

Applying Hölder's inequality for the right hand side of the last inequality gives

$$\begin{aligned}
&\|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1 \\
&\qquad \leq \|\mathbf{T}u\|_2 \left\{32 \sum_{j=1}^{p} \sum_{k=1}^{d_j} |\gamma_{j,k+1} - \gamma_{j,k}|^2 + 2|K_j| \|\gamma_{j,\bullet}\|_\infty^2 \Delta_{\min,K_j}^{-1}\right\}^{1/2}.
\end{aligned}$$

This completes the proof of Lemma 5.C.3. □

Now, using Assumption 2 and Lemma 5.C.3, we establish a compatibility condition satisfied by the product of matrices $\boldsymbol{X}^B\mathbf{T}$.

Lemma 5.C.4 *Let Assumption 2 holds. Let $\gamma \in \mathbb{R}_+^d$ be a given vector of weights, and $K = [K_1, \dots, K_p]$ such that K_j is given by (5.22) for all $j = 1, \dots, p$. Then, one has*

$$\inf_{u \in \mathscr{C}_{1,\hat{w}}(K) \backslash \{\mathbf{0}_d\}} \left\{\frac{\|\boldsymbol{X}^B\mathbf{T}u\|_2}{\sqrt{n}\left|\|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1\right|}\right\} \geq \kappa_{\mathbf{T},\gamma}(K)\kappa(K), \tag{5.23}$$

where

$$\mathscr{C}_{1,\hat{w}}(K) = \left\{ u \in \mathbb{R}^d : \sum_{j=1}^{p} \|(u_{j,\bullet})_{K_j^\complement}\|_{1,\hat{w}_{j,\bullet}} \leq 2 \sum_{j=1}^{p} \|(u_{j,\bullet})_{K_j}\|_{1,\hat{w}_{j,\bullet}} \right\}, \tag{5.24}$$

with $\|\cdot\|_{1,a}$ *denoting the weighted* ℓ_1*-norm.*

Proof. By Lemma 5.C.3, we have that

$$\frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n}\Big|\|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1\Big|} \geq \kappa_{\mathbf{T},\gamma}(K) \frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n}\|\mathbf{T} u\|_2}.$$

Now, we note that if $u \in \mathscr{C}_{1,\hat{w}}(K)$, then $\mathbf{T}u \in \mathscr{C}_{\mathrm{TV},\hat{w}}(K)$. Hence, by Assumption 2, we get

$$\frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n}\Big|\|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1\Big|} \geq \kappa_{\mathbf{T},\gamma}(K)\kappa(K).$$

□

5.C.4 Connection between empirical Kullback-Leibler divergence and the empirical squared norm.

To compare the empirical Kullback-Leibler divergence and the empirical squared norm, we use Lemma 1 in Bach [2010], that we recall here.

Lemma 5.C.5 *Let* $\varphi : \mathbb{R} \to \mathbb{R}$ *be a convex three times differentiable function such that for all* $t \in \mathbb{R}$, $|\varphi'''(t)| \leq M|\varphi''(t)|$ *for some* $M \geq 0$*. Then, for all* $t \geq 0$*, one has*

$$\frac{\varphi''(0)}{M^2}\psi(-Mt) \leq \varphi(t) - \varphi(0) - \varphi'(0)t \leq \frac{\varphi''(0)}{M^2}\psi(Mt),$$

with $\psi(u) = e^u - u - 1$.

Now, we give a version of the previous Lemma in our setting.

Lemma 5.C.6 *Under Assumption 1, for all* $\theta \in B_d(\rho)$*, one has*

$$\frac{L_n \psi(-2(C_n+\rho))}{\phi(2(C_n+\rho))^2}\frac{1}{n}\|m^0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \leq KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$
$$\frac{U_n \psi(2(C_n+\rho))}{\phi(2(C_n+\rho))^2}\frac{1}{n}\|m^0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \geq KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})).$$

Proof. Let us consider the function $G_n : \mathbb{R} \to \mathbb{R}$ defined by $G_n(t) = R_n(m^0 + tm_\eta)$, then

$$G_n(t) = \frac{1}{n}\sum_{i=1}^{n} b(m^0(x_i) + tm_\eta(x_i)) - \frac{1}{n}\sum_{i=1}^{n} y_i(m^0(x_i) + tm_\eta(x_i)).$$

By differentiating G_n three times with respect to t, we obtain

$$\begin{aligned} G_n'(t) &= \frac{1}{n}\sum_{i=1}^{n} m_\eta(x_i)b'(m^0(x_i) + tm_\eta(x_i)) - \frac{1}{n}\sum_{i=1}^{n} y_i m_\eta(x_i), \\ G_n''(t) &= \frac{1}{n}\sum_{i=1}^{n} m_\eta^2(x_i)b''(m^0(x_i) + tm_\eta(x_i)), \\ \text{and} \quad G_n'''(t) &= \frac{1}{n}\sum_{i=1}^{n} m_\eta^3(x_i)b'''(m^0(x_i) + tm_\eta(x_i)). \end{aligned}$$

In all the considered models, we have $|b'''(z)| \leq 2|b''(z)|$, see the following table

Model	ϕ	$b(z)$	$b'(z)$	$b''(z)$	$b'''(z)$	L_n	U_n
Normal	σ^2	$\frac{z^2}{2}$	z	1	0	1	1
Logistic	1	$\log(1+e^z)$	$\frac{e^z}{1+e^z}$	$\frac{e^z}{(1+e^z)^2}$	$\frac{1-e^z}{1+e^z}b''(z)$	$\frac{e^{C_n}}{(1+e^{C_n})^2}$	$\frac{1}{4}$
Poisson	1	e^z	e^z	e^z	$b''(z)$	e^{-C_n}	e^{C_n}

.

Then, we get

$$|G_n'''(t)| \leq 2\|m_\eta\|_\infty |G_n''(t)|$$

where

$$\|m_\eta\|_\infty := \max_{i=1,\dots,n} |m_\eta(x_i)|.$$

Applying Lemma 5.C.5 with $M = 2\|m_\eta\|_\infty$, we obtain

$$G_n''(0)\frac{\psi(-2\|m_\eta\|_\infty t)}{4\|m_\eta\|_\infty^2} \leq G_n(t) - G_n(0) - tG_n'(0) \leq G_n''(0)\frac{\psi(2\|m_\eta\|_\infty t)}{4\|m_\eta\|_\infty^2}.$$

for all $t \geq 0$. Taking $t = 1$ leads to

$$\begin{aligned} G_n''(0)\frac{\psi(-2\|m_\eta\|_\infty)}{4\|m_\eta\|_\infty^2} &\leq R_n(m^0 + m_\eta) - R_n(m^0) - G_n'(0), \\ G_n''(0)\frac{\psi(2\|m_\eta\|_\infty)}{4\|m_\eta\|_\infty^2} &\geq R_n(m^0 + m_\eta) - R_n(m^0) - G_n'(0). \end{aligned}$$

A short calculation gives that

$$-G_n'(0) = \frac{1}{n}\sum_{i=1}^{n} m_\eta(x_i)\Big(y_i - b'(m^0(x_i))\Big), \text{ and } G_n''(0) = \frac{1}{n}\sum_{i=1}^{n} m_\eta^2(x_i)b''(m_\vartheta(x_i)).$$

It is clear that $\mathbb{E}_{\mathbb{P}_{\mathbf{y}|X}}[-G_n'(0)] = 0$. Then

$$G_n''(0)\frac{\psi(-2\|m_\eta\|_\infty)}{4\|m_\eta\|_\infty^2} \leq R(m^0 + m_\eta) - R(m^0) \leq G_n''(0)\frac{\psi(2\|m_\eta\|_\infty)}{4\|m_\eta\|_\infty^2}.$$

Now choose $m_\eta = m_\theta - m^0$, and using Assumption 1 and Equation(5.8), we have

$$2\|m_\eta\|_\infty \leq 2\max_{i=1,\dots,n}\Big(|\langle x_i^B, \theta\rangle| + |m^0(x_i)|\Big) \leq 2(\rho + C_n).$$

Hence, we obtain

$$G_n''(0)\frac{\psi(-2(C_n+\rho))}{C_n^2(\rho,p)} \leq R(m_\theta) - R(m_0) = \phi KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$
$$G_n''(0)\frac{\psi(2(C_n+\rho))}{C_n^2(\rho,p)} \geq R(m_\theta) - R(m_0) = \phi KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$

with $G_n''(0) = n^{-1}\sum_{i=1}^n \Big(m_\theta(x_i) - m^0(x_i)\Big)^2 b''(m^0(x_i))$. It entails that

$$\frac{L_n\psi(-2(C_n+\rho))}{\phi(2(C_n+\rho))^2}\frac{1}{n}\|m^0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \leq KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X}))$$
$$\leq \frac{U_n\psi(2(C_n+\rho))}{\phi(2(C_n+\rho))^2}\frac{1}{n}\|m^0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2.$$

□

5.C.5 Proof of Theorem 5.3.1.

Recall that for all $\theta \in \mathbb{R}^d$,

$$R_n(m_\theta) = \frac{1}{n}\sum_{i=1}^n b(m_\theta(x_i)) - \frac{1}{n}\sum_{i=1}^n y_i m_\theta(x_i)$$

and

$$\hat{\theta} \in \operatorname{argmin}_{\theta\in B_d(\rho)}\Big\{R_n(\theta) + \operatorname{bina}(\theta)\Big\}. \tag{5.25}$$

According to Proposition 5.C.2, Equation (5.25) involves that there is

$$\hat{h} = (\hat{h}_{j,\bullet})_{j=1,\dots,p} \in \partial\Big(\|\hat{\theta}\|_{\mathrm{TV},\hat{w}}\Big),$$
$$\hat{g} = (\hat{g}_{j,\bullet})_{j=1,\cdots,p} \in \Big(\partial\Big(\delta_1(\hat{\theta}_{j,\bullet})\Big)\Big)_{j=1,\dots,p},$$
$$\text{and } \hat{f} = (\hat{f}_{j,\bullet})_{j=1,\dots,p} \in \partial\Big(\delta_{B_d(\rho)}(\hat{\theta})\Big)$$

such that

$$\langle \frac{1}{n}(\boldsymbol{X}^B)^\top \Big(b'(m_{\hat{\theta}}(\boldsymbol{X})) - \boldsymbol{y}\Big) + \hat{h} + \hat{g} + \hat{f}, \hat{\theta} - \theta \rangle = 0$$

for all $\theta \in \mathbb{R}^d$, which can be written

$$\frac{1}{n}\langle b'(m_{\hat{\theta}}(\boldsymbol{X})) - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle$$
$$- \frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle + \langle \hat{h} + \hat{g} + \hat{f}, \hat{\theta} - \theta\rangle = 0.$$

For any $\theta \in B_d(\rho)$ such that, for all j, $\mathbf{1}^\top \theta_{j,\bullet} = 0$, and $h \in \partial\Big(\|\theta\|_{\mathrm{TV},\hat{w}}\Big)$, the monotony of the subdifferential mapping implies $\langle \hat{h}, \theta - \hat{\theta}\rangle \leq \langle h, \theta - \hat{\theta}\rangle$, $\langle \hat{g}, \theta - \hat{\theta}\rangle \leq 0$, and $\langle \hat{f}, \theta - \hat{\theta}\rangle \leq 0$. Therefore

$$\frac{1}{n}\langle b'(m_{\hat{\theta}}(\boldsymbol{X})) - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle \tag{5.26}$$
$$\leq \frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle - \langle h, \hat{\theta} - \theta\rangle.$$

We consider now the function $H_n : \mathbb{R} \to \mathbb{R}$, defined by

$$H_n(t) = \frac{1}{n}\sum_{i=1}^n b(m_{\hat{\theta}+t\eta}(x_i)) - \frac{1}{n}\sum_{i=1}^n b'(m^0(x_i)) m_{\hat{\theta}+t\eta}(x_i)$$

By differentiating H_n three times with respect t, we obtain

$$H_n'(t) = \frac{1}{n}\sum_{i=1}^n m_\eta(x_i) b'(m_{\hat{\theta}+t\eta}(x_i)) - \frac{1}{n}\sum_{i=1}^n b'(m^0(x_i)) m_\eta(x_i),$$
$$H_n''(t) = \frac{1}{n}\sum_{i=1}^n m_\eta^2(x_i) b''(m_{\hat{\theta}+t\eta}(x_i)),$$
$$\text{and} \quad H_n'''(t) = \frac{1}{n}\sum_{i=1}^n m_\eta^3(x_i) b'''(m_{\hat{\theta}+t\eta}(x_i)).$$

In the way as in the proof of Lemma 5.C.6, we have $|H_n'''(t)| \leq 2\rho |H_n''(t)|$, applying now Lemma 5.C.5, we obtain

$$H_n''(0)\frac{\psi(-t2\rho)}{4\rho^2} \leq H_n(t) - H_n(0) - tH_n'(0) \leq H_n''(0)\frac{\psi(t2\rho)}{4\rho^2},$$

for all $t \geq 0$. Taking $t = 1$ and $\eta = \theta - \hat{\theta}$ implies

$$H_n(1) = \frac{1}{n}\sum_{i=1}^n b(m_\theta(x_i)) - \frac{1}{n}\sum_{i=1}^n b'(m^0(x_i)) m_\theta(x_i) = R(m_\theta),$$
$$\text{and} \quad H_n(0) = \frac{1}{n}\sum_{i=1}^n b(m_{\hat{\theta}}(x_i)) - \frac{1}{n}\sum_{i=1}^n b'(m^0(x_i)) m_{\hat{\theta}}(x_i) = R(m_{\hat{\theta}}).$$

Moreover, we have

$$\begin{aligned} H_n'(0) &= \frac{1}{n}\sum_{i=1}^{n}\langle x_i^B, \theta-\hat{\theta}\rangle b'(m_{\hat{\theta}}(x_i)) - \frac{1}{n}\sum_{i=1}^{n} b'(m^0(x_i))\langle x_i^B, \hat{\theta}-\theta\rangle \\ &= \frac{1}{n}\langle b'(m_{\hat{\theta}}(\boldsymbol{X})) - b'(m^0(\boldsymbol{X})), \boldsymbol{X}^B(\theta-\hat{\theta})\rangle, \\ \text{and} \quad H_n''(0) &= \frac{1}{n}\sum_{i=1}^{n}\langle x_i^B, \hat{\theta}-\theta\rangle^2 b''(m_{\hat{\theta}}(x_i)). \end{aligned}$$

Then, we deduce that

$$\begin{aligned} H_n''(0)\frac{\psi(-2\rho)}{4\rho^2} &\leq R(m_\theta) - R(m_{\hat{\theta}}) - \frac{1}{n}\langle b'(m_{\hat{\theta}}(\boldsymbol{X})) - b'(m^0(\boldsymbol{X})), \boldsymbol{X}^B(\theta-\hat{\theta})\rangle \\ &= \phi KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) - \phi KL_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\ &\quad + \frac{1}{n}\langle b'(m_{\hat{\theta}}(\boldsymbol{X})) - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle. \end{aligned}$$

Then, with Equation (5.26), one has

$$\begin{aligned} \phi KL_n(m^0(\boldsymbol{X}),&m_{\hat{\theta}}(\boldsymbol{X})) + H_n''(0)\frac{\psi(-2\rho)}{4\rho^2} \\ &\leq \phi KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) \\ &\quad + \frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle - \langle h, \hat{\theta}-\theta\rangle. \end{aligned} \tag{5.27}$$

As $H_n''(0) \geq 0$, it implies that

$$\begin{aligned} \phi KL_n(m^0(\boldsymbol{X}),&m_{\hat{\theta}}(\boldsymbol{X})) \leq \phi KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) \\ &\quad + \frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle - \langle h, \hat{\theta}-\theta\rangle. \end{aligned} \tag{5.28}$$

If

$$\frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta}-\theta)\rangle - \langle h, \hat{\theta}-\theta\rangle < 0,$$

it follows that

$$KL_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$

then Theorem 5.3.1 holds. From now on, let us assume that

$$\frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle - \langle h, \hat{\theta}-\theta\rangle \geq 0.$$

We first derive a bound on

$$\frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle.$$

Using $\mathbf{D}^{-1} = \mathbf{T}$, we focus on finding out a bound of

$$\frac{1}{n}\langle (\boldsymbol{X}^B\mathbf{T})^\top(\boldsymbol{y} - b'(m^0(\boldsymbol{X}))), \mathbf{D}(\hat{\theta} - \theta)\rangle.$$

In one hand, one has

$$\begin{aligned}\frac{1}{n}\langle (\boldsymbol{X}^B)^\top(\boldsymbol{y} - b'(m^0(\boldsymbol{X}), \hat{\theta} - \theta\rangle &= \frac{1}{n}\langle (\boldsymbol{X}^B\mathbf{T})^\top(\boldsymbol{y} - b'(m^0(\boldsymbol{X}), \mathbf{D}(\hat{\theta} - \theta)\rangle \\ &\leq \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{d_j}\Big|\langle (\boldsymbol{X}^B_{\bullet,j}T_j)_{\bullet,k}, \boldsymbol{y} - b'(m^0(\boldsymbol{X}))\rangle\Big|\Big|\Big(D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})\Big)_k\Big|,\end{aligned}$$

where

$$(\boldsymbol{X}^B_{\bullet,j}T_j)_{\bullet,k} = \Big((\boldsymbol{X}^B_{\bullet,j}T_j)_{1,k}, \dots, (\boldsymbol{X}^B_{\bullet,j}T_j)_{n,k}\Big)^\top \in \mathbb{R}^n$$

is the k-th column of the matrix $(\boldsymbol{X}^B_{\bullet,j}T_j)$. Let us consider the event

$$\mathcal{E}_n = \bigcap_{j=1}^{p}\bigcap_{k=2}^{d_j}\mathcal{E}_{n,j,k},$$

where

$$\mathcal{E}_{n,j,k} = \Big\{\frac{1}{n}\Big|\langle (\boldsymbol{X}^B_{\bullet,j}T_j\Big)_{\bullet,k}, \boldsymbol{y} - b'(m^0(\boldsymbol{X}))\rangle\Big| \leq \hat{w}_{j,k}\Big\}.$$

Then, on $\mathcal{E}_n$, we have

$$\begin{aligned}\frac{1}{n}\langle (\boldsymbol{X}^B)^\top(\boldsymbol{y} - b'(m^0(\boldsymbol{X})), \hat{\theta} - \theta\rangle &\leq \sum_{j=1}^{p}\sum_{k=1}^{d_j}\hat{w}_{j,k}\Big|\Big(D_j\Big(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet}\Big)\Big)_k\Big| \\ &\leq \sum_{j=1}^{p}\|\hat{w}_{j,\bullet} \odot D_j\Big(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet}\Big)\|_1. \qquad (5.29)\end{aligned}$$

In another hand, from the definition of the subgradient $(h_{j,\bullet})_{j=1,\dots,p} \in \partial\Big(\|\theta\|_{\mathrm{TV},\hat{w}}\Big)$ (see Equation (5.17)), one can choose h such that

$$h_{j,k} = \Big(D_j^\top\Big(\hat{w}_{j,\bullet} \odot \mathrm{sign}(D_j\theta_{j,\bullet})\Big)\Big)_k$$

for all $k = 1, \dots, J_j(\theta)$ and

$$h_{j,k} = \Big(D_j^\top\Big(\hat{w}_{j,\bullet} \odot \mathrm{sign}\Big(D_j\hat{\theta}_{j,\bullet}\Big)\Big)\Big)_k = \Big(D_j^\top\Big(\hat{w}_{j,\bullet} \odot \mathrm{sign}\Big(D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})\Big)\Big)\Big)_k$$

for all $k = 1, \dots, J_j^{\complement}(\theta)$. Using a triangle inequality and the fact that

$$\langle \mathrm{sign}(x), x\rangle = \|x\|_1,$$

we obtain

$$
\begin{aligned}
-\langle h, \hat{\theta} - \theta \rangle &\leq \sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)} \odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_1 \\
&\quad - \sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j^{\complement}(\theta)} \odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j^{\complement}(\theta)}\|_1 \\
&\leq \sum_{j=1}^{p} \|(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} - \sum_{j=1}^{p} \|(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j^{\complement}(\theta)}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}. \quad (5.30)
\end{aligned}
$$

Combining inequalities (5.29) and (5.30), we get

$$
\sum_{j=1}^{p} \|(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j^{\complement}(\theta)}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} \leq 2 \sum_{j=1}^{p} \|(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}
$$

on $\mathcal{E}_n$. Hence

$$
\sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j^{\complement}(\theta)} \odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j^{\complement}(\theta)}\|_1 \leq 2 \sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)} \odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_1.
$$

This means that

$$
\hat{\theta} - \theta \in \mathscr{C}_{\mathrm{TV},\hat{w}}(J(\theta)) \text{ and } \mathbf{D}(\hat{\theta} - \theta) \in \mathscr{C}_{1,\hat{w}}(J(\theta)), \quad (5.31)
$$

see (6.3) and (5.24). Now, going back to (5.28) and taking into account (5.31), the compatibility of $\boldsymbol{X}^B\mathbf{T}$ (see (5.23)), on $\mathcal{E}_n$ the following holds

$$
\begin{aligned}
&\phi KL_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\
&\quad \leq \phi KL_n(m^0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X})) + 2 \sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)} \odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_1.
\end{aligned}
$$

Then

$$
\begin{aligned}
&KL_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\
&\quad \leq KL_n(m^0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X})) + \frac{\|m_{\hat{\theta}}(\boldsymbol{X}) - m_{\theta}(\boldsymbol{X})\|_2}{\sqrt{n}\, \phi\, \kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))}, \quad (5.32)
\end{aligned}
$$

where $\hat{\gamma} = (\hat{\gamma}_{1,\bullet}^{\top}, \ldots, \hat{\gamma}_{p,\bullet}^{\top})^{\top}$ such that

$$
\hat{\gamma}_{j,k} = \begin{cases} 2\hat{w}_{j,k} & \text{if } k \in J_j(\theta), \\ 0 & \text{if } k \in J_j^{\complement}(\theta), \end{cases}
$$

for all $j = 1, \ldots, p$ and

$$
\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta)) = \left\{ 32 \sum_{j=1}^{p} \sum_{k=1}^{d_j} |\hat{\gamma}_{j,k+1} - \hat{\gamma}_{j,k}|^2 + 2|J_j(\theta)| \|\hat{\gamma}_{j,\bullet}\|_{\infty}^2 \Delta_{\min,J_j(\theta)}^{-1} \right\}^{-1/2}.
$$

Next, we find an upper bound for

$$\left(\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))\right)^{-1}.$$

We have

$$\frac{1}{\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))} = 32\sum_{j=1}^{p}\sum_{k=1}^{d_j}|\hat{\gamma}_{j,k+1}-\hat{\gamma}_{j,k}|^2 + 2|J_j(\theta)|\|\hat{\gamma}_{j,\bullet}\|^2_\infty \Delta^{-1}_{\min,J_j(\theta)}.$$

Note that one has

$$\|\hat{\gamma}_{j,\bullet}\|_\infty \leq 2\|\hat{w}_{j,\bullet}\|_\infty.$$

We write the set

$$J_j(\theta) = \left\{k_j^1,\ldots,k_j^{|J_j(\theta)|}\right\}$$

and we set

$$B_r = [\![k_j^{r-1}, k_j^r[\![= \{k_j^{r-1}, k_j^{r-1}+1,\ldots,k_j^r-1\}$$

for $r = 1,\ldots,|J_j(\theta)|+1$ with the convention that $k_j^0 = 0$ and $k_j^{|J_j(\theta)|+1} = d_j+1$. Then

$$\begin{aligned}
\sum_{k=1}^{d_j}|\hat{\gamma}_{j,k+1}-\hat{\gamma}_{j,k}|^2 &= \sum_{r=1}^{|J_j(\theta)|+1}\sum_{k\in B_r}|\hat{\gamma}_{j,k+1}-\hat{\gamma}_{j,k}|^2\\
&= \sum_{r=1}^{|J_j(\theta)|+1}|\hat{\gamma}_{j,k_j^{r-1}+1}-\hat{\gamma}_{j,k_j^{r-1}}|^2 + |\hat{\gamma}_{j,k_j^r}-\hat{\gamma}_{j,k_j^r-1}|^2\\
&= \sum_{r=1}^{|J_j(\theta)|+1}\hat{\gamma}^2_{j,k_j^{r-1}} + \hat{\gamma}^2_{j,k_j^r}\\
&= \sum_{r=1}^{|J_j(\theta)|} 2\ \hat{\gamma}^2_{j,k_j^r}\\
&\leq 8\ |J_j(\theta)|\ \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty.
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{1}{\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))} &\leq 32\sum_{j=1}^{p}\left\{8\ |J_j(\theta)|\ \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty\right\} + 8\ |J_j(\theta)|\ \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty\Delta^{-1}_{\min,J_j(\theta)}\\
&\leq (32\times 8)\sum_{j=1}^{p}\left\{1+\frac{1}{\Delta_{\min,J_j(\theta)}}\right\}|J_j(\theta)|\|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty\\
&\leq 512|J(\theta)|\max_{j=1,\ldots,p}\|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty.
\end{aligned}$$

Remark 5.C.1 *For the case of least squares regression where $y_i|x_i$ has Gaussian distribution with mean $m^0(x_i)$ and variance $\phi = \sigma^2$. Using inequalities (5.27) and (5.32), we get*

$$\begin{aligned}
&\phi KL_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) + \frac{\psi(-2\rho)}{4\rho^2}\frac{1}{n}\|m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \\
&\quad \leq \phi KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + \frac{\|m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2}{\sqrt{n}\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))} \\
&\quad \leq \phi KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) \\
&\qquad + 2\frac{\sqrt{\psi(-2\rho)}}{2\rho}\frac{1}{\sqrt{n}}\|m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2 \frac{2\rho}{\sqrt{\psi(-2\rho)}\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))}
\end{aligned}$$

Using the fact that $2uv \leq u^2 + v^2$ it yields

$$\phi KL_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq \phi KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + \frac{4\rho^2}{\psi(-2\rho)\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa^2(J(\theta))}$$

Hence, we derive the following sharp oracle inequality

$$R(m_{\hat{\theta}}) - R(m^0) \leq \inf_{\theta \in B_d(\rho)} \left\{ R(m_\theta) - R(m^0) + \frac{\xi|J(\theta)|}{\kappa^2(J(\theta))} \max_{j=1,\ldots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2 \right\},$$

where

$$\xi = \frac{5124\rho^2}{\psi(-2\rho)}.$$

Now for generalized linear models, we use the connection between the empirical norm and the Kullback-Leibler divergence. First, We have

$$\begin{aligned}
&\frac{\|m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2}{\sqrt{n}\phi\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))} \\
&\quad \leq \frac{1}{\phi\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))}\left(\frac{1}{\sqrt{n}}\|m_{\hat{\theta}}(\boldsymbol{X}) - m^0(\boldsymbol{X})\|_2 + \frac{1}{\sqrt{n}}\|m^0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2\right).
\end{aligned}$$

Therefore, by Lemma 5.C.6, we get

$$\begin{aligned}
&\frac{\|m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2}{\sqrt{n}\phi\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))} \\
&\qquad \leq \frac{2}{\sqrt{\phi}\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))}\left(\sqrt{C_n(\rho, L_n)^{-1} KL_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X}))}\right. \\
&\qquad\qquad \left. + \sqrt{C_n(\rho, L_n)^{-1} KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X}))}\right).
\end{aligned}$$

We now use the elementary inequality

$$2uv \leq \epsilon u^2 + v^2/\epsilon$$

with $\epsilon > 0$.

Therefore (5.32) becomes

$$\begin{aligned} KL_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + \frac{\epsilon}{\phi \kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa^2(J(\theta))} \\ + 2\Big(\epsilon\, C_n(\rho, L_n)\Big)^{-1} KL_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\ + 2\Big(\epsilon\, C_n(\rho, L_n)\Big)^{-1} KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})). \end{aligned}$$

By choosing

$$2\Big(\epsilon\, C_n(\rho, L_n)\Big)^{-1} < 1,$$

we get

$$\begin{aligned} KL_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) &\leq \frac{1 + 2\Big(\epsilon\, C_n(\rho, L_n)\Big)^{-1}}{1 - 2\Big(\epsilon\, C_n(\rho, L_n)\Big)^{-1}} KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) \\ &\quad + \frac{\epsilon^2}{\Big(1 - 2(\epsilon\, C_n(\rho, L_n))^{-1}\Big)\phi \kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa^2(J(\theta))} \\ &\leq \frac{\epsilon\, C_n(\rho, L_n) + 2}{\epsilon\, C_n(\rho, L_n) - 2} KL_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) \\ &\quad + \frac{\epsilon^2\, C_n(\rho, L_n)}{(\epsilon\, C_n(\rho, L_n) - 2)\phi \kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa^2(J(\theta))}. \end{aligned}$$

Setting

$$\frac{\epsilon\, C_n(\rho, L_n) + 2}{\epsilon\, C_n(\rho, L_n) - 2} = 1 + \frac{4}{\epsilon\, C_n(\rho, L_n) - 2} = 1 + \zeta,$$

we get the desired result in (5.12).

Finally, we have to compute the probability of the complementary of the event $\mathcal{E}_n$. This is given by the following.

$$\begin{aligned} \mathbb{P}[\mathcal{E}_n^{\complement}] &\leq \sum_{j=1}^{p} \sum_{k=2}^{d_j} \mathbb{P}\left[\frac{1}{n}\left|\langle \left(\boldsymbol{X}^B_{\bullet,j} T_j\right)_{\bullet,k}, \boldsymbol{y} - b'(m^0(\boldsymbol{X}))\rangle\right| \geq \hat{w}_{j,k}\right] \\ &\leq \sum_{j=1}^{p} \sum_{k=2}^{d_j} \mathbb{P}\left[\sum_{i=1}^{n} \left|(\boldsymbol{X}^B_{\bullet,j} T_j)_{i,k}(y_i - b'(m^0(x_i)))\right| \geq n\hat{w}_{j,k}\right]. \end{aligned}$$

Let

$$\xi_{i,j,k} = (\boldsymbol{X}^B_{\bullet,j} T_j)_{i,k},$$

and

$$Z_i = y_i - b'(m^0(x_i)).$$

Note that conditionally on x_i, the random variables (Z_i) are independent. It can be easily shown (see Theorem 5.10 in Lehmann and Casella [1998]) that the moment generating function of Z (copy of Z_i) is given by

$$\mathbb{E}[\exp(tZ)] = \exp\Big(\phi^{-1}\Big\{b(m^0(x)+t) - tb'(m^0(x) - b(m^0(x)))\Big\}\Big). \tag{5.33}$$

Applying Lemma 6.1 in Rigollet [2012], using (5.33) and Assumption 1, we can derive the following Chernoff-type bounds

$$\mathbb{P}\Big[\sum_{i=1}^{n} |\xi_{i,j,k} Z_i| \geq n\hat{w}_{j,k}\Big] \leq 2\exp\Big(-\frac{n^2\hat{w}_{j,k}^2}{2U_n\phi\|\xi_{\bullet,j,k}\|_2^2}\Big), \tag{5.34}$$

where

$$\xi_{\bullet,j,k} = (\xi_{1,j,k}, \ldots, \xi_{n,j,k})^\top \in \mathbb{R}^n.$$

We have

$$\boldsymbol{X}^B_{\bullet,j} T_j = \begin{bmatrix} 1 & \sum_{k=2}^{d_j} x^B_{1,j,k} & \sum_{k=3}^{d_j} x^B_{1,j,k} & \cdots & \sum_{k=d_{j-1}}^{d_j} x^B_{1,j,k} & x^B_{1,j,d_j} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & \sum_{k=2}^{d_j} x^B_{n,j,k} & \sum_{k=3}^{d_j} x^B_{n,j,k} & \cdots & \sum_{k=d_{j-1}}^{d_j} x^B_{n,j,k} & x^B_{n,j,d_j} \end{bmatrix}.$$

Therefore,

$$\|\xi_{\bullet,j,k}\|_2^2 = \sum_{i=1}^{n} (\boldsymbol{X}^B_{\bullet,j} T_j)^2_{\bullet,k} = \#\Big(\Big\{i \in [n] : x_{i,j} \in \bigcup_{r=k}^{d_j} I_{j,r}\Big\}\Big) = n\hat{\pi}_{j,k}. \tag{5.35}$$

Using weights $\hat{w}_{j,k}$ (see (5.11) in Theorem 5.3.1), and (5.34) together with (5.35), we find that the probability of the complementary event $\mathcal{E}_n^\complement$ is smaller than $2e^{-A}$. This concludes the proof of Theorem 5.3.1. □

Chapitre 6

Binacox : automatic cut-points detection in high-dimensional Cox model

Sommaire

Abstract. Determining significant prognostic biomarkers is of increasing importance in many areas of medicine. In order to translate a continuous biomarker into a clinical decision, it is often necessary to determine cut-points. There is so far no standard method to help evaluate how many cut-points are optimal for a given feature in a survival analysis setting. Moreover, most existing methods are univariate, hence not well suited for high-dimensional frameworks. This chapter introduces a prognostic method called *binacox* to deal with the problem of detecting multiple cut-points per features in a multivariate setting where a large number of continuous features are available. It is based on the Cox model and combines one-hot encodings with the binarsity penalty. This penalty uses total-variation regularization together with an extra linear constraint to avoid collinearity between the one-hot encodings and enable feature selection. A non-asymptotic oracle inequality is established. The statistical performance of the method is then examined on an extensive Monte Carlo simulation study, and finally illustrated on three publicly available genetic cancer datasets with high-dimensional features. On this datasets, our proposed methodology significantly outperforms the state-of-the-art survival models regarding risk prediction in terms of C-index, with a computing time orders of magnitude faster. In addition, it provides powerful interpretability by automatically pinpointing significant cut-points on relevant features from a clinical point of view.

Résumé. La détermination de biomarqueurs pronostiques significatifs est d'un intérêt croissant dans de nombreux domaines de la médecine. Pour traduire la donnée d'un biomarqueur continu en décision clinique, il est souvent nécessaire de déterminer des seuils. Il n'y a jusqu'à présent aucune méthode standard pour aider à évaluer le nombre optimal de seuils pour une covariable donnée dans un contexte d'analyse de survie. De plus, la plupart des méthodes existantes sont univariées, donc peu adaptées au cadre de la grande dimension. Ce chapitre introduit une méthode pronostique appelée *binacox* pour traiter le problème de la détection de multiples seuils par covariable dans un cadre multivarié, et lorsqu'un grand nombre de covariables continues sont disponibles. Celle-ci est basée sur le modèle de Cox et combine l'encodage "one-hot" avec la pénalité binarsity. Cette pénalité utilise la régularisation par variation totale ainsi qu'une contrainte linéaire pour traiter le problème de colinéarité dans les encodages binaires, et permet de faire de la sélection de variables. Une inégalité oracle non-asymptotique est établie. Les performances de la méthode sont ensuite examinées lors d'une étude de simulation de Monte Carlo, et finalement illustrées sur trois jeux de données génétiques publiques de grande dimension en cancérologie. Sur ces jeux de données, notre méthode surpasse significativement les modèles de survie usuels en terme de prédiction du risque, évalué par le C-index, tout en ayant des temps de calcul bien moindre. En outre, la méthode fournit une interprétabilité puissante en identifiant automatiquement des seuils significatifs sur des covariables qui s'avèrent pertinentes d'un point de vue clinique.

6.1 Introduction

In any medical applications, the effects of certain clinical variables on the prognostic are sometimes known, but their precise roles remain to be clearly established. For instance, in a breast cancer study, there is reasonable agreement that younger patients have higher risk of an unfavourable outcome, but there is little agreement on the exact nature of the relationship between age and prognosis. Similar issues occur in genetic oncology studies where gene expressions effects on survival times are often non-linear.

The cut-points detection problem. A simple and popular way to treat this problem consists in determining cut-off values, or cut-points, of the continuous features (*e.g.* the age or the gene expressions in the previous examples). This technique brings to light potential non-linearities on feature effects that most models cannot detect. It also offers the ability to classify patients into several groups regarding its features values relatively to the cut-points. More importantly, it can lead to a better understanding of the features effects on the outcome under study. A convenient tool to find optimal cut-points is, therefore, of high interest.

Hence, cut-points detection is a widespread issue in many medical studies and multiple methods have been proposed to determine a single cut-point for a given feature . They range from choosing the mean or median value to methods based on distribution of values or association with clinical outcomes, such as the minimal p-value of multiple log-rank tests, see for instance Camp et al. [2004], Moul et al. [2007], Rota et al. [2015] among many others. However, the choice of the actual cut-points is not a straightforward problem, even for one single cut-point [Lausen and Schumacher, 1992, Klein and Wu, 2003, Contal and O'Quigley, 1999].

While many studies have been devoted to find one optimal cut-point, there is often need in practical medicine to determine not only one, but multiple cut-points. Some method deal with multiple cut-points detection for one-dimensional signals (see for instance Bleakley and Vert [2011] and Harchaoui and Lévy-Leduc [2010] that use a group fused lasso and total variation penalties respectively) or for multivariate time series (see Cho and Fryzlewicz [2015]). Whereas cut-points detection is known to be a paramount issue in survival analysis also [Faraggi and Simon, 1996], the corresponding developed methods are looking only at a single feature at a time (*e.g.* Motzer et al. [1999] or LeBlanc and Crowley [1993] with the survival trees). To our knowledge,no multivariate survival analysis method well suited to detect multiple cut-points per feature in a high-dimensional setting has yet been proposed.

General framework. Let us consider the usual survival analysis framework. Following Andersen et al. [1993], let non-negative random variables T and C stand

for the times of the event of interest and censoring times respectively, and X denotes the p-dimensional vector of features (*e.g.* patients characteristics, therapeutic strategy, omics features). The event of interest could be for instance survival time, re-hospitalization, relapse or disease progression. Conditional on X, T and C are assumed to be independent, which is classical in survival analysis [Klein and Moeschberger, 2005]. We then denote Z the right-censored time and Δ the censoring indicator, defined as

$$Z = T \wedge C \quad \text{and} \quad \Delta = \mathbb{1}(\{T \leq C\}),$$

where $a \wedge b$ denotes the minimum between two numbers a and b, and $\mathbb{1}(\cdot)$ the indicator function taking the value 1 if the condition in $(\cdot)$ is satisfied and 0 otherwise.

The Cox Proportional Hazards (PH) model [Cox, 1972] is by far the most widely used in survival analysis. It is a regression model that describes the relation between intensity of events and features, given by

$$\lambda(t|X = x) = \lambda_0(t)e^{x^\top \beta^{\text{cox}}},$$

where λ_0 is a baseline intensity function describing how the event risk changes over time at baseline levels of features, and $\beta^{\text{cox}} \in \mathbb{R}^p$ is a vector quantifying the multiplicative impact on the hazard ratio of each feature.

High-dimensional survival analysis. High-dimension settings are becoming increasingly frequent, in particular for genetic data applications where the cut-points estimation is a common problem (see for instance Harvey et al. [1999], Shirota et al. [2001], Cheang et al. [2009]), but also in other contexts where the number of available features to consider as potential risk factors is tremendous, especially with the development of electronic health records. A penalized version of the Cox PH model well suited for such settings is proposed in Simon et al. [2011], but it cannot model nonlinearities. Other methods have being developed to deal with this problem in such settings, like boosting Cox PH models [Li and Luan, 2005] or random survival forests [Ishwaran et al., 2008]. But none of them identify cut-points values, which is of major interest for both interpretations and clinical benefits.

The proposed method. In this chapter, we propose a method called *binacox* for estimating multiple cut-points in a Cox PH model with high-dimensional features. First, the binacox one-hot encodes the continuous input features [Wu and Coggeshall, 2012] through a mapping to a new binarized space of much higher dimension, and then trains the Cox PH model in this space, regularized with the binarsity penalty [Alaya et al., 2017] that combines total-variation regularization with an extra sum-to-zero constraint to avoid collinearity between the one-hot encodings and enable feature selection. Cut-points of the initial continuous input features are then

detected by the jumps in the regression coefficient vector, that the binarsity penalty enforces to be piecewise constant.

Organization of the chapter. The main contribution of this chapter is then the idea of using a total-variation penalization, with an extra linear constraint, on the weights of a Cox PH model trained on a binarization of the raw continuous features, leading to a procedure that selects multiple cut-points per feature, looking at all features simultaneously and that also selects relevant features. A precise description of the model is given in Section 6.2. Section 6.3 highlights the good theoretical properties of the binacox by establishing a fast oracle inequality in prediction. Section 6.4 presents the simulation procedure used to evaluate the performances and compares it with existing methods. In Section 6.5, we apply our method to high-dimensional genetic datasets. Finally, we discuss the obtained results in Section 6.6.

Notations. Throughout the chapter, for every $q > 0$, we denote by $\|v\|_q$ the usual ℓ_q-quasi norm of a vector $v \in \mathbb{R}^m$, namely

$$\|v\|_q = (\sum_{k=1}^{m} |v_k|^q)^{1/q}$$

and

$$\|v\|_\infty = \max_{k=1,\dots,m} |v_k|.$$

We write $\mathbf{1}$ (resp. $\mathbf{0}$) the vector having all coordinates equal to one (resp. zero). We also denote $|A|$ the cardinality of a finite set A. If I is an interval, $|I|$ stands for its Lebesgue measure. Finally, for any $u \in \mathbb{R}^m$ and any $L \subset \{1, \dots, m\}$, we denote u_L as the vector in $\mathbb{R}^m$ satisfying $(u_L)_k = u_k$ for $k \in L$ and $(u_L)_k = 0$ for $k \in L^\complement = \{1, \dots, m\} \backslash L$. Let M be a matrix of size $k \times k'$, $M_{j,\bullet}$ denotes its j-th row and $M_{\bullet,l}$ its l-th column.

6.2 Method

Cox PH model with cut-points. Consider a training dataset of n independent and identically distributed (i.i.d.) examples

$$(x_1, z_1, \delta_1), \dots, (x_n, z_n, \delta_n) \in [0,1]^p \times \mathbb{R}_+ \times \{0,1\},$$

where the condition $x_i \in [0,1]^p$ for all $i \in \{1, \dots, n\}$ is always true after an appropriate rescaling preprocessing step, with no loss of generality. Let us denote

$$\boldsymbol{X} = [x_{i,j}]_{1 \le i \le n; 1 \le j \le p}$$

the $n \times p$ design matrix vertically stacking the n samples of p raw features. Let $\boldsymbol{X}_{\bullet,j}$ be the j-th feature column of $\boldsymbol{X}$ and $\boldsymbol{X}_{i,\bullet}$ the i-th row example. In order to simplify presentation of our results, we assume in the chapter that all raw features $\boldsymbol{X}_{\bullet,j}$ are continuous. Assume that intensity of events for patient i is given by

$$\lambda^\star(t|\boldsymbol{X}_{i,\bullet} = x_i) = \lambda_0^\star(t)e^{f^\star(x_i)}, \tag{6.1}$$

where $\lambda_0^\star(t)$ is the baseline hazard function, and

$$f^\star(x_i) = \sum_{j=1}^{p} \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \mathbb{1}(x_{i,j} \in I_{j,k}^\star), \tag{6.2}$$

with $I_{j,k}^\star = (\mu_{j,k-1}^\star, \mu_{j,k}^\star]$ for $k \in \{1, \ldots, K_j^\star + 1\}$. Since our model defined in (6.1) is not identifiable, we choose to impose a sum-to-zero constraint in each $\beta^\star$'s block, that is

$$\sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star |I_{j,k}^\star| = 0 \text{ for all } j \in \{1, \ldots, p\},$$

then re-defining the baseline in (6.1) as

$$\lambda_0^\star(t) := \lambda_0^\star(t) \exp(\sum_{j=1}^{p} \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star).$$

Here, $\mu_{j,k}^\star$ for $k \in \{1, \cdots, K_j^\star\}$ denote the so-called cut-points of feature $j \in \{1, \ldots, p\}$ that are such that

$$\mu_{j,1}^\star < \mu_{j,2}^\star < \cdots < \mu_{j,K_j^\star}^\star,$$

with the conventions $\mu_{j,0}^\star = 0$ and $\mu_{j,K_j^\star+1}^\star = 1$. Denoting $K^\star = \sum_{j=1}^{p} K_j^\star$, the vector of regression coefficients $\beta^\star \in \mathbb{R}^{K^\star+p}$ is given by

$$\beta^\star = ({\beta_{1,\bullet}^\star}^\top, \ldots, {\beta_{p,\bullet}^\star}^\top)^\top = (\beta_{1,1}^\star, \ldots, \beta_{1,K_1^\star+1}^\star, \ldots, \beta_{p,1}^\star, \ldots, \beta_{p,K_p^\star+1}^\star)^\top,$$

while the cut-points vector $\mu^\star \in \mathbb{R}^{K^\star}$ is given by

$$\mu^\star = ({\mu_{1,\bullet}^\star}^\top, \ldots, {\mu_{p,\bullet}^\star}^\top)^\top = (\mu_{1,1}^\star, \ldots, \mu_{1,K_1^\star}^\star, \ldots, \mu_{p,1}^\star, \ldots, \mu_{p,K_p^\star}^\star)^\top.$$

Our goal is to estimate simultaneously $\mu^\star$ and $\beta^\star$, which also requires an estimation of unknown $K_j^\star$ for $j \in \{1, \ldots, p\}$. To this end, the first step of our proposed methodology is to map the features space to a much higher space of binarized features.

Binarization. The binarized matrix $\boldsymbol{X}^B$ is a sparse matrix with an extended number $p+d$ of columns, typically with $d \gg p$, where features are one-hot encoded [Wu and Coggeshall, 2012, Liu et al., 2002]. The j-th column $\boldsymbol{X}_{\bullet,j}$ is then replaced by $d_j + 1 \geq 2$ columns $\boldsymbol{X}^B_{\bullet,j,1}, \ldots, \boldsymbol{X}^B_{\bullet,j,d_j+1}$ containing only zeros and ones and the i-th row $x_i^B \in \mathbb{R}^{p+d}$ is written

$$x_i^B = (x^B_{i,1,1}, \ldots, x^B_{i,1,d_1+1}, \ldots, x^B_{i,p,1}, \ldots, x^B_{i,p,d_p+1})^\top.$$

To be more precise, we consider a partition of intervals $I_{j,1}, \ldots, I_{j,d_j+1}$ where $I_{j,l} = (\mu_{j,l-1}, \mu_{j,l}]$ for $l \in \{1, \ldots, d_j+1\}$, with $\mu_{j,0} = 0$ and $\mu_{j,d_j+1} = 1$ by convention. Then for $i \in \{1, \ldots, n\}$ and $l \in \{1, \ldots, d_j+1\}$, we define

$$x^B_{i,j,l} = \begin{cases} 1 & \text{if } x_{i,j} \in I_{j,l}, \\ 0 & \text{otherwise.} \end{cases}$$

A natural choice of intervals $I_{j,l}$ is given by a uniform grid $\mu_{j,l} = l/(d_j+1)$.

To each binarized feature $\boldsymbol{X}^B_{\bullet,j,l}$ corresponds a parameter $\beta_{j,l}$ and the vectors associated to the binarization of the j-th feature are naturally denoted

$$\begin{aligned} & \beta_{j,\bullet} = (\beta_{j,1}, \ldots, \beta_{j,d_j+1})^\top \\ \text{and} \quad & \mu_{j,\bullet} = (\mu_{j,1}, \ldots, \mu_{j,d_j})^\top. \end{aligned}$$

Hence, we define

$$f_\beta(x_i) = \beta^\top x_i^B = \sum_{j=1}^{p} f_{\beta_{j,\bullet}}(x_i) \tag{6.3}$$

where

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

for all $j \in \{1, \ldots, p\}$. Thus, f_β is a candidate for the estimation of $f^\star = f_{\beta^\star}$ defined in (6.2).

The full parameters vectors of size $p+d$ and d respectively, where $d = \sum_{j=1}^p d_j$, are simply obtained by concatenation of the vectors $\beta_{j,\bullet}$ and $\mu_{j,\bullet}$, that is

$$\beta = (\beta_{1,\bullet}^\top, \ldots, \beta_{p,\bullet}^\top)^\top = (\beta_{1,1}, \ldots, \beta_{1,d_1+1}, \ldots, \beta_{p,1}, \ldots, \beta_{p,d_p+1})^\top,$$

and

$$\mu = (\mu_{1,\bullet}^\top, \ldots, \mu_{p,\bullet}^\top)^\top = (\mu_{1,1}, \ldots, \mu_{1,d_1}, \ldots, \mu_{p,1}, \ldots, \mu_{p,d_p})^\top.$$

Estimation procedure. In the sequel of the chapter, for a fixed vector μ of quantization, we define the binarized partial negative log-likelihood (rescaled by $1/n$) as follows

$$\ell_n(f_\beta) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\Big\{f_\beta(x_i) - \log\sum_{i':z_{i'}\geq z_i} e^{f_\beta(x_{i'})}\Big\}. \tag{6.4}$$

Our approach consists in minimizing the function ℓ_n plus the binarsity penalization term introduced in the Chapter 5. The resulting optimization problem is

$$\hat{\beta} \in \operatorname{argmin}_{\beta\in\mathscr{B}_{p+d}(R)}\Big\{\ell_n(f_\beta) + \mathrm{bina}(\beta)\Big\} \tag{6.5}$$

where

$$\mathscr{B}_{p+d}(R) = \{\beta\in\mathbb{R}^{p+d} : \|\beta\|_2 \leq R\}$$

is the ℓ_2-ball of radius $R>0$ in $\mathbb{R}^{p+d}$ and

$$\mathrm{bina}(\beta) = \sum_{j=1}^{p}\Big(\sum_{l=2}^{d_j+1}\omega_{j,l}|\beta_{j,l} - \beta_{j,l-1}| + \delta_1(\beta_{j,\bullet})\Big),$$

where

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

and the weights $\omega_{j,l}$ are of order

$$\omega_{j,l} = \mathcal{O}\left(\sqrt{\frac{\log(p+d)}{n}}\right),$$

see Section 6.B.1 for their explicit form.

It turns out that the binarsity penalty is well suited for our problem. First, it tackles the problem that $\boldsymbol{X}^B$ is not full rank by construction, since

$$\sum_{l=1}^{d_j+1} x^B_{i,j,l} = 1$$

for all $j\in\{1,\dots,p\}$, which means that the columns of each block sum to $\mathbf{1}$. This problem is solved since the penalty imposes the linear constraint $\sum_{l=1}^{d_j+1}\beta_{j,l}=0$ in each block with the $\delta_1(\cdot)$ term. Then, the other term in the penalty consists in a within block weighted total-variation penalization

$$\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} = \sum_{l=2}^{d_j+1}\omega_{j,l}|\beta_{j,l} - \beta_{j,l-1}|, \tag{6.6}$$

that takes advantage on the fact that within each block, binary features are ordered. The effect is then to keep the number of different values taken by $\beta_{j,\bullet}$ to a minimal level, which makes significant cut-points appear, as detailed hereafter.

Let us make a first assumption required for being sure to detect all cut-points.

Assumption 3 *We choose* d_j *such that* $\min_{1\leq k\leq K_j^\star+1}|I_{j,k}^\star| \geq \max_{1\leq l\leq d_j+1}|I_{j,l}|$ *for all* $j \in \{1,\dots,p\}$.

This assumption ensures that for all features $j \in \{1,\dots,p\}$, there exists a unique interval $I_{j,l}$ containing cut-point $\mu_{j,k}^\star$, which we denote

$$I_{j,l_{j,k}^\star} = (\mu_{j,l_{j,k}^\star-1}, \mu_{j,l_{j,k}^\star}]$$

for all $k \in \{1,\dots,K_j^\star\}$. Note that in practice, one can always work under Assumption 3 by increasing d_j.

For all $\beta \in \mathbb{R}^{p+d}$, let $\mathcal{A}(\beta) = \left[\mathcal{A}_1(\beta),\dots,\mathcal{A}_p(\beta)\right]$ be the concatenation of the support sets relative to the total-variation penalization, namely

$$\mathcal{A}_j(\beta) = \left\{l : \beta_{j,l} \neq \beta_{j,l-1},\ \text{for } l = 2,\dots,d_j+1\right\}$$

for all $j = 1,\dots,p$. Similarly, we denote $\mathcal{A}^{\complement}(\beta) = \left[\mathcal{A}_1^{\complement}(\beta),\dots,\mathcal{A}_p^{\complement}(\beta)\right]$ the complementary set of $\mathcal{A}(\beta)$. We then denote

$$\mathcal{A}_j(\hat{\beta}) = \{\hat{l}_{j,1},\dots,\hat{l}_{j,s_j}\}, \tag{6.7}$$

where $\hat{l}_{j,1} < \cdots < \hat{l}_{j,s_j}$ and $s_j = |\mathcal{A}_j(\hat{\beta})|$. Finally, we obtain the following estimator

$$\widehat{\mu}_{j,\bullet} = (\mu_{j,\hat{l}_{j,1}},\dots,\mu_{j,\hat{l}_{j,s_j}})^\top \tag{6.8}$$

for $\mu_{j,\bullet}^\star$ and $j = 1,\dots,d$. By construction, $K_j^\star$ is estimated by

$$\widehat{K}_j = s_j,$$

see Section 6.B.1 for its explicit form. Details on the algorithm used to solve the regularization problem (6.5) are given in Section 6.A.1.

6.3 Theoretical guarantees

This paragraph is devoted to our theoretical result. In order to evaluate the prediction error, we construct an (empirical) Kullback-Leibler divergence KL_n between the true function $f^\star$ and any other candidate f as

$$KL_n(f^\star, f) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \log\left\{\frac{e^{f^\star(X_i)}\sum_{i=1}^n Y_i(t)e^{f(X_i)}}{e^{f(X_i)}\sum_{i=1}^n Y_i(t)e^{f^\star(X_i)}}\right\} Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}\mathrm{d}t, \tag{6.9}$$

where we denote $Y_i(t) = \mathbb{1}(Z_i \geq t)$ the at-risk process. This divergence has been introduced in Senoussi [1990]. The oracle inequality in Theorem 6.3.1 is expressed

in terms of compatibility factor [van de Geer and Bühlmann, 2009] satisfied by the following nonnegative symmetric matrix

$$\Sigma_n(f^\star,\tau)=\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big(X_i^B-\bar{X}_n(s)\Big)\Big(X_i^B-\bar{X}_n(s)\Big)^\top y_i(s)e^{f^\star(X_i)}\lambda_0^\star(s)ds, \qquad (6.10)$$

where

$$\bar{X}_n(s)=\frac{\sum_{i=1}^n X_i^B y_i(s)e^{f^\star(X_i)}}{\sum_{i=1}^n y_i(s)e^{f^\star(X_i)}} \quad \text{and} \quad y_i(s)=\mathbb{E}[Y_i(s)|X_i].$$

For any concatenation of index subsets $L=[L_1,\dots,L_p]$, we define the compatibility factor

$$\kappa_\tau(L)=\inf_{\beta\in\mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}}\frac{\sqrt{\beta^\top\Sigma_n(f^\star,\tau)\beta}}{\|\beta_L\|_2},$$

where

$$\mathscr{C}_{\mathrm{TV},\omega}(L)=\Big\{\beta\in\mathscr{B}_{p+d}(R):\sum_{j=1}^p\|(\beta_{j,\bullet})_{L_j^\complement}\|_{\mathrm{TV},\omega_{j,\bullet}}\le 3\sum_{j=1}^p\|(\beta_{j,\bullet})_{L_j}\|_{\mathrm{TV},\omega_{j,\bullet}}\Big\}$$

is a cone composed by all vectors with similar support L.

Assumption 4 *Let $\varepsilon\in(0,1)$, and define*

- $f_\infty^\star=\max_{i=1,\dots,n}|f^\star(X_i)|\le\sum_{j=1}^p\|\beta_{j,\bullet}^\star\|_\infty$,
- $r_\tau=(1/n)\mathbb{E}[\sum_{i=1}^n Y_i(\tau)e^{f^\star(X_i)}]$,
- $\Lambda_0^\star(\tau)=\int_0^\tau\lambda_0^\star(s)ds$,
- $t_{n,p,d,\varepsilon}$ *as the solution of* $(p+d)^2\exp\{-nt^2_{n,p,d,\varepsilon}/(2+2t_{n,p,d,\varepsilon}/3)\}=\varepsilon/2.221$.

For any concatenation set $L=[L_1,\dots,L_p]$ such that $\sum_{j=1}^p|L_j|\le K^\star$, assume that

$$\kappa_\tau^2(L)>\Xi_\tau(L)$$

where

$$\Xi_\tau(L)=4|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\Big\{\Big(1+e^{2f_\infty^\star}\Lambda_0^\star(\tau)\Big)\sqrt{(2/n)\log\Big(2(p+d)^2/\varepsilon\Big)}+\Big(2e^{2f_\infty^\star}\Lambda_0^\star(\tau)/r_\tau\Big)t^2_{n,p,d,\varepsilon}\Big\}.$$

Note that $\kappa_\tau^2(L)$ is the smallest eigenvalue of a population integrated covariance matrix defined in (6.10), so it is reasonable to treat it as a constant. Moreover, $t^2_{n,p,d,\varepsilon}$ is of order

$$\frac{1}{n}\log\left(\frac{(p+d)^2}{\varepsilon}\right).$$

If

$$\frac{|L|\log(p+d)}{n}$$

is sufficiently small, then Assumption 4 is verified. With these preparations, let us now state the oracle inequality satisfied by our estimator of $f^\star$ which is by construction given by $\hat{f} = f_{\hat{\beta}}$ (see (6.3)).

Theorem 6.3.1 *Let* $c_{p,R,f^\star_\infty} = \sqrt{p}R + f^\star_\infty$, $\psi(u) = e^u - u - 1$, $\varrho > 2c^2_{p,R,f^\star_\infty}/\psi(-c_{p,R,f^\star_\infty})$ *and* $\xi = 2/\Big(\varrho\psi(-c_{p,R,f^\star_\infty})/2c^2_{p,R,f^\star_\infty} - 1\Big)$. *The following inequality*

$$KL_n(f^\star, f_{\hat{\beta}}) \leq (1+\xi) \inf_{\substack{\beta\in\mathscr{B}_{p+d}(R)\\ |\mathcal{A}(\beta)|\leq K^\star\\ \forall j, \mathbf{1}^\top\beta_{j,\bullet}=0}} \Bigg\{ KL_n(f^\star, f_\beta) + \frac{512\varrho}{1 - 2c^2_{p,R,f^\star_\infty}/\varrho\psi(-c_{p,R,f^\star_\infty})} \frac{|\mathcal{A}(\beta)| \max_{j=1,\dots,p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|^2_\infty}{\kappa^2_\tau\Big(\mathcal{A}(\beta)\Big) - \Xi_\tau\Big(\mathcal{A}(\beta)\Big)} \Bigg\} \tag{6.11}$$

holds with probability greater than $1 - 28.55e^{-c} - e^{-nr^2_\tau/(8e^{2f^\star_\infty})} - 3\varepsilon$, *for some* $c > 0$.

The proof of the theorem is postponed to Section 6.B.3. The second term in the right-hand side of (6.11) can be viewed as a variance term, and its dominant term satisfies

$$\frac{|\mathcal{A}(\beta)| \max_{j=1,\dots,p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|^2_\infty}{\kappa^2_\tau\Big(\mathcal{A}(\beta)\Big) - \Xi_\tau\Big(\mathcal{A}(\beta)\Big)} \lesssim \frac{|\mathcal{A}(\beta)|}{\kappa^2_\tau\Big(\mathcal{A}(\beta)\Big) - \Xi_\tau\Big(\mathcal{A}(\beta)\Big)} \frac{\log(p+d)}{n} \tag{6.12}$$

where the symbol $\lesssim$ means that the inequality holds up to multiplicative constant. The complexity term in (6.12) depends on both the sparsity of the vector β relatively to the total-variation penalization (through $|\mathcal{A}(\beta)|$) and the compatibility factor. Finally, the rate of convergence of the estimator $\hat{f} = f_{\hat{\beta}}$ has the expected shape

$$\mathcal{O}\left(\frac{\log(p+d)}{n}\right).$$

Remark 6.3.1 *As in generalized linear models, the minimization problem has to be localized to achieve a bound as in Theorem 6.3.1. We chose to localize the candidates* $\beta \in \mathbb{R}^{p+d}$ *in a* ℓ_2*-ball. Other possible choices include the localisation around the "true" values, that is in the set*

$$\Big\{ \max_{i=1,\dots,n} |f_\beta(X_i) - f^\star(X_i)| \leq R \Big\},$$

see e.g. Bühlmann and van de Geer [2011], and in this case the constant $c_{p,R,f_\infty^\star}$ *equals* R*. An other possibility is to consider the localization in the convex set*

$$\Big\{ \max_{i=1,\dots,n} |f_\beta(X_i)| = \max_{i=1,\dots,n} |\langle x_i^B, \beta \rangle| \leq \sum_{j=1}^{p} \|\beta_{j,\cdot}\|_\infty \leq R \Big\},$$

as in Ivanoff et al. [2016]. In the later case, $c_{p,R,f_\infty^\star} = R + f_\infty^\star$*. With these other localizations, the dependence in* p *does not appear in the oracle bound, but the proofs do not change.*

6.4 Performance evaluation

6.4.1 Practical details

Let us give some details about binacox's use in practice. First, instead of taking the uniform grid for the intervals $I_{j,l}$ that makes theoretical results easier to state, we choose the estimated quantiles

$$\mu_{j,l} = q_j\Big(\frac{l}{d_j + 1}\Big)$$

where $q_j(u)$ denotes an empirical quantile of order u for $\boldsymbol{X}_{\bullet,j}$. This choice provides two major practical advantages : 1) the resulting grid is data-driven and follows the distribution of $\boldsymbol{X}_{\bullet,j}$ and 2) there is no need to tune hyper-parameters d_j (number of bins for the one-hot encoding of raw feature j). Indeed, if d_j is "large enough" (we take $d_j = 50$ for all $j \in \{1, \dots, p\}$ in practice), increasing d_j barely changes the results since the cut-points selected by the penalization do not change any more, and the size of each block automatically adapts itself to the data : depending on the distribution of $\boldsymbol{X}_{\bullet,j}$, ties may appear in the corresponding empirical quantiles (for more details on this last point, see Alaya et al. [2017]).

Then, let us precise that the binacox is proposed in the `tick` library [Bacry et al., 2017], we provide sample code for its use in Figure 6.1. For practical convenience, we take all weights $\omega_{j,l} = \gamma$ and select the hyper-parameter γ using a V-fold cross-validation procedure with $V = 10$, taking the negative partial log-likelihood defined in (6.4) as a score computed after a refit of the model on the binary space obtained by the estimated cut-points, and with the sum-to-zero constraint only (without the TV penalty, which actually gives a fair $\beta^\star$ estimate in practice), which intuitively makes sense. Figure 6.A.1 in Section 6.A.2 gives the learning curves obtained with this cross-validation procedure on an example.

We also add a simple de-noising step in the cut-point detection phase which is useful in practice. Indeed, it is usual to observe two consecutive $\hat{\beta}$'s jumps in the neighbourhood of a true cut-point, leading to an over-estimation of $K^\star$. This can

be viewed as a clustering problem. We tried different clustering methods but in practice, nothing works better than this simple routine : if $\hat{\beta}$ has three consecutive different coefficients within a group, then only the largest jump is considered as a "true" jump. Figure 6.A.2 in Section 6.A.2 illustrates this last point.

```python
from tick.simulation import SimuCoxRegWithCutPoints
from tick.preprocessing.features_binarizer import FeaturesBinarizer
from tick.inference import CoxRegression

# Generate data
simu = SimuCoxRegWithCutPoints(n_samples=1000, n_features=20)
X, Y, delta = simu.simulate()

# Binarize features
binarizer = FeaturesBinarizer(n_cuts=50)
X_bin = binarizer.fit_transform(X)

# Fit the model with a penalty strength equal to `C`
learner = CoxRegression(penalty='binarsity',
                        blocks_start=binarizer.blocks_start,
                        blocks_length=binarizer.blocks_length,
                        C=10)
learner.fit(X_bin, Y, delta)

# Obtain the estimated vector
beta = learner.coeffs
```

FIGURE 6.1 Sample python code for the use of binacox in the `tick` library, with the use of the `FeaturesBinarizer` transformer for features binarization.

6.4.2 Simulation

Design. In order to assess the methods, we perform an extensive Monte Carlo simulation study. We first take

$$[x_{ij}] \in \mathbb{R}^{n \times p} \sim \mathcal{N}\Big(0, \Sigma(\rho)\Big),$$

with $\Sigma(\rho)$ a $(p \times p)$ Toeplitz covariance matrix [Mukherjee and Maiti, 1988] with correlation $\rho \in (0, 1)$.

For each feature $j \in \{1, \dots, p\}$, we sample the cut-points $\mu^\star_{jk}$ uniformly without replacement among the estimated quantiles $q_j(u/10)$ for $u \in \{1, \dots, 9\}$ for $k \in \{1, \dots, K^\star_j\}$. This way, we avoid having undetectable cut-points (with very few examples above the cut-point value) as well as two cut-points indissociable because too close. We choose the same $K^\star_j$ values for all $j \in \{1, \dots, p\}$. Now that the true cut-points $\mu^\star$ are generated, one can compute the corresponding binarized version

of the features that we denote $x_i^{B^\star}$ for example i. Then, we generate

$$c_{jk} \sim (-1)^k |\mathcal{N}(1, 0.5)|$$

for $k \in \{1, \ldots, K_j^\star + 1\}$ and $j \in \{1, \ldots, p\}$ to make sure we create "real" cut-points, and take

$$\beta_{jk}^\star = c_{jk} - (K_j^\star + 1)^{-1} \sum_{k=1}^{K_j^\star+1} c_{jk}$$

to impose the sum-to-zero constraint of the true coefficients in each block. We also induce a sparsity aspect by uniformly selecting a proportion r_s of features $j \in \mathcal{S}$ with no cut-point effect, that is features for which we enforce $\beta_{jk}^\star = 0$ for all $k \in \{1, \ldots, K_j^\star + 1\}$. Finally, we generate survival times using Weibull distributions, which is a common choice in survival analysis [Klein and Moeschberger, 2005], that is

$$T_i \sim \nu^{-1} \Big[-\log(U_i) \exp\Big(-(x_i^{B^\star})^\top \beta_i^\star \Big) \Big]^{1/\varsigma}$$

with $\nu > 0$ and $\varsigma > 0$ the scale and shape parameters respectively, $U_i \sim \mathcal{U}([0, 1])$ and where $\mathcal{U}([a, b])$ stands for the uniform distribution on a segment $[a, b]$. The distribution of the censoring variable C_i is geometric $\mathcal{G}(\alpha_c)$, where $\alpha_c \in (0, 1)$ is empirically tuned to maintain a desired censoring rate $r_c \in [0, 1]$.

The choices of the hyper-parameters is driven by the applications on real data presented in Section 6.5 and are summarized in Table 6.1. Figure 6.2 gives an example of data generated according to the design we just described, with $\beta^\star$ plotted in Figure 6.3.

TABLE 6.1 Hyper-parameters choice for simulation.

n	p	ρ	$K_j^\star$	ν	ς	r_c	r_s
(200, 4000)	50	0.5	$\{1, 2, 3\}$	2	0.1	0.3	0.2

Metrics. We evaluate the considered methods using two metrics. The first one assesses the estimation of the cut-points values by

$$m_1 = |\mathcal{S}'|^{-1} \sum_{j \in \mathcal{S}'} \mathcal{H}(\mathcal{M}_j^\star, \widehat{\mathcal{M}}_j)$$

where $\mathcal{M}_j^\star = \{\mu_{j,1}^\star, \ldots, \mu_{j,K_j^\star}^\star\}$ (resp. $\widehat{\mathcal{M}}_j = \{\hat{\mu}_{j,1}, \ldots, \hat{\mu}_{j,\widehat{K}_j}\}$) is the set of true (resp. estimated) cut-points for feature j,

$$\mathcal{S}' = \Big\{ j,\ j \notin \mathcal{S} \cap \{l,\ \widehat{\mathcal{M}}_l = \emptyset\} \Big\}$$

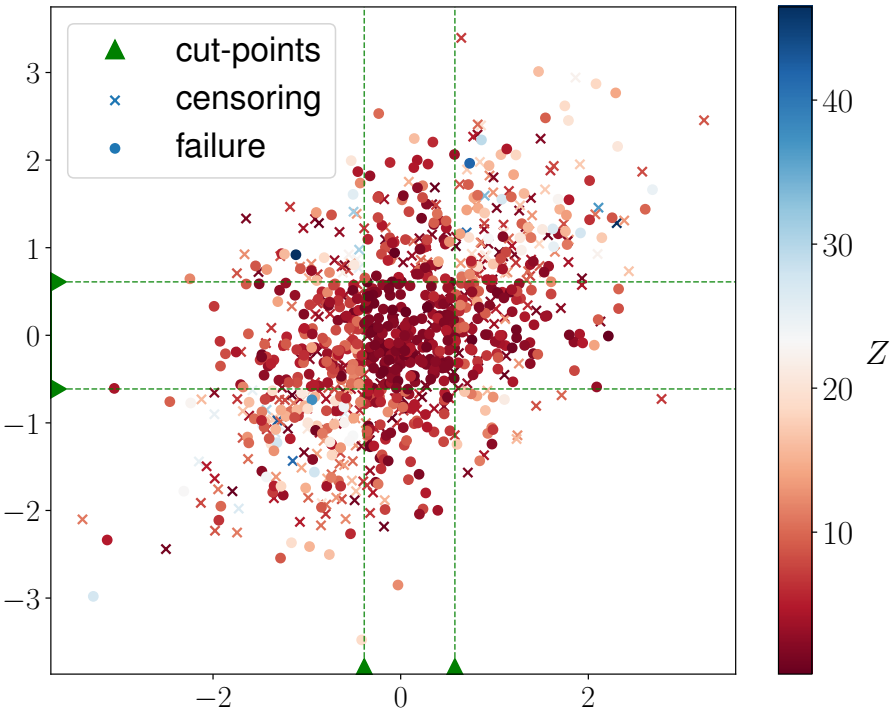


FIGURE 6.2 Illustration of data simulated with $p = 2$, $K_1^\star = K_2^\star = 2$ and $n = 1000$. Dots represent failure times ($z_i = t_i$) while crosses represent censoring times ($z_i = c_i$), and the colour gradient represents the z_i values (red for low and blue for high values). The $\beta^\star$ used to generate the data is plotted in Figure 6.3.

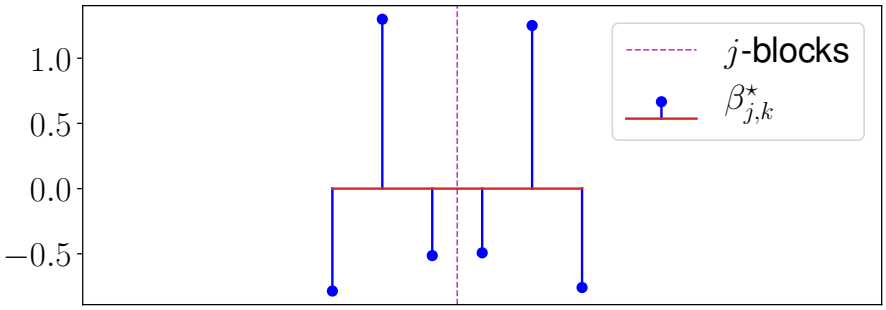


FIGURE 6.3 Illustration of the $\beta^\star$ used in Figure 6.2, with a doted line to demarcate the two blocks (since $p = 2$).

is the indexes corresponding to features with at least one true cut-point and one detected cut-point, and $\mathcal{H}(A, B)$ is the Hausdorff distance between the two sets A and B, that is

$$\mathcal{H}(A, B) = \max\Big(\mathcal{E}(A||B), \mathcal{E}(B||A)\Big)$$

with

$$\mathcal{E}(A||B) = \sup_{b\in B} \inf_{a\in A} |a-b|$$

for two sets A and B. This is inspired by Harchaoui and Lévy-Leduc [2010], except that in our case, both $\mathcal{M}_j^\star$ and $\widehat{\mathcal{M}}_j$ can be empty, which explain the use of $\mathcal{S}'$. The second metric we use is precisely focused on the sparsity aspect : it assesses the ability for each method to detect features with no cut-points and is defined by

$$m_2 = |\mathcal{S}|^{-1} \sum_{j\in\mathcal{S}} \widehat{K}_j.$$

6.4.3 Competing methods

To the best of our knowledge, existing algorithms and methods are based on multiple log-rank tests in univariate models. These methods are widely used and among recent implementations are the web applications `Cutoff Finder` and `Findcutoffs` described respectively in Budczies et al. [2012] and Chang et al. [2017].

We describe in what follows the principle of the univariate log-rank tests. Consider one of the initial variable $\boldsymbol{X}_{\bullet,j} = (x_{1,j}, \ldots, x_{n,j}\}$, and define its 10th and 90th quantiles as $x_{10th,j}$ and $x_{90th,j}$. Define then a grid $\{g_{j,1}, \ldots, g_{j,\kappa_j}\}$. In most implementations, the $g_{j,k}$'s are chosen at the original observation points and such that $x_{10th,j} \leq g_{j,k} \leq x_{90th,j}$. For each $g_{j,k}$, the p-value $\mathrm{pv}_{j,k}$ of the log-rank test associated to the univariate Cox model

$$\lambda_0(t) \exp\left(\beta_j \mathbb{1}(x \leq g_{j,k})\right)$$

is computed (via the `python` package `lifelines` in our implementation). For each initial variable $\boldsymbol{X}_{\bullet,j}$, κ_j p-values are available at this stage. The choice of the size κ_j of the grid depends on the implementation and ranges for several dozens to all observed values between $x_{10th,j}$ and $x_{90th,j}$.

In Figure 6.4, the values $-\log(\mathrm{pv}_{j,k})$ for $k = 1, \ldots, \kappa_j$ (denoted by "MT" for multiple testing) are represented, for the simulated example described in Figure 6.2. Notice that the level $-\log(\alpha) = -\log(0.055)$ is exceed at numerous $g_{j,k}$'s. A common approach is to consider the maximal value $-\log(\mathrm{pv}_{j,\hat{k}})$ and then define the cut-point for variable j as $g_{j,\hat{k}}$. As argued in Altman et al. [1994], this is obviously "associated with an inflation of type I error", for this reason we do not consider this approach.

To cope with the multiple testing (MT) problem at hand, a multiple test correction has to be applied. We consider two corrections. This first is the well-known Bonferroni p-values correction, referred to as MT-B. We insist on the fact that, although commonly used, this method is not correct in this situation as the p-values are correlated. Note also that in this context, the Benjamini–Hochberg (BH) procedure would result in the same detection as MT-B (with FDR=α), since we only

consider as a cut-point candidate the points with minimal p-value. Indeed, applying the classical BH procedure would select far too many cut-points. The second, denoted MT-LS, is the correction proposed in Lausen and Schumacher [1992], based on asymptotic theoretical considerations. Figure 6.4 illustrates how these methods behave on a simulated example. A third correction we could think of would be a bootstrap based MaxT procedure (or MinP) developed in Dudoit and Van Der Laan [2007] or Westfall et al. [1993], but this would be intractable in our high-dimensional setting (see Figure 6.5 that only considers a single feature, and a bootstrap procedure based on MT would dramatically increase the required computing time).

6.4.4 Results of simulation

Example. Figure 6.4 illustrates how the considered methods behave on the data illustrated in Figure 6.2. Through this example, one can visualise the good performances of the binacox method : the position, strength and number of cut-points are well estimated. The MT-B and MT-LS methods can only detect one cut-point by construction. Both methods detect "the most significant" cut-point for the 2 features, namely the one corresponding to the higher jump in $\beta^\star_{j,\bullet}$ (see Figure 6.2) : $\mu^\star_{1,1}$ and $\mu^\star_{2,2}$.

With regard to the shape of the p-value curves, one can see that for each of the two features, the two "main" local maxima correspond to the true cut-points. One could consider a method for detecting those preponderant maxima, but it is beyond the scope of the article (plus it would still be based on the MT methods, which has high computational cost, as detailed hereafter).

Computing times. Let us focus on the computing times required for the considered methods. The multiple testing related methods being univariate, one can directly parallelize their computation on the dimension p (which is what we did), and we consider here a single feature X ($p = 1$). Following the method explained in Section 6.4.3, we have to compute all log-rank test p-values computed on the two populations $\{y_i : x_i > \mu\}$ and $\{y_i : x_i \leq \mu\}$ for $i \in \{1, \dots, n\}$, for μ taking all x_i values between the 10-th and 90-th empirical quantile of X. We denote "MT all" this method in Figure 6.5 that compares its computing times with the binacox one for varying n, and where we add the "MT grid" method that only computes the p-values for candidates $\mu_{j,l}$ used in the binacox method.

Since the number of candidates does not change with n for the MT grid method, the computing time ratio between MT all and MT grid naturally increases, and goes roughly from one to two orders of magnitude higher when n goes from 300 to 4000. Hence to make computations much faster, we consider the MT grid for all multiple testing related method in the sequel of the chapter without mentioning it. The resulting loss of precision in the MT related methods is negligible for a high

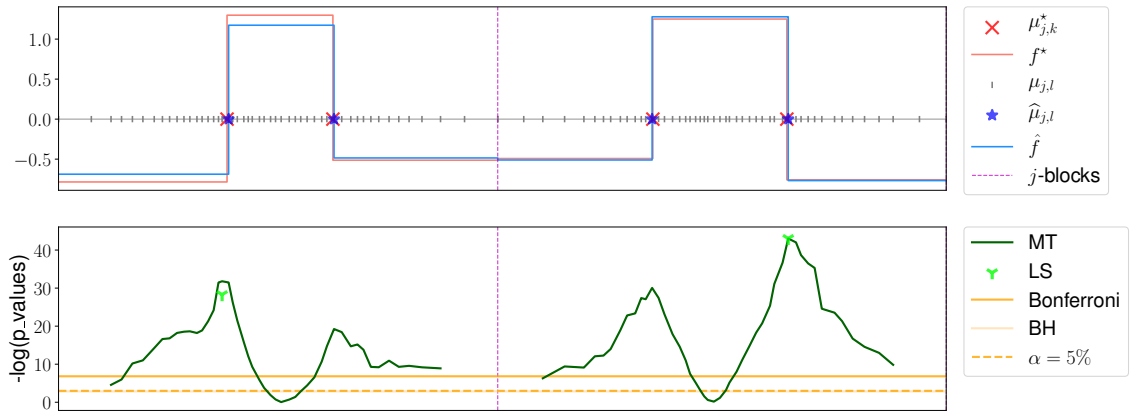


FIGURE 6.4 Illustration of the main quantities involved in the binacox on top, with estimation obtained on the data presented in Figure 6.2. Our algorithm detects the correct number of cut-points $\widehat{K}_j = 2$, and estimates their position very accurately, as well as their strength. At the bottom, one observe the results on the same data using the multiple testing related methods presented in Section 6.4.3. Here the BH threshold lines overlap the one corresponding to $\alpha = 5\%$. The BH procedure would consider as cut-point all $\mu_{j,l}$ value for which the corresponding darkgreen (MT) line value is above, then detecting far too many cut-points.

enough d_j value (we take 50 in practice).

Then, let us stress the fact that the binacox is still roughly 5 times faster than the MT grid method, and it remains very fast when we increase the dimension, as shown in Figure 6.6. It turns out that the computational time grows roughly logarithmically with p.

Performances comparison. Let us compare now the results of simulations in terms of m_1 and m_2 metrics introduced in Section 6.4.2. Figure 6.7 gives a comparison of the considered methods on the cut-points estimation aspect, hence in terms of m_1 score. It appears that the binacox outperforms the multiple testing related methods when $K^\star_j > 1$, and is competitive when $K^\star_j = 1$ except for small values of n. This is due to an overestimation by the binacox in the number of cut-points (see see Figure 6.8) when p is high for small n, which gives higher m_1 values, even if the "true" cut-point is actually well estimated. Note that for such p value, the binacox is way faster than the multiple testing related methods.

Figure 6.8, on the other hand, assesses the ability for each method to detect features with no cut-points using the m_2 metric, that is to estimate $\hat{K}^\star_j = 0$ for $j \in \mathcal{S}$. The binacox appears to have a strong ability to detect features with no cut-point when n takes a high enough value compared to p, which is not the case for the multiple testing related methods.

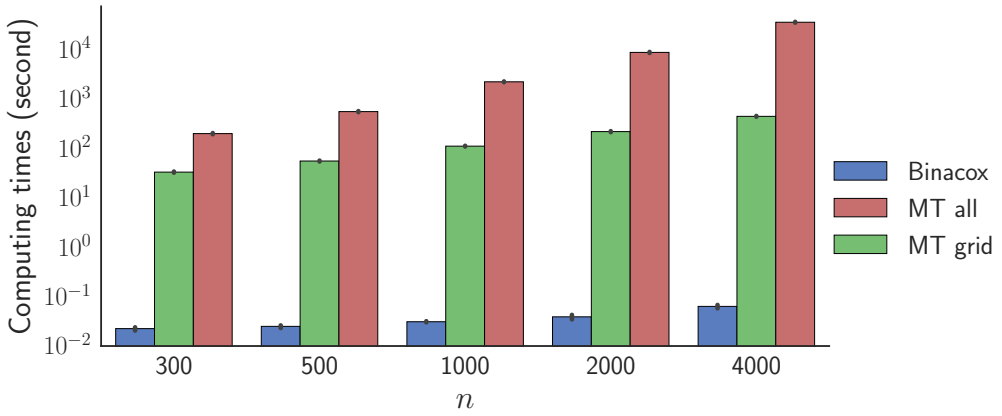


FIGURE 6.5 Average computing times in second (with the black lines representing ± the standard deviation) obtained on 100 simulated datasets (according to Section 6.4.2 with $p = 1$ and $K^\star = 2$) for training the binacox VS the multiple testing method where cut-points candidates are either all x_i values between the 10-th and 90-th empirical quantile of X (MT all), or the same candidates as the grid considered by the binacox (MT grid).

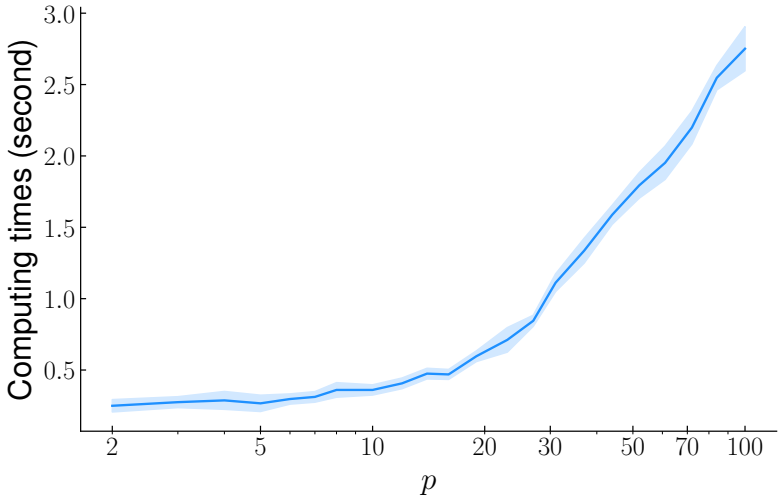


FIGURE 6.6 Average (bold) computing times in second and standard deviation (bands) obtained on 100 simulated datasets (according to Section 6.4.2 with $K_j^\star = 2$) for training the binacox when increasing the dimension p up to 100. Our method remains very fast in a high-dimensional setting.

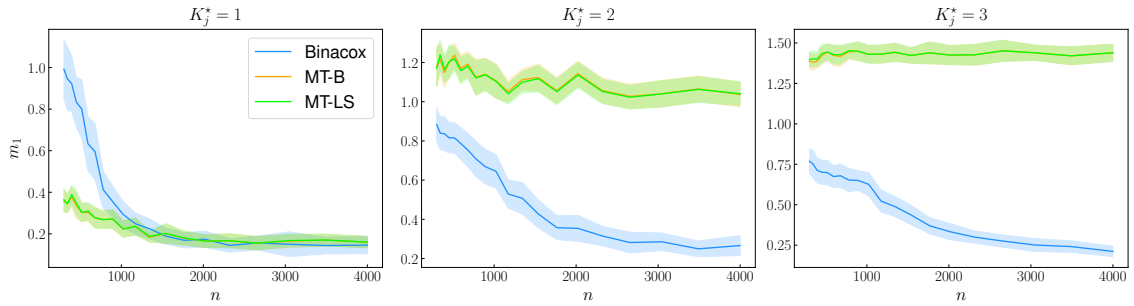


FIGURE 6.7 Average (bold) m_1 scores and standard deviation (bands) obtained on 100 simulated datasets according to Section 6.4.2 with $p = 50$ and $K_j^\star$ equals to 1, 2 and 3 (for all $j \in \{1, \ldots, p\}$) for the left, center and right sub-figures respectively) for varying n. The lower m_1 the best result : the binacox outperforms clearly other methods when there are more than one cut-point, and is competitive with other methods when there is only one cut-points with poorer performances when n is small because of an overestimation of $K_j^\star$ in this case.

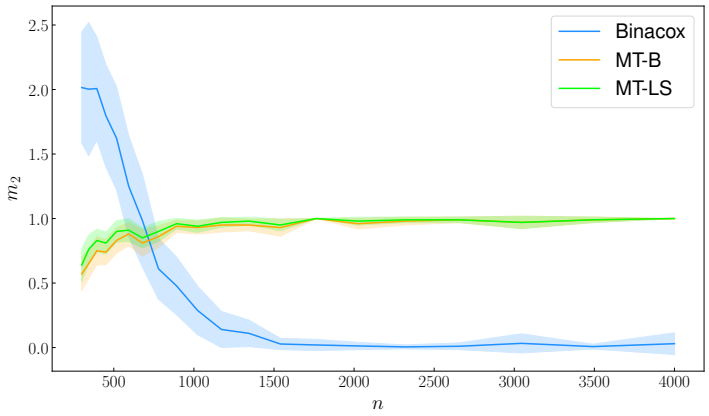


FIGURE 6.8 Average (bold) m_2 scores and standard deviation (bands) obtained on 100 simulated datasets according to Section 6.4.2 with $p = 50$ for varying n. It turns out that MT-B and MT-LS tend to detect a cut-point while there is not (no matter the value of n), and that the binacox overestimates the number of cut-points for small n values but detects well $\mathcal{S}$ for $p = 50$ on the simulated data when $n > 1000$.

6.5 Application to genetic data

In this section, we apply our method on three biomedical datasets. We extracted normalized expression data and survival times Z in days from breast invasive carcinoma (BRCA, $n = 1211$), glioblastoma multiforme (GBM, $n = 168$) and kidney renal clear cell carcinoma (KIRC, $n = 605$). These datasets are available on The Cancer Genome Atlas (TCGA) platform, which aims at accelerating the understanding of the molecular basis of cancer through the application of genomic technologies, including large-scale genome sequencing. A more precise description of these datasets is given in Section A.2.2. For each patients, 20531 features corresponding to the normalized gene expressions are available.

As we saw in Section 6.4.4, the multiple testing related methods are intractable in such high dimension. We therefore make a screening step to select the portion of features the most relevant for our problem among the 20531 ones. To do so, we fit our method on each block j separately and we compute the resulting $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ as a score that roughly assess the propensity for feature j to get one (or more) relevant cut-point. We then select the features corresponding to the top$-P$ values with $P = 50$, this choice being suggested by the distribution of the obtained scores given in Figure 6.A.3 of Section 6.A.3.

Estimation results. Let us present in Figure 6.1 the results obtained by the considered methods on the GBM cancer dataset for the top-10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values. One can observe that all cut-points detected by the univariate multiple testing methods with Bonferroni (MT-B) or Lausen and Schumacher (MT-LS) correction are also detected by the multivariate binacox that detects more cut-points, which is summarized in Table 6.1. It turns out that among the 20531 initial genes, the resulting top-10 are very relevant for a study on GBM cancer (being the most aggressive cancer that begins within the brain). For instance, the first gene SOD3 is related to an antioxidant enzyme that may protect in particular the brain from oxidative stress, which is believed to play a key role in tumour formation [Rajaraman et al., 2008]. Some other genes (like C11orf63 or HOXA1) are known to be directly related to the brain development [Canu et al., 2009].

Similar results are obtained on the KIRC and BRCA cancers and are postponed in Section 6.A.4.

Risk prediction. Let us now investigate how performances in terms of risk prediction are impacted when account is taken of the detected cut-points, namely let us compare predictions when training a Cox PH model either on the original continuous feature space versus on the $\widehat{\mu}$-binarized space constructed with the cut-points estimates.

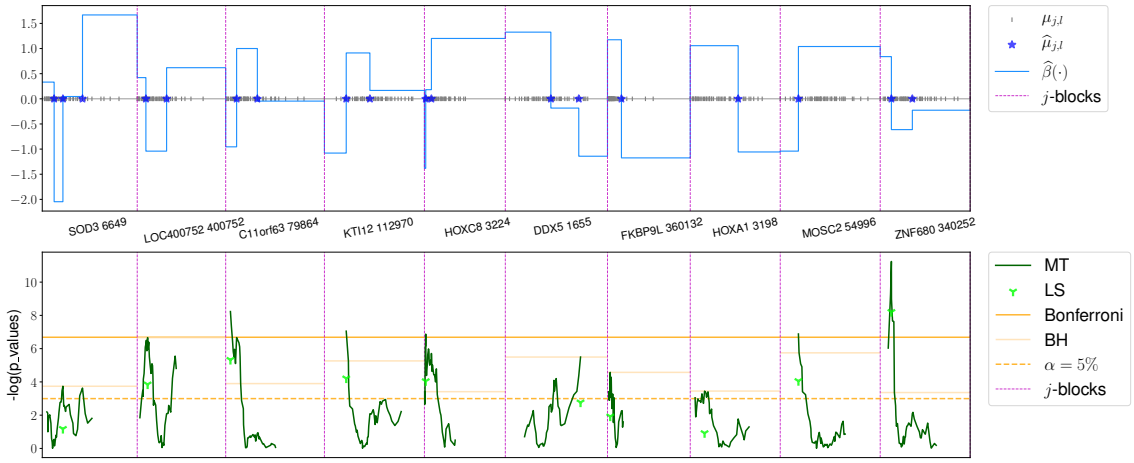


FIGURE 6.1 Illustration of the results obtained on the top−10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the GBM dataset. The binacox detects multiple cut-points and sheds light on non-linear effects for various genes. The BH thresholds are plotted for informational purposes, but are unusable in practice.

TABLE 6.1 Estimated cut-points values for each method on the top−10 genes presented in Figure 6.1 for the GBM cancer. Dots (·) mean "no cut-point detected". The binacox identifies much more cut-points than the univariate MT-B and MT-LS methods. But all cut-points detected by those two methods are also detected by the binacox.

Genes	Binacox	MT-B	MT-LS
SOD3 6649	200.87, 326.40, 606.48	·	·
LOC 400752	31.46, 62.50	·	34.04
C11orf63 79864	40.30, 109.67	19.65	19.65
KTI12 112970	219.60, 305.70	219.60	219.60
HOXC8 3224	3.30, 15.75	3.30	3.30
DDX5 1655	10630.11, 13094.89	·	·
FKBP9L 360132	111.72	·	·
HOXA1 3198	67.28	·	·
MOSC2 54996	107.53	107.53	107.53
ZNF680 340252	385.85, 638.06	385.85	385.85

In a classical Cox PH model, $R_i = \exp(X_i^\top \hat{\beta})$ is known as the predicted risk for patient i measured at $t = 0$. A common metric to evaluate risk prediction performances in a survival setting is the C-index [Heagerty and Zheng, 2005]. It is defined

by

$$\mathcal{C}_\tau = \mathbb{P}[R_i > R_j | Z_i < Z_j, Z_i < \tau],$$

with $i \neq j$ two independent patients and τ the follow-up period. A Kaplan-Meier estimator for the censoring distribution leads to a non-parametric and consistent estimator of $\mathcal{C}_\tau$ [Uno et al., 2011], already implemented in the `python` package `lifelines`. We randomly split the three datasets into a training and a testing sets (30% for testing) and compare the C-index on the test sets in Table 6.2 when the $\widehat{\mu}$-binarized space is constructed based on $\widehat{\mu}$ obtained either from the binacox, MT-B or MT-LS. We also compare performances obtained by two nonlinear multivariate methods known to perform well in high-dimension : the boosting Cox PH (CoxBoost) [Li and Luan, 2005] used with 500 boosting steps, and the random survival forests (RSF) [Ishwaran et al., 2008] used with 500 trees, respectively implemented in the `R` packages `CoxBoost` and `randomForestSRC`. The binacox method clearly improves risk prediction compare to classical Cox PH, as well as MT-B and MT-LS methods. Moreover, it also significantly outperforms both CoxBoost and RSF methods. Figure 6.2 compares the computing times of the considered methods. It appears that the binacox is by far the most computationally efficient.

TABLE 6.2 C-index comparison for Cox PH model trained on continuous features vs. on its binarized version constructed using the considered methods cut-points estimates, and the CoxBoost and RSF methods. On the three datasets, the binacox method gives by far the best results (in bold).

Cancer	Continuous	Binacox	MT-B	MT-LS	CoxBoost	RSF
GBM	0.660	**0.806**	0.753	0.768	0.684	0.691
KIRC	0.682	**0.727**	0.663	0.663	0.679	0.686
BRCA	0.713	**0.849**	0.741	0.738	0.723	0.746

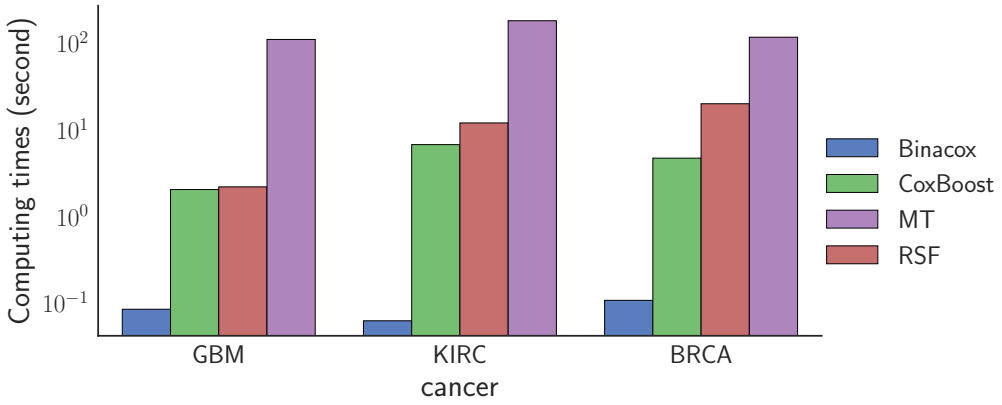


FIGURE 6.2 Comparison of the computing times required by the considered method on the three datasets. The binacox method is orders of magnitude faster.

6.6 Concluding remarks

In this chapter, we introduced the binacox designed for estimating multiple cut-points in a Cox PH model with high-dimensional features. We illustrated the good theoretical properties of the model by establishing non-asymptotic oracle inequality. An extensive Monte Carlo simulation study has been carried out to evaluate the performance of the developed estimation procedure. It showed that our approach outperforms existing methods with a computing time orders of magnitude faster. Moreover, it succeeds in detecting automatically multiple cut-points per feature. The proposed methodology has then been applied on three high-dimensional genetic public datasets. Many gene expressions pinpointed by the model are relevant for medical interpretations (*e.g.* the gene SOD3 for the GBM cancer), whilst others must involve further investigations in the genetic research community. Furthermore, the binacox outperformed the classical Cox PH model in terms of risk prediction performances evaluated through the C-index metric. It can then be an interesting alternative to more classical methods found in the medical literature to deal with prognosis studies in a high dimensional framework, providing a new way to model non-linear features associations. More importantly, our method provides powerful interpretation aspects that could be useful in both clinical research and daily practice. Indeed, in addition to its raw feature selection ability, the estimated cut-points could directly be used in clinical routine. For instance, the binacox directly estimates the impact on the survival risk for a feature (gene expression in our application) to be in a relevant interval through the estimated coefficient corresponding to this interval. Our study lays the groundwork for the development of powerful methods which could help provide personalized care.

Acknowledgments

Mokhtar Z. Alaya is grateful for a grant from DIM Math Innov Région Ile-de-France http://www.dim-mathinnov.fr. Agathe Guilloux' work has been supported by the INCA-DGOS grant PTR-K 2014. The results shown in this chapter are based upon data generated by the TCGA Research Network and freely available from http://cancergenome.nih.gov. *Conflict of Interest* : None declared.

Software

All the methodology discussed in this chapter is implemented in `Python/C++`. The code that generates all figures is available from https://github.com/SimonBussy/binacox in the form of annotated programs, together with notebook tutorials.

Appendices

6.A Additional details

6.A.1 Algorithm.

To solve the regularization problem (6.5), we are first interested in the proximal operator of binarsity [Alaya et al., 2017]. It turns out that it can be computed very efficiently, using an algorithm [Condat, 2013] that we modify in order to include weights $\omega_{j,k}$. It applies in each group the proximal operator of the total-variation since binarsity penalty is block separable, followed by a centering within each block to satisfy the sum-to-zero constraint, see Algorithm 2 in Section 5.2, and Algorithm 3 in Section 5.B for the weighted total-variation proximal operator.

6.A.2 Implementation

Figure 6.A.1 gives the learning curves obtained during the V-fold cross-validation procedure detailed in Section 6.4.3 with $V = 10$ for fine-tuning parameter γ, being the strength of the binarsity penalty. We randomly split the data into a training and a validation set (30% for validation, cross-validation being done on the training). Recall that the score we use is the negative partial log-likelihood defined in (6.4) computed after a refit of the model on the binary space obtained by the estimated cut-points, with the sum-to-zero constraint in each block but without the TV penalty.

Figure 6.A.2 illustrates the denoising step when detecting the cut-points when looking at the $\hat{\beta}$ support relatively to the TV norm. The $\hat{\beta}$ vector plotted here corresponds to the data generated in Figure 6.2 of Section 6.4.2 and where final estimation results are presented in Figure 6.4 of Section 6.4.4. Since it is usual to observe three consecutive $\hat{\beta}$'s jumps in the neighbourhood of a true cut-point, which is the case in Figure 6.A.2 for the first and the last jumps, this could lead to an over-estimation of $K^\star$. To bypass this problem, we then use the following rule : if $\hat{\beta}$ has three consecutive different coefficients within a group, then only the largest jump is considered as a "true" jump.

6.A.3 TCGA genes screening

Figure 6.A.3 illustrates the screening procedure followed to reduce the high-dimension of the TCGA datasets to make the multiple testing related methods tractable. We then fit an univariate binacox on each block j separately and compute the resulting $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ to assess the propensity for feature j to get one (or more)

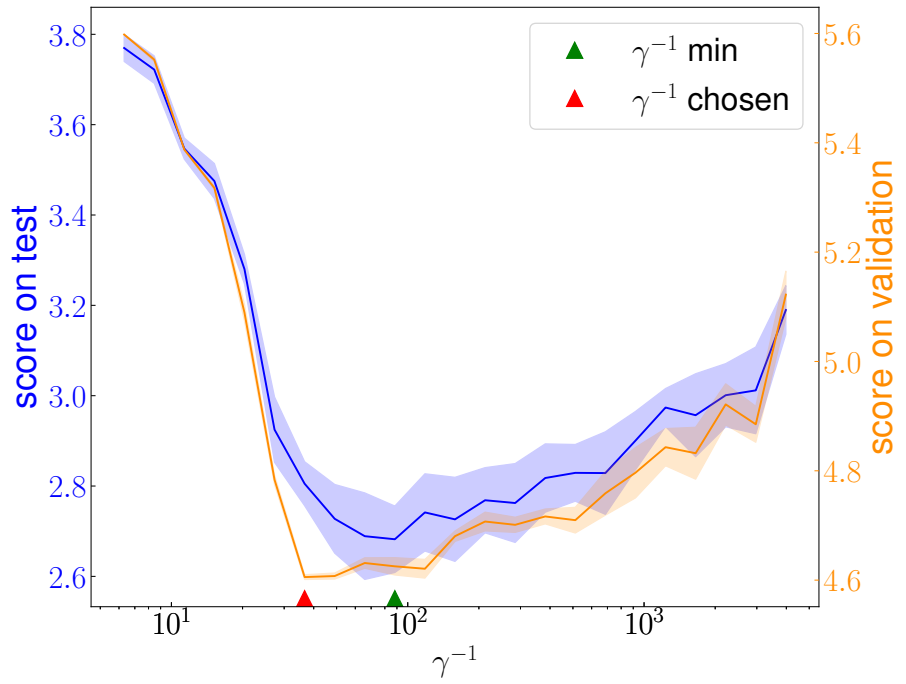


FIGURE 6.A.1 Learning curves obtained for various γ, in blue on the changing test sets of the cross-validation, and in orange on the validation set. Bold lines represent average scores on the folds and bands represent Gaussian 95% confidence intervals. The green triangle points out the value of γ^{-1} that gives the minimum score (best training score), while the γ^{-1} value we automatically select (the red triangle) is the smallest value such that the score is within one standard error of the minimum, wich is a classical trick [Simon et al., 2011] that favors a slightly higher penalty strength (smaller γ^{-1}), to avoid an over-estimation of $K^\star$ in our case.

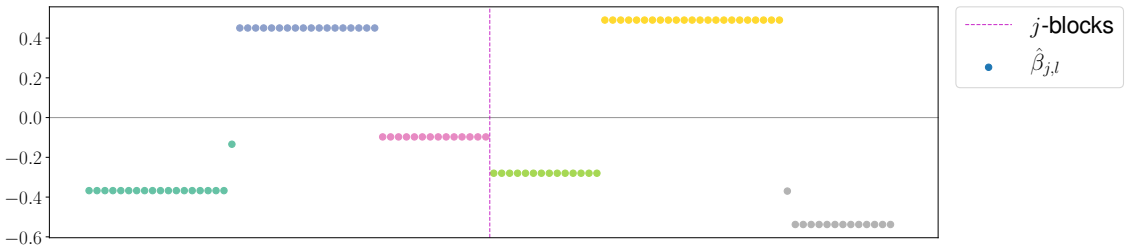


FIGURE 6.A.2 Illustration of the denoising step on the cut-points detection phase. Within a block (separated with the dotted pink line), the different colors represent $\hat{\beta}_{j,l}$ with corresponding $\mu_{j,l}$ in distinct estimated $I^\star_{j,k}$. When a $\hat{\beta}_{j,l}$ is "isolated", it is assigned to its "closest" group.

relevant cut-point. It appears that taking the top$-P$ features with $P = 50$ is a reasonable choice for each considered dataset.

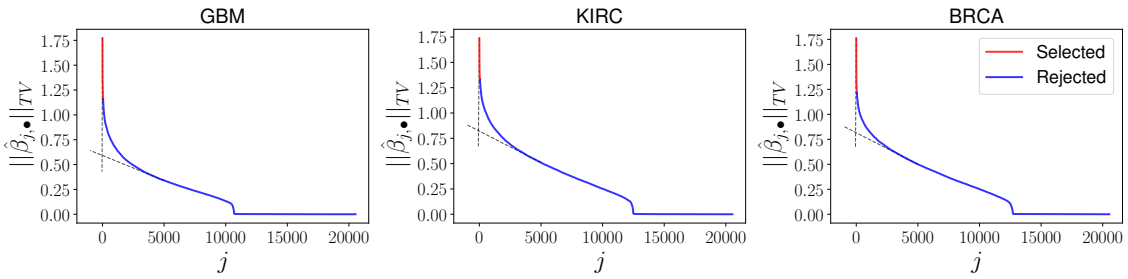


FIGURE 6.A.3 $\|\hat{\beta}_{j,\bullet}\|_{\text{TV}}$ obtained on univariate binacox fits for the three considered datasets. Top$-P$ selected features appear in red, and it turns out that taking $P = 50$ coincides with the elbow (represented with the dotted grey lines) in each three curves.

6.A.4 Results on BRCA and KIRC data

Figure 6.A.4 presents the results obtained by the considered methods on the BRCA cancer dataset for the top-10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\text{TV}}$ values. Table 6.A.1 summarizes the detected cut-points values for each method. It turns out that the selected genes are very relevant for cancer studies (for instance, NPRL2 is a tumor suppressor gene [Huang et al., 2016]), and more particularly for breast cancer studies : for instance, HBS1L expression is known for being predictive of breast cancer survival [Antonov et al., 2014, Antonov, 2011, BioProfiling, 2009], while FOXA1 and PPFIA1 are highly related to breast cancer, see Badve et al. [2007] and Dancau et al. [2010] respectively.

Finally, Figure 6.A.5 gives the results obtained by the considered methods on the KIRC cancer dataset for the top-10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\text{TV}}$ values and Table 6.A.2 summarizes the detected cut-points values for each method. Once again, the selected genes are relevant for cancer studies including kidney cancer. For instance, EIF4EBP2 is related to cancer proliferation [Mizutani et al., 2016]), RGS17 is known to be overexpressed in various cancers [James et al., 2009], and both COL7A1 and NUF2 are known to be related to renal cell carcinoma (see [Csikos et al., 2003] and [Kulkarni et al., 2012] respectively).

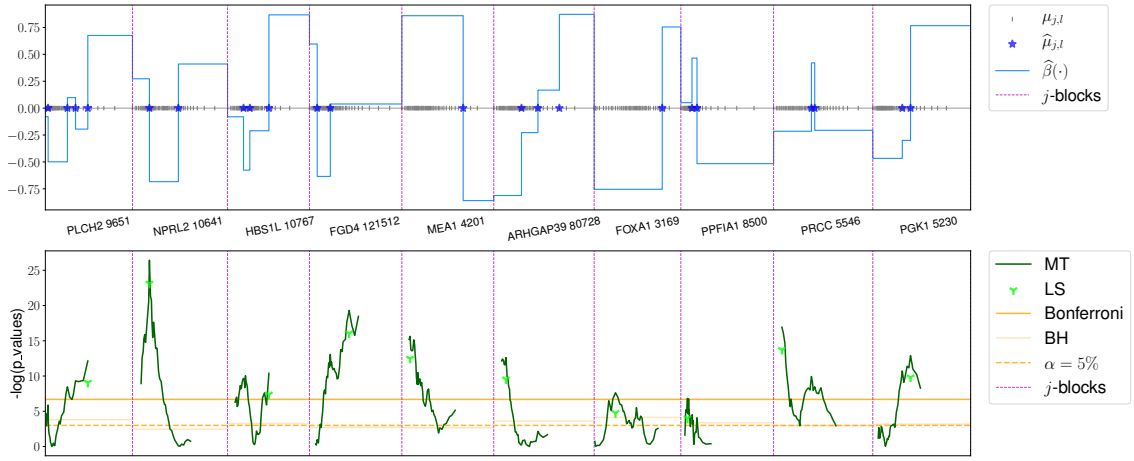


FIGURE 6.A.4 Illustration of the results obtained on the top−10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the BRCA dataset.

TABLE 6.A.1 Estimated cut-points values for each method on the top−10 genes presented in Figure 6.A.4 for the BRCA cancer.

Genes	Binacox	MT-B	MT-LS
PLCH2 9651	28.43, 200.74, 273.04, 382.87	382.87	382.87
NPRL2 10641	330.64, 568.06	330.64	330.64
HBS1L 10767	1023.91, 1212.54, 1782.77	1782.77	1782.77
FGD4 121512	163.59, 309.24	517.90	517.90
MEA1 4201	2199.21	786.29	786.29
ARHGAP39 80728	493.01, 734.37, 1049.04	265.26	265.26
FOXA1 3169	11442.32	3586.03	3586.03
PPFIA1 8500	1500.02, 1885.27	1152.98	1152.98
PRCC 5546	2091.16, 2194.08	1165.49	1165.49
PGK1 5230	10205.72, 12036.29	12036.29	12036.29

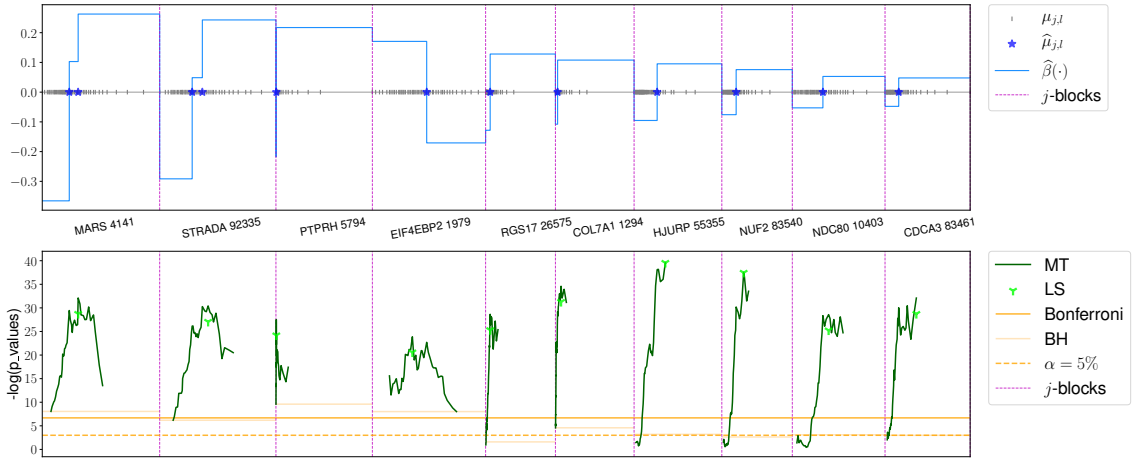


FIGURE 6.A.5 Illustration of the results obtained on the top−10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the KIRC dataset.

TABLE 6.A.2 Estimated cut-points values for each method on the top−10 genes presented in Figure 6.A.5 for the KIRC cancer.

Genes	Binacox	MT-B	MT-LS
MARS 4141	1196.21, 1350.00	1350.00	1350.00
STRADA 92335	495.24, 553.73	586.88	586.88
PTPRH 5794	3.32	3.32	3.32
EIF4EBP2 1979	6504.80	5455.59	5455.59
RGS17 26575	4.30	4.30	4.30
COL7A1 1294	44.19	113.08	113.08
HJURP 55355	99.83	134.31	134.31
NUF2 83540	42.18	63.09	63.09
NDC80 10403	91.39	107.53	107.53
CDCA3 83461	52.03	110.18	110.18

6.B Proofs

In this section, we provide the proofs of the main theoretical results. Before that, we derive some preliminaries that will be used in the proofs.

6.B.1 Preliminaries to the proofs

Additional notations. For $u, v \in \mathbb{R}^m$, we denote by $u \odot v$ the Hadamard product

$$u \odot v = (u_1 v_1, \ldots, u_m v_m)^\top.$$

We denote by $\mathrm{sign}(u)$ the subdifferential of the function $u \mapsto |u|$, that is

$$\mathrm{sign}(u) = \begin{cases} \{1\} & \text{if } u > 0, \\ [-1, 1] & \text{if } u = 0, \\ \{-1\} & \text{if } u < 0. \end{cases}$$

We write $\partial(\phi)$ the subdifferential mapping of a convex functional ϕ.

We adopt in the proofs counting processes notations. We then define the observed-failure counting process

$$N_i(t) = \mathbb{1}(Z_i \leq t, \Delta_i = 1),$$

the at-risk process

$$Y_i(t) = \mathbb{1}(Z_i \geq t),$$

and

$$\bar{N}(t) = n^{-1} \sum_{i=1}^{n} N_i(t).$$

For every vector v, let $v^{\otimes 0} = 1$, $v^{\otimes 1} = v$, and $v^{\otimes 2} = vv^\top$ (outer product). Let $\tau > 0$ be the finite study duration.

Weights. For a given numerical constant $c > 0$, the weights $\omega_{j,l}$ have an explicit form given by the following :

$$\omega_{j,l} = 5.64 \sqrt{\frac{c + \log(p+d) + \mathscr{L}_{n,c}}{n}} + 18.62 \frac{(c + \log(p+d) + 1 + \mathscr{L}_{n,c})}{n} \tag{6.13}$$

where $\mathscr{L}_{n,c} = 2 \log\log\Big((2en + 24ec) \vee e\Big)$.

Properties of binarsity penalty. We define $\omega = (\omega_{1,\bullet}, \ldots, \omega_{p,\bullet})$ the weights vector, with $\omega_{j,1} = 0$ for all $j = 1, \ldots, p$. Then, we rewrite the total variation part in binarsity as follows : let us define the $(d_j + 1) \times (d_j + 1)$ matrix D_j by

$$D_j = \begin{bmatrix} 1 & 0 & & 0 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{d_j+1} \times \mathbb{R}^{d_j+1}.$$

We remark that for all $\beta_{j,\bullet} \in \mathbb{R}^{d_j+1}$,

$$\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} = \|\omega_{j,\bullet} \odot D_j \beta_{j,\bullet}\|_1,$$

where $\odot$ denotes the component-wise product (Hadamard product). Moreover, note that the matrix D_j is invertible. We denote its inverse T_j, which is defined by the $(d_j + 1) \times (d_j + 1)$ lower triangular matrix with entries $(T_j)_{r,s} = 0$ if $r < s$ and $(T_j)_{r,s} = 1$ otherwise. We set

$$\mathbf{D} = \mathrm{diag}(D_1, \cdots, D_p) \quad \text{and} \quad \mathbf{T} = \mathrm{diag}(T_1, \cdots, T_p). \tag{6.14}$$

We further prove that binarsity is a sub-additive penalty (see Kutateladze [2013] for the definition of sub-additive).

Lemma 6.B.1 *For all $\beta, \beta' \in \mathbb{R}^{p+d}$, one has*

$$\mathrm{bina}(\beta + \beta') \leq \mathrm{bina}(\beta) + \mathrm{bina}(\beta') \quad \textit{and} \quad \mathrm{bina}(-\beta) \leq \mathrm{bina}(\beta).$$

Proof of Lemma 6.B.1. The hyperplane

$$\mathrm{span}\{u \in \mathbb{R}^{d_j+1} : \mathbf{1}_{d_j+1}{}^\top u = 0\}$$

is a convex cone, then the indicator function δ_1 is sublinear (i.e., positively homogeneous + subadditive [Kutateladze, 2013]). Furthermore, the total variation penalization satisfies triangular inequality, which gives the first statement of Lemma 6.B.1. To prove the second one, we use the fact that $\delta_1(\beta_{j,\bullet}) + \delta_1(-\beta_{j,\bullet}) \geq 0$, then we obtain

$$\mathrm{bina}(-\beta) = \sum_{j=1}^{p} \Big(\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} + \delta_1(-\beta_{j,\bullet}) \Big) \leq \sum_{j=1}^{p} \Big(\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} + \delta_1(\beta_{j,\bullet})) \Big),$$

which concludes the proof of Lemma 6.B.1.

□

Additional usefull quantities. The Doob-Meyer decomposition [Aalen, 1978] implies that, for all $i = 1, \dots, n$ and all $t \geq 0$

$$dN_i(t) = Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}dt + dM_i(t)$$

where the martingales M_i are square integrable and orthogonal.

With this notations, we define, for all $t \geq 0$ and f, the processes

$$S_n^{(r)}(f,t) = \sum_{i=1}^n Y_i(t)e^{f(X_i)}(X_i^B)^{\otimes r},$$

for $r = 0, 1, 2$ and where X_i^B is the i-th row of the binarized matrix $\boldsymbol{X}^B$.

The empirical loss ℓ_n can then be rewritten as

$$\ell_n(f) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big\{f(X_i) - \log\Big(S_n^{(0)}(f,t)\Big)\Big\}dN_i(t).$$

Together with this loss, we introduced the loss

$$\begin{aligned}\ell(f) &= -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big\{f(X_i) - \log\Big(S_n^{(0)}(f,t)\Big)\Big\}Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}dt \\ &= -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \log\Big(e^{f(X_i)}/S_n^{(0)}(f,t)\Big)Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}dt.\end{aligned}$$

We will use the fact that, for a function f_β of the form

$$f_\beta(X_i) = \beta^\top X_i^B = \sum_{j=1}^p f_{\beta_j,\bullet}(X_i),$$

the Doob-Meyer decomposition implies that

$$\begin{aligned}\nabla\ell_n(f_\beta) &= -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{X_i^B - \frac{S_n^{(1)}(f_\beta,t)}{S_n^{(0)}(f_\beta,t)}\right\}dN_i(t) \\ &= \nabla\ell(f_\beta) + H_n(f_\beta) \qquad (6.15)\end{aligned}$$

where $H_n(f_\beta)$ is an error term defined by

$$H_n(f_\beta) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big\{X_i^B - S_n^{(1)}(f_\beta,t)/S_n^{(0)}(f_\beta,t)\Big\}dM_i(t) \qquad (6.16)$$

We introduce also the empirical ℓ_2-norm defined for any function f as

$$\|f\|_n^2 = \int_0^\tau \sum_{i=1}^n \Big(f(X_i) - \bar{f}(t)\Big)^2 \frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star,t)}d\bar{N}(t),$$

with

$$\bar{f}(t) = \sum_{i=1}^n Y_i(t)e^{f^\star(X_i)}f(X_i)/S_n^{(0)}(f^\star,t).$$

Lemma 6.B.3 below connects it to our empirical divergence.

6.B.2 Lemmas

Thereafter are some lemmas useful for the proof of our theorem. Their proofs are postponed to Section 6.B.4

The following lemma is a consequence of the Karush-Kuhn-Tucker (KKT) optimality conditions [Boyd and Vandenberghe, 2004] for a convex optimization and the monotony of subdifferential mapping.

Lemma 6.B.2 *Let $\beta \in \mathscr{B}_{p+d}(R)$ such that $\mathbf{1}^\top \beta_{j,\bullet} = 0$, and $h = (h_{1,\bullet}^\top, \dots, h_{p,\bullet}^\top)^\top$ with $h_{j,\bullet} \in \partial\big(\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}}\big)$ for all $j \in \{1, \dots, p\}$, the following holds*

$$\langle \nabla \ell(f_{\hat\beta}), \hat\beta - \beta \rangle \leq -\langle H_n(f_{\hat\beta}), \hat\beta - \beta \rangle - \langle h, \hat\beta - \beta \rangle.$$

The following lemma is derived from the self-concordance definition and Lemma 1 in Bach [2010].

Lemma 6.B.3 *Let $\hat\beta$ be defined by Equation (6.5) and $\beta \in \mathscr{B}_{p+d}(R)$, the following inequalities hold almost-surely*

$$KL_n(f^\star, f_\beta) - KL_n(f^\star, f_{\hat\beta}) + (\hat\beta - \beta)^\top \nabla \ell(f_{\hat\beta}) \geq 0$$

$$\|f^\star - f_\beta\|_n^2 \frac{\psi(-\|f^\star - f_\beta\|_\infty)}{\|f^\star - f_\beta\|_\infty^2} \leq KL_n(f^\star, f_\beta) \leq \|f^\star - f_\beta\|_n^2 \frac{\psi(\|f^\star - f_\beta\|_\infty)}{\|f^\star - f_\beta\|_\infty^2}. \quad (6.17)$$

Let us define the nonnegative definite matrix

$$\widehat\Sigma_n(f^\star, \tau) = \sum_{i=1}^n \int_0^\tau \Big(X_i^B - \breve X_n(t)\Big)^{\otimes 2} \frac{Y_i(t) \exp f^\star(X_i)}{S_n^{(0)}(f^\star, t)} d\bar N(t),$$

where

$$\breve X_n(t) = \frac{\sum_{i=1}^n X_i^B Y_i(t) e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i(t) e^{f^\star(X_i)}}.$$

This matrix is linked to our empirical norm via the relation

$$\|f_\beta\|_n^2 = \beta^\top \widehat\Sigma_n(f^\star, \tau) \beta.$$

The proof of our main theorem requires for the matrix $\widehat\Sigma_n(f^\star, \tau)$ to fulfill a compability condition. The following lemma shows that this is true with a large probability as long as Assumption 4 is true.

Lemma 6.B.4 *Let $\zeta \in \mathbb{R}_+^{p+d}$ be a given vector of weights and $L = [L_1, \dots, L_p]$ a concatenation of index subsets. Set for all $j \in \{1, \dots, p\}$*

$$L_j = \{a_j^1, \dots a_j^{b_j}\} \subset \{1, \dots, d_j + 1\},$$

with the convention that $a_j^0 = 0$, *and* $a_j^{b_j+1} = d_j + 2$. *Then, with a probability greater than* $1 - e^{-nr_\tau^2/(8e^{2f_\infty^\star})} - 3\varepsilon$, *one has*

$$\inf_{u \in \mathscr{C}_{1,\omega}(L)\setminus\{\mathbf{0}\}} \frac{(\mathbf{T}u)^\top \widehat{\Sigma}_n(f^\star, \tau)\mathbf{T}u}{|\|u_L \odot \zeta_L\|_1 - \|u_{L^\complement} \odot \zeta_{L^\complement}\|_1|^2} \geq (\kappa_\tau^2(L) - \Xi_\tau(L))\kappa_{\mathbf{T},\zeta}^2(L),$$

where

$$\Xi_\tau(L) = 4|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\{(1+e^{2f_\infty^\star}\Lambda_0^\star(\tau))\sqrt{2/n\log(2(p+d)^2/\varepsilon)} + (2e^{2f_\infty^\star}\Lambda_0^\star(\tau)/r_\tau)t_{n,p,d,\varepsilon}^2\},$$

$$\kappa_{\mathbf{T},\zeta}(L) = \Big(32\sum_{j=1}^p\sum_{l=1}^{d_j+1}|\zeta_{j,l+1}-\zeta_{j,l}|^2 + (b_j+1)\|\zeta_{j,\bullet}\|_\infty^2\Big\{\min_{1\leq b\leq b^j}|a_j^b - a_j^{b-1}|\Big\}^{-1}\Big)^{-\frac{1}{2}}$$

and

$$\mathscr{C}_{1,\omega}(L) = \Big\{u \in \mathscr{B}_{p+d}(R) : \sum_{j=1}^p \|(u_{j,\bullet})_{L_j^\complement}\|_{1,\omega_{j,\bullet}} \leq 3\sum_{j=1}^p \|(u_{j,\bullet})_{L_j}\|_{1,\omega_{j,\bullet}}\Big\}.$$

We now state a technical result connecting the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on $\mathscr{C}_{\mathrm{TV},\omega}(L)$.

Lemma 6.B.5 *Let* Σ *and* $\tilde{\Sigma}$ *be two non-negative matrix of same size. For any* $L = [L_1, \ldots, L_p]$ *concatenation of index subsets, then*

$$\inf_{\beta\in\mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}}\frac{\beta^\top\tilde{\Sigma}\beta}{\|\beta_L\|_2^2} \geq \inf_{\beta\in\mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}}\frac{\beta^\top\Sigma\beta}{\|\beta_L\|_2^2} - |L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\max_{j,l}|\Sigma_{j,l}-\tilde{\Sigma}_{j,l}|.$$

6.B.3 Proof of Theorem 6.3.1

Combining Lemmas 6.B.2 and 6.B.3, we get

$$KL_n(f^\star, f_{\hat{\beta}}) \leq KL_n(f^\star, f_\beta) + (\hat{\beta}-\beta)^\top\nabla\ell(f_{\hat{\beta}}) \leq KL_n(f^\star, f_\beta) - \langle H_n(f_{\hat{\beta}}), \hat{\beta}-\beta\rangle - \langle h, \hat{\beta}-\beta\rangle.$$

If $-\langle H_n(f_{\hat{\beta}}), \hat{\beta}-\beta\rangle - \langle h, \hat{\beta}-\beta\rangle < 0$, the theorem holds. Let us assume for now that $-\langle H_n(f_{\hat{\beta}}), \hat{\beta}-\beta\rangle - \langle h, \hat{\beta}-\beta\rangle \geq 0$.

Bound for $-\langle H_n(f_{\hat\beta}), \hat\beta - \beta\rangle - \langle h, \hat\beta - \beta\rangle$. From the definition of the sub-gradient $\hat h = (\hat h_{1,\bullet}^\top, \dots, \hat h_{p,\bullet}^\top)^\top \in \partial\big(\|\hat\beta\|_{\mathrm{TV},\omega}\big)$, one can choose h such that,

$$h_{j,l} = \begin{cases} 2D_j^\top\big(\omega_{j,\bullet} \odot \operatorname{sign}(D_j\beta_{j,\bullet})\big) \text{ if } l \in \mathcal{A}_j(\beta), \\ 2D_j^\top\Big(\omega_{j,\bullet} \odot \operatorname{sign}\big(D_j(\hat\beta_{j,\bullet} - \beta_{j,\bullet})\big)\Big) \text{ if } l \in \mathcal{A}_j^{\complement}(\beta). \end{cases}$$

This gives

$$\begin{aligned} -\langle h, \hat\beta - \beta\rangle &= -\sum_{j=1}^p \langle h_{j,\bullet}, \hat\beta_{j,\bullet} - \beta_{j,\bullet}\rangle \\ &= \sum_{j=1}^p \langle (-h_{j,\bullet})_{\mathcal{A}_j(\beta)}, (\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\rangle - \sum_{j=1}^p \langle (h_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}, (\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\rangle \\ &= 2\sum_{j=1}^p \langle (-\omega_{j,\bullet} \odot \operatorname{sign}(D_j\beta_{j,\bullet}))_{\mathcal{A}_j(\beta)}, D_j(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\rangle \\ &\quad - 2\sum_{j=1}^p \langle (\omega_{j,\bullet} \odot \operatorname{sign}\big(D_j(\hat\beta_{j,\bullet} - \beta_{j,\bullet})\big)_{\mathcal{A}_j^{\complement}(\beta)}, D_j(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\rangle. \end{aligned}$$

Using the fact that $\langle \operatorname{sign}(u), u\rangle = \|u\|_1$, we have that

$$\begin{aligned} -\langle h, \hat\beta - \beta\rangle &\le 2\sum_{j=1}^p \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)} \odot D_j(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_1 \\ &\quad - 2\sum_{j=1}^p \|(\omega_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)} \odot D_j(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_1 \\ &= 2\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} - 2\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} \quad (6.18) \end{aligned}$$

Inequality (6.18) therefore becomes

$$\begin{aligned} KL_n(f^\star, f_{\hat\beta}) \le KL_n(f^\star, f_\beta) - \langle H_n(f_{\hat\beta}), \hat\beta - \beta\rangle + 2\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} \\ - 2\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}. \end{aligned}$$

Using the fact that $\mathbf{TD} = \mathbf{I}_{p+d}$ (see their definitions in Equation (6.14)), we get

$$\begin{aligned} KL_n(f^\star, f_{\hat\beta}) \le KL_n(f^\star, f_\beta) - \langle \mathbf{T}^\top H_n(f_{\hat\beta}), \mathbf{D}(\hat\beta - \beta)\rangle \\ + 2\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} - 2\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}. \end{aligned}$$

On the event

$$\mathscr{E}_n := \Big\{ |\mathbf{T}^\top H_n(f_{\hat\beta})| \leq (\omega_{1,1}, \dots, \omega_{p,d_p+1}) \Big\}$$

(the vector comparison has to be understood element by element), we have

$$\begin{aligned} KL_n(f^\star, f_{\hat\beta}) \leq KL_n(f^\star, f_\beta) &+ \sum_{j=1}^{p} \sum_{l=1}^{d_j+1} \omega_{j,l} |\Big(\mathbf{D}(\hat\beta - \beta)\Big)_{j,l}| \\ &+ 2\sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} - 2\sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}. \end{aligned}$$

Hence,

$$\begin{aligned} KL_n(f^\star, f_{\hat\beta}) &\leq KL_n(f^\star, f_\beta) + \sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} \\ &\quad + \sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} + 2\sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} \\ &\quad - 2\sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} \\ &\leq KL_n(f^\star, f_\beta) + 3\sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} \\ &\quad - \sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}. \end{aligned}$$

One therefore has

$$KL_n(f^\star, f_{\hat\beta}) \leq KL_n(f^\star, f_\beta) + 3\sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}. \tag{6.19}$$

On the event $\mathscr{E}_n$, the following also holds

$$\sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} \leq 3\sum_{j=1}^{p} \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}},$$

this means that

$$\begin{aligned} \hat\beta - \beta &\in \mathscr{C}_{\mathrm{TV},\omega}(\mathcal{A}(\beta)) \\ \text{and} \quad \mathbf{D}(\hat\beta - \beta) &\in \mathscr{C}_{1,\omega}(\mathcal{A}(\beta)) \end{aligned}$$

Now returning to (6.19), by Lemma 6.B.4 and under Assumption 4, we get

$$KL_n(f^\star, f_{\hat\beta}) \leq KL_n(f^\star, f_\beta) + \frac{\|f_{\hat\beta} - f_\beta\|_n}{\sqrt{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa_{\mathbf{T},\hat\zeta}(\mathcal{A}(\beta))}}, \tag{6.20}$$

where

$$\hat{\zeta}_{j,l} = \begin{cases} 3\omega_{j,l} & \text{if } l \in \mathcal{A}(\beta) \\ 0 & \text{if } l \in \mathcal{A}^{\complement}(\beta). \end{cases}$$

The second term in the right hand side of (6.20) fulfills

$$\frac{\|f_{\hat{\beta}} - f_\beta\|_n}{\sqrt{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))}} \leq \frac{\|f^\star - f_{\hat{\beta}}\|_n + \|f^\star - f_\beta\|_n}{\sqrt{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))}}.$$

By (6.17) in Lemma 6.B.3, we get that

$$\|f^\star - f_\beta\|_n \leq \sqrt{\frac{\|f^\star - f_\beta\|_\infty^2}{\psi(-\|f^\star - f_\beta\|_\infty)} KL_n(f^\star, f_\beta)},$$
$$\text{and } \|f^\star - f_{\hat{\beta}}\|_n \leq \sqrt{\frac{\|f^\star - f_{\hat{\beta}}\|_\infty^2}{\psi(-\|f^\star - f_{\hat{\beta}}\|_\infty)} KL_n(f^\star, f_{\hat{\beta}})}.$$

In addition, one can easily check that

$$\max_{i=1,\dots,n} \sup_{\beta \in \mathscr{B}_{p+d}(R)} |f_\beta(X_i)| \leq \sqrt{p} R,$$

hence

$$\|f^\star - f_\beta\|_\infty \leq \max_{i=1,\dots,n} \Big\{ |f^\star(X_i)| + |f_\beta(X_i)| \Big\} \leq c_{p,R,f^\star_\infty} \text{ and}$$
$$\|f^\star - f_{\hat{\beta}}\|_\infty \leq \max_{i=1,\dots,n} \Big\{ |f^\star(X_i)| + |f_\beta(X_i)| \Big\} \leq c_{p,R,f^\star_\infty}.$$

Now, using the fact that the function $u \mapsto \psi(-u)/u^2$ is decreasing, we get

$$\|f^\star - f_\beta\|_n \leq \sqrt{\frac{c^2_{p,R,f^\star_\infty}}{\psi(-c_{p,R,f^\star_\infty})} KL_n(f^\star, f_\beta)},$$
$$\text{and } \|f^\star - f_{\hat{\beta}}\|_n \leq \sqrt{\frac{c^2_{p,R,f^\star_\infty}}{\psi(-c_{p,R,f^\star_\infty})} KL_n(f^\star, f_{\hat{\beta}})}.$$

With theses bounds inequality (6.20) yields

$$KL_n(f^\star, f_{\hat{\beta}}) \leq KL_n(f^\star, f_\beta) + \sqrt{\frac{c^2_{p,R,f^\star_\infty}}{\psi(-c_{p,R,f^\star_\infty})}} \frac{\sqrt{KL_n(f^\star, f_\beta)} + \sqrt{KL_n(f^\star, f_{\hat{\beta}})}}{\sqrt{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))}}.$$

We now use an elementary inequality $2uv \leq \varrho u^2 + v^2/\varrho$ with $\varrho > 0$. We get

$$KL_n(f^\star, f_{\hat{\beta}}) \leq KL_n(f^\star, f_\beta) + \frac{\varrho}{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa^2_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))} + \frac{2c^2_{p,R,f^\star_\infty}}{\varrho\psi(-c_{p,R,f^\star_\infty})} KL_n(f^\star, f_\beta) + KL_n(f^\star, f_{\hat{\beta}}))$$

and

$$\Big(1 - \frac{2c^2_{p,R,f^\star_\infty}}{\varrho\psi(-c_{p,R,f^\star_\infty})}\Big) KL_n(f^\star, f_{\hat{\beta}}) \leq \Big(1 + \frac{2c^2_{p,R,f^\star_\infty}}{\varrho\psi(-c_{p,R,f^\star_\infty})}\Big) KL_n(f^\star, f_\beta) + \frac{\varrho}{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa^2_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))}.$$

By choosing $\varrho > 2c^2_{p,R,f^\star_\infty}/\psi(-c_{p,R,f^\star_\infty})$, we obtain

$$KL_n(f^\star, f_{\hat{\beta}}) \leq (1+\xi) KL_n(f^\star, f_\beta) + \frac{1}{1 - \frac{2c^2_{p,R,f^\star_\infty}}{\varrho\psi(-c_{p,R,f^\star_\infty})}} \frac{\varrho}{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa^2_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))}.$$

where

$$1 + \xi = \frac{\frac{\varrho\psi(-c_{p,R,f^\star_\infty})}{2c^2_{p,R,f^\star_\infty}} + 1}{\frac{\varrho\psi(-c_{p,R,f^\star_\infty})}{2c^2_{p,R,f^\star_\infty}} - 1} = 1 + \frac{2}{\frac{\varrho\psi(-c_{p,R,f^\star_\infty})}{2c^2_{p,R,f^\star_\infty}} - 1}.$$

On the other hand, by definition of $\kappa^2_{\mathbf{T},\zeta}$ (see Lemma 6.B.4), we know that

$$\frac{1}{\kappa^2_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))} \leq 512|\mathcal{A}(\beta)| \max_{j=1,\dots,p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|^2_\infty.$$

Finally,

$$KL_n(f^\star, f_{\hat{\beta}}) \leq (1+\xi) KL_n(f^\star, f_\beta) + \frac{512\varrho}{\Big(1 - \frac{2c^2_{p,R,f^\star_\infty}}{\varrho\psi(-c_{p,R,f^\star_\infty})}\Big)} \frac{|\mathcal{A}(\beta)| \max_{j=1,\dots,p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|^2_\infty}{\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta))}.$$

Therefore, on the event $\mathscr{E}_n$, we get the desired result.

Computation of $\mathbb{P}[\mathscr{E}_n^\complement]$. From the definition of H_n in Equation (6.16), $\mathbf{T}^\top H_n(f_{\hat{\beta}})$ has the form

$$(\mathbf{T}^\top H_n(f_{\hat{\beta}})) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big(\mathbf{T}^\top X_i^B - \mathbf{T}^\top \frac{S_n^{(1)}(f_{\hat{\beta}}, t)}{S_n^{(0)}(f_{\hat{\beta}}, t)}\Big) dM_i(t)$$

So each component of this vector has the form needed to applied Theorem 3 from Gaïffas and Guilloux [2012]. We recall that H_n and $\mathbf{T}^\top H_n$ have a block structure : they are vectors of p blocks of lengths d_j+1, $j=1,\dots,p$. We then denote by $\left(\mathbf{T}^\top H_n\right)_{j,l}$ the l-th component of the jth block.

In addition, due to the definition of X_i^B, we know that each coefficient of $\mathbf{T}^\top X_i^B$ is less than 1. As a consequence, for all $t \leq \tau$

$$\left|\left(\mathbf{T}^\top X_i^B - \mathbf{T}^\top \frac{S_n^{(1)}(f_{\hat\beta},t)}{S_n^{(0)}(f_{\hat\beta},t)}\right)_{j,k}\right| \leq \left|\left(\mathbf{T}^\top X_i^B\right)_{j,k}\right| + \left|\left(\mathbf{T}^\top \frac{S_n^{(1)}(f_{\hat\beta},t)}{S_n^{(0)}(f_{\hat\beta},t)}\right)_{j,k}\right| \leq 2.$$

We now use the Theorem 3 from Gaïffas and Guilloux [2012], hence we obtain

$$\mathbb{P}\left[|(\mathbf{T}^\top H_n(f_{\hat\beta},t))_{j,l}| \geq 5.64\sqrt{\frac{c+\mathscr{L}_{n,c}}{n}} + 18.62\frac{(c+1+\mathscr{L}_{n,c})}{n}\right] \leq 28.55e^{-c},$$

Then by choosing the $\omega_{j,l}$ as in (6.13), we conclude that $\mathbb{P}[\mathscr{E}_n^\complement] \leq 28.55e^{-c}$ for some $c>0$.

□

6.B.4 Proof of the Lemmas

a) Proof of Lemma 6.B.2

To characterize the solution of the problem (6.5), the following result can be sraightforwardly obtained using the Karush-Kuhn-Tucker (KKT) optimality conditions [Boyd and Vandenberghe, 2004] for a convex optimization. A vector $\hat\beta \in \mathbb{R}^{p+d}$ is an optimum of the objective function in (6.5) if and only if there exists three sequences of subgradients $\hat h = (\hat h_{j,\bullet})_{j=1,\dots,p}$ with

$$\hat h_{j,\bullet} \in \partial\Big(\|\hat\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}}\Big),$$

$\hat g = (\hat g_{j,\bullet})_{j=1,\dots,p}$ with

$$\hat g_{j,\bullet} \in \partial\Big(\delta_1(\hat\beta_{j,\bullet})\Big)$$

and $\hat k \in \partial\Big(\delta_{\mathscr{B}_{p+d}(R)}(\hat\beta)\Big)$ such that

$$\nabla \ell_n(f_{\hat\beta}) + \hat h + \hat g + \hat k = \mathbf{0}, \tag{6.21}$$

where

$$\hat h_{j,l}\begin{cases} = \left(D_j^\top\left(\omega_{j,\bullet}\odot \mathrm{sign}(D_j\hat\beta_{j,\bullet})\right)\right)_l & \text{if } l \in \mathcal{A}_j(\hat\beta),\\ \in \left(D_j^\top\left(\omega_{j,\bullet}\odot [-1,+1]^{d_j+1}\right)\right)_l & \text{if } l \in \mathcal{A}_j^\complement(\hat\beta),\end{cases}$$

and where $\mathcal{A}(\hat{\beta})$ is the active set of $\hat{\beta}$, see (6.7). The subgradient $\hat{g}_{j,\bullet}$ belongs to

$$\partial\Big(\delta_1(\hat{\beta}_{j,\bullet})\Big) = \Big\{v \in \mathbb{R}^{d_j+1} : \langle \hat{\beta}_{j,\bullet} - \beta_{j,\bullet}, v\rangle \geq 0, \text{ for all } \beta \text{ such that } \mathbf{1}^\top \beta_{j,\bullet} = 0\Big\},$$

and $\hat{k}$ to

$$\partial\Big(\delta_{\mathscr{B}_{p+d}(R)}(\hat{\beta})\Big) = \Big\{v \in \mathbb{R}^{p+d} : \langle \hat{\beta} - \beta, v\rangle \geq 0, \text{ for all } \beta \text{ such that } \|\beta\|_2 \leq R\Big\},$$

From the Equality (6.21), consider a $\beta \in \mathbb{R}^{p+d}$, we obtain

$$\langle \nabla \ell_n(f_{\hat{\beta}}), \hat{\beta} - \beta\rangle + \langle \hat{h} + \hat{g} + \hat{k}, \hat{\beta} - \beta\rangle = 0$$

and, with Equation (6.15)

$$\langle \nabla \ell(f_{\hat{\beta}}), \hat{\beta} - \beta\rangle + \langle H_n(f_{\hat{\beta}}), \hat{\beta} - \beta\rangle + \langle \hat{h} + \hat{g} + \hat{k}, \hat{\beta} - \beta\rangle = 0$$

Consider now a $\beta \in \mathscr{B}_{p+d}(R)$ and such that $\mathbf{1}^\top \beta_{j,\bullet} = 0$ for all $j \in \{1, \dots, p\}$, and $h \in \partial\Big(\|\beta\|_{\mathrm{TV},\omega}\Big)$ then the fact that the monotony of sub-differential mapping (this is an immediate consequence of its definition, see Rockafellar [1970]) gives the conclusion.

□

b) Proof of Lemma 6.B.3

Let us consider the function $G : \mathbb{R} \to \mathbb{R}$ defined by $G(\eta) = \ell(f_1 + \eta f_2)$, i.e.,

$$\begin{aligned} G(\eta) = &-\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (f_1 + \eta f_2)(X_i) Y_i(t) e^{f^\star(X_i)} \lambda_0^\star(t) dt \\ &+ \frac{1}{n}\int_0^\tau \log\Big\{S_n^{(0)}(f_1 + \eta f_2, t)\Big\} S_n^{(0)}(f^\star, t)\lambda_0^\star(t) dt. \end{aligned}$$

By differentiating G with respect to the variable η we get :

$$\begin{aligned} G'(\eta) = &-\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau f_2(X_i) Y_i(t) e^{f^\star(X_i)} \lambda_0^\star(t) dt \\ &+ \frac{1}{n}\int_0^\tau \frac{\sum_{i=1}^n f_2(X_i) Y_i(t) \exp(f_1(X_i) + \eta f_2(X_i))}{\sum_{i=1}^n Y_i(t)\exp(f_1(X_i) + \eta f_2(X_i))} S_n^{(0)}(f^\star, t)\lambda_0^\star(t) dt, \end{aligned}$$

and

$$\begin{aligned} G''(\eta) = &\frac{1}{n}\int_0^\tau \frac{\sum_{i=1}^n f_2^2(X_i) Y_i(t) \exp(f_1(X_i) + \eta f_2(X_i))}{\sum_{i=1}^n Y_i(t)\exp(f_1(X_i) + \eta f_2(X_i))} S_n^{(0)}(f^\star, t)\lambda_0^\star(t) dt \\ &- \int_0^\tau \left(\frac{\sum_{i=1}^n f_2(X_i) Y_i(t) \exp(f_1(X_i) + \eta f_2(X_i))}{\sum_{i=1}^n Y_i(t)\exp(f_1(X_i) + \eta f_2(X_i))}\right)^2 S_n^{(0)}(f^\star, t)\lambda_0^\star(t) dt. \end{aligned}$$

For a $t \geq 0$, we now consider the discrete random variable U_t that takes the values $f_2(X_i)$ with probability

$$\mathbb{P}(U_t = f_2(X_i)) = \pi_{t,f_1,f_2,\eta}(i) = \frac{Y_i(t)\exp(f_1(X_i) + \eta f_2(X_i))}{\sum_{i=1}^n Y_i(t)\exp(f_1(X_i) + \eta f_2(X_i))}.$$

We observe that for all $k = 0, 1, 2 \ldots$

$$\frac{\sum_{i=1}^n f_2^k(X_i) Y_i(t)\exp(f_1(X_i) + \eta f_2(X_i))}{\sum_{i=1}^n Y_i(t)\exp(f_1(X_i) + \eta f_2(X_i))} = \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t^k].$$

Then

$$G'(\eta) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau f_2(X_i) Y_i(t) e^{f^\star(X_i)} \lambda_0^\star(t) dt + \frac{1}{n}\int_0^\tau \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t] S_n^{(0)}(f^\star, t)\lambda_0^\star(t) dt,$$

and

$$\begin{aligned} G''(\eta) &= \frac{1}{n}\int_0^\tau \left(\mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t^2] - \left(\mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t]\right)^2\right) S_n^{(0)}(f^\star, t)\lambda_0^\star(t) dt \\ &= \frac{1}{n}\int_0^\tau \mathbb{V}_{\pi_{t,f_1,f_2,\eta}}[U_t] S_n^{(0)}(f^\star, t)\lambda_0^\star(t) dt. \end{aligned}$$

Differenciating again, we obtain

$$G'''(\eta) = \frac{1}{n}\int_0^\tau \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}\left[\left(U_t - \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t]\right)^3\right] S_n^{(0)}(f^\star, t)\lambda_0^\star(t) dt.$$

Therefore, we have

$$\begin{aligned} G'''(\eta) &\leq \frac{1}{n}\int_0^\tau \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}\left[\left|U_t - \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t]\right|^3\right] S_n^{(0)}(f^\star, t)\lambda_0^\star(t) dt \\ &\leq \frac{1}{n} 2\|f_2\|_\infty \int_0^\tau \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}\left[\left(U_t - \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t]\right)^2\right] S_n^{(0)}(f^\star, t)\lambda_0^\star(t) dt \\ &\leq 2\|f_2\|_\infty G''(\eta), \end{aligned}$$

where $\|f_2\|_\infty := \max_{i=1,\ldots,n} |f_2(X_i)|$. Lemma 1 in Bach [2010] to G, we obtain for all $\eta \geq 0$

$$G''(0)\frac{\psi(-\|f_2\|_\infty)}{\|f_2\|_\infty^2} \leq G(\eta) - G(0) - \eta G'(0) \leq G''(0)\frac{\psi(\|f_2\|_\infty)}{\|f_2\|_\infty^2}. \tag{6.22}$$

We will apply inequalities in (6.22) in two situations :

— Case #1 : $\eta = 1$, $f_1 = f_{\hat{\beta}}$ and $f_2 = f_\beta - f_{\hat{\beta}}$

— Case #2 : $\eta = 1$, $f_1 = f^\star$ and $f_2 = f_\beta - f^\star$.

In case #1,

$$G'(0) = -(\beta - \hat{\beta})^\top \frac{1}{n} \sum_{i=1}^{n} \Big\{ \int_0^\tau X_i^B Y_i(t) e^{f^\star(X_i)} \lambda_0^\star(t) dt - \int_0^\tau X_i^B Y_i(t) e^{f_{\hat{\beta}}(X_i)} \frac{S_n^{(0)}(f^\star, t)}{S_n^{(0)}(f_{\hat{\beta}}, t)} \lambda_0^\star(t) dt \Big\}$$
$$= (\beta - \hat{\beta})^\top \nabla \ell(f_{\hat{\beta}}),$$

so

$$G(1) - G(0) - G^{'}(0) = \ell(f_\beta) - \ell(\hat{f}_\beta) + (\hat{\beta} - \beta)^\top \nabla \ell(f_{\hat{\beta}}).$$

With the left bound of the self-concordance inequality (6.22), we get result 1 of lemma 6.B.3.

In case# 2, one gets

$$G^{'}(0) = 0, \text{ and}$$
$$G^{''}(0) = \frac{1}{n} \int_0^\tau \frac{\sum_{i=1}^n (f_\beta(X_i) - f^\star(X_i))^2 Y_i(t) \exp(f^\star(X_i)}{\sum_{i=1}^n Y_i(t) \exp(f^\star(X_i)} S_n^{(0)}(f^\star, t) \lambda_0^\star(t) dt$$
$$- \frac{1}{n} \int_0^\tau \left(\frac{\sum_{i=1}^n (f_\beta(X_i) - f^\star(X_i)) Y_i(t) \exp(f^\star(X_i))}{\sum_{i=1}^n Y_i(t) \exp(f^\star(X_i))} \right)^2 S_n^{(0)}(f^\star, t) \lambda_0^\star(t) dt$$
$$= \|f^\star - f_\beta\|_n^2$$

which gives result 2 of Lemma 6.B.3.

□

c) Proof of Lemma 6.B.4

For any concatenation of index sets $L = [L_1, \ldots, L_p]$, we define

$$\hat{\kappa}_\tau(L) = \inf_{\beta \in \mathscr{C}_{\mathrm{TV},\omega}(L) \setminus \{\mathbf{0}\}} \frac{\sqrt{\beta^\top \hat{\Sigma}_n(f^\star, \tau) \beta}}{\|\beta_L\|_2}.$$

To prove Lemma 6.B.4, we will first establish the following Lemma 6.B.6, which assures that if Assumption 4 is fulfilled our random bound $\hat{\kappa}_\tau(L)$ is bounded away from 0 with large probability. It bears resemblance with Theorem 4.1 of Huang et al. [2013] apart from the fact that we work here in a fixed design setting.

Lemma 6.B.6 *Let $L = [L_1, \ldots, L_p]$ be a concatenation of index sets, then the following*

$$\hat{\kappa}_\tau^2(L) \geq \kappa_\tau^2(L) - 4|L| \Big(\frac{8 \max_j (d_j + 1) \max_{j,l} \omega_{jl}}{\min_{j,l} \omega_{j,l}} \Big)^2$$
$$\times \Big\{ (1 + e^{2f_\infty^\star} \Lambda_0^\star(\tau)) \sqrt{2/n \log(2(p+d)^2/\varepsilon)} + (2 e^{2f_\infty^\star} \Lambda_0^\star(\tau)/r_\tau) t_{n,p,d,\varepsilon}^2 \Big\},$$

holds with at least probability $1 - e^{-nr_\tau^2/(8e^{2f^\star_\infty})} - 3\varepsilon$.

Proof of Lemma 6.B.6. The proof is adapted from the proof of Theorem 4.1 in Huang et al. [2013] and it is divided in 3 steps.

Step 1. By replacing $d\bar{N}(t)$ by its compensator $n^{-1}S_n^0(f^\star,t)\lambda_0^\star(t)dt$, an approximation of $\widehat{\Sigma}_n(f^\star,\tau)$ can be defining

$$\bar{\Sigma}_n(f^\star,\tau) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left(X_i^B - \breve{X}_n(s)\right)^{\otimes 2} Y_i(s) e^{f^\star(X_i)} \lambda_0^\star(s) ds.$$

The (m,m') component of

$$\sum_{i=1}^n \left(X_i^B - \breve{X}_n(s)\right)^{\otimes 2} \frac{Y_i(s)e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i(s)e^{f^\star(X_i)}}$$

is given by

$$\sum_{i=1}^n (\{X_i^B\}_m - \{\breve{X}_n(s)\}_m)(\{X_i^B\}_{m'} - \{\breve{X}_n(s)\}_{m'}) \frac{Y_i(s)e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i(s)e^{f^\star(X_i)}},$$

which, in our case,is bounded by 4. We moreover know that

$$\int_0^\tau Y_i(t) dN_i(t) \leq 1 \quad \text{for all} \quad i = 1, \ldots, n.$$

So Lemma 3.3 in Huang et al. [2013] applies and

$$\mathbb{P}\left[\{\widehat{\Sigma}_n(f^\star,\tau) - \bar{\Sigma}_n(f^\star,\tau)\}_{m,m'} > 4x\right] \leq 2e^{-nx^2/2}.$$

Via an union bound, we get that

$$\mathbb{P}\Big[\max_{m,m'}\{\widehat{\Sigma}_n(f^\star,\tau) - \bar{\Sigma}_n(f^\star,\tau)\}_{m,m'} > 4\sqrt{2/n\log(2(p+d)^2/\varepsilon)}\Big] \leq \varepsilon.$$

Let

$$\bar{\kappa}_\tau^2(L) = \inf_{\beta \in \mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}} \frac{\sqrt{\beta^\top \bar{\Sigma}_n(f^\star,\tau)\beta}}{\|\beta_L\|_2}.$$

Lemma 6.B.5 implies that

$$\mathbb{P}\left[\hat{\kappa}_\tau^2(L) \geq \bar{\kappa}_\tau^2(L) - 4|L|\left(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\right)^2 \sqrt{2/n\log(2(p+d)^2/\varepsilon)}\right] \geq 1-\varepsilon. \tag{6.23}$$

Step 2. Let

$$\widetilde{\Sigma}_n(f^\star,\tau)=\frac{1}{n}\sum_{i=1}^n\int_0^\tau\left(X_i^B-\bar{X}_n(s)\right)^{\otimes 2}Y_i(s)\;e^{f^\star(X_i)}\lambda_0^\star(s)ds$$

and

$$\tilde{\kappa}_\tau(L)=\inf_{\beta\in\mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}}\frac{\sqrt{\beta^\top\widetilde{\Sigma}_n(f^\star,\tau)\beta}}{\|\beta_L\|_2}.$$

We will now compare $\bar{\kappa}_\tau^2(L)$ and $\tilde{\kappa}_\tau^2(L)$. Straightforward computations lead to the following equality

$$\begin{aligned}&\sum_{i=1}^n\left(X_i^B-\bar{X}_n(s)\right)^{\otimes 2}Y_i(s)e^{f^\star(X_i)}-\frac{1}{n}\sum_{i=1}^n\left(X_i^B-\breve{X}_n(s)\right)^{\otimes 2}Y_i(s)e^{f^\star(X_i)}\\&=S_n^{(0)}(f^\star,s)(\breve{X}_n(s)-\bar{X}_n(s))^{\otimes 2}.\end{aligned}$$

Hence

$$\bar{\Sigma}_n(f^\star,\tau)=\widetilde{\Sigma}_n(f^\star,\tau)-\frac{1}{n}\int_0^\tau S_n^{(0)}(f^\star,s)(\breve{X}_n(s)-\bar{X}_n(s))^{\otimes 2}\lambda_0^\star(s)ds. \quad (6.24)$$

We first boud the second term on the right-hand side of (6.24). Let

$$\Delta_n(s)=\frac{1}{n}S_n^{(0)}(f^\star,s)(\breve{X}_n(s)-\bar{X}_n(s))=\frac{1}{n}\sum_{i=1}^n Y_i(s)e^{f^\star(X_i)}(X_i^B-\bar{X}_n(s))$$

so that for each (m,m')

$$\left(\frac{1}{n}\int_0^\tau S_n^{(0)}(f^\star,s)(\breve{X}_n(s)-\bar{X}_n(s))^{\otimes 2}\lambda_0^\star(s)ds\right)_{m,m'}\le\left(\frac{\int_0^\tau\Delta_n^{\otimes 2}(s)\lambda_0^\star(s)ds}{n^{-1}S_n^{(0)}(f^\star,\tau)}\right)_{m,m'}.$$

In our setting, for each i and all $t\le\tau$, $Y_i(t)\exp(f^\star(X_i)\le e^{f_\infty^\star}$. By Hoeffding inequality implies

$$\mathbb{P}[n^{-1}S_n^{(0)}(f^\star,\tau)<r_\tau/2]\le e^{-nr_\tau^2/(8e^{2f_\infty^\star})}.$$

Furthermore, we have

$$\mathbb{E}[\Delta_n(s)|X]=\frac{1}{n}\sum_{i=1}^n y_i(s)e^{f^\star(X_i)}\left(X_i^B-\frac{\sum_{i=1}^n X_i^B y_i(s)e^{f^\star(X_i)}}{\sum_{i=1}^n y_i(s)e^{f^\star(X_i)}}\right)=\mathbf{0},$$

and the (m,m') component of $\Delta_n^{\otimes 2}(s)$ is given by

$$\begin{aligned}\{\Delta_n^{\otimes 2}(s)\}_{m,m'}=\frac{1}{n^2}\sum_{i=1}^n\sum_{i'=1}^n&Y_i(s)Y_{i'}(s)e^{f^\star(X_i)}e^{f^\star(X_{i'})}\\&\times(\{X_i^B\}_m-\{\bar{X}_n(s)\}_m)(\{X_{i'}^B\}_{m'}-\{\bar{X}_n(s)\}_{m'}).\end{aligned}$$

Therefore, $\int_0^\tau \{\Delta_n^{\otimes 2}(s)\}_{m,m'}\lambda_0^\star(s)ds$ is a V-statistic for each (m, m'). Moreover,

$$\int_0^\tau |\{\Delta_n^{\otimes 2}(s)\}_{m,m'}|\lambda_0^\star(s)ds \leq 4e^{2f_\infty^\star}\Lambda_0^\star(\tau),$$

where $\Lambda_0^\star(\tau) = \int_0^\tau \lambda_0^\star(s)ds$.

By Lemma 4.2 in Huang et al. [2013], we obtain

$$\mathbb{P}\Big[\max_{1\leq m,m'\leq p+d} \pm \int_0^\tau |\{\Delta_n^{\otimes 2}(s)\}_{m,m'}|\lambda_0^\star(s)ds > 4e^{2f_\infty^\star}\Lambda_0^\star(\tau)x^2\Big]$$
$$\leq 2.221(p+d)^2 \exp\Big(\frac{-nx^2/2}{1+x/3}\Big).$$

Thanks to (6.24), Lemma 6.B.5, and the above two probability bounds, we known that

$$\bar{\kappa}_\tau^2(L) \geq \tilde{\kappa}_\tau^2(L) - 8e^{2f_\infty^\star}\Lambda_0^\star(\tau)|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2 \frac{t_{n,p,d,\varepsilon}^2}{r_\tau}, \tag{6.25}$$

with probability $1 - e^{-nr_\tau^2/(8e^{2f_\infty^\star})} - \varepsilon$.

Step 3. Now, $\widetilde{\Sigma}_n(f^\star, \tau)$ is an average of independent matrices with mean $\Sigma_n(f^\star, \tau)$ and $\{\widetilde{\Sigma}_n(f^\star, \tau)\}_{m,m'}$ are uniformly bounded by $4e^{2f_\infty^\star}\Lambda_0^\star(\tau)$ so that Hoeffding inequality assures that

$$\mathbb{P}[\max_{m,m'} |\{\widetilde{\Sigma}_n(f^\star, \tau)\}_{m,m'} - \{\Sigma_n(f^\star, \tau)\}_{m,m'} > 4e^{2f_\infty^\star}\Lambda_0^\star(\tau)x] \leq (p+d)^2 e^{-nx^2/2}.$$

Again Lemma 6.B.5 implies that, with a probability larger than $1-\varepsilon$

$$\tilde{\kappa}_\tau^2(L) \geq \kappa_\tau^2(L) - 4e^{2f_\infty^\star}\Lambda_0^\star(\tau)|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2 \sqrt{2/n\log(2(p+d)^2/\varepsilon)}, \tag{6.26}$$

Finally, the conclusion follows from (6.23), (6.25) and (6.26).This finishes the proof of Lemma 6.B.6.

□

Going back to the proof of Lemma 6.B.4, following Lemma 5 in Alaya et al. [2017], for any u in

$$\mathscr{C}_{1,\omega}(K) = \Big\{u \in \mathbb{R}^d : \sum_{j=1}^p \|(u_{j,\bullet})_{K_j^\complement}\|_{1,\omega_{j,\bullet}} \leq 3\sum_{j=1}^p \|(u_{j,\bullet})_{K_j}\|_{1,\omega_{j,\bullet}}\Big\},$$

the following holds

$$\frac{(\mathbf{T}u)^\top \widehat{\Sigma}_n(f^\star,\tau)\mathbf{T}u}{|\|u_L \odot \zeta_L\|_1 - \|u_{L^\complement} \odot \zeta_{L^\complement}\|_1|^2} \geq \kappa^2_{\mathbf{T},\zeta}(L)\frac{(\mathbf{T}u)^\top \widehat{\Sigma}_n(f^\star,\tau)\mathbf{T}u}{(\mathbf{T}u)^\top \mathbf{T}u}$$

Now, we note that if $u \in \mathscr{C}_{1,\omega}(K)$, then $\mathbf{T}u \in \mathscr{C}_{\mathrm{TV},\omega}(K)$. Hence, by the definition of $\hat{\kappa}_\tau(L)$ and Lemma 6.B.6 we get the desired result.

□

d) Proof of Lemma 6.B.5

We have that

$$|\beta^\top \tilde{\Sigma}\beta - \beta^\top \Sigma \beta| \leq \|\beta\|_1^2 \max_{j,l} |\tilde{\Sigma}_{j,l} - \Sigma_{j,l}|.$$

Then, we get

$$\beta^\top \tilde{\Sigma}\beta \geq \beta^\top \Sigma \beta - \|\beta\|_1^2 \max_{j,l} |\tilde{\Sigma}_{j,l} - \Sigma_{j,l}|.$$

So to get the desired result, it sufficient to control $\|\beta\|_1$ using the cone $\mathscr{C}_{\mathrm{TV},\omega}$. Note that for all $j = 1,\dots,p$, we have $T_j D_j = I_{d_j+1}$, where I_{d_j+1} denotes the identiy matrix in $\mathbb{R}^{d_j+1}$. Then, we have for any β

$$\begin{aligned}
\|\beta\|_1 &= \sum_{j=1}^p \|T_j D_j \beta_{j,\bullet}\| \\
&= \sum_{j=1}^p \sum_{l=1}^{d_j+1} \Big| \sum_{r=1}^{l} (D_j \beta_{j,\bullet})_r \Big| \\
&\leq \sum_{j=1}^p (d_j+1) \sum_{l=1}^{d_j+1} \Big| (D_j \beta_{j,\bullet})_l \Big| \\
&\leq \frac{\max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \sum_{j=1}^p \sum_{l=1}^{d_j+1} \omega_{j,l} \Big| (D_j \beta_{j,\bullet})_l \Big| \\
&\leq \frac{\max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \sum_{j=1}^p \|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}}.
\end{aligned}$$

For any concatenation of index subsets $L = [L_1,\dots,L_p] \subset \{1,\dots,p+d\}$, it yields

$$\|\beta\|_1 \leq \frac{\max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \Big(\sum_{j=1}^p \|(\beta_{j,\bullet})_{L_j}\|_{\mathrm{TV},\omega_{j,\bullet}} + \sum_{j=1}^p \|(\beta_{j,\bullet})_{L_j^\complement}\|_{\mathrm{TV},\omega_{j,\bullet}} \Big).$$

Now, if $\beta \in \mathscr{C}_{\mathrm{TV},\omega}(L)$, we obtain

$$\|\beta\|_1 \leq \frac{4 \max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \sum_{j=1}^p \|(\beta_{j,\bullet})_{L_j}\|_{\mathrm{TV},\omega_{j,\bullet}}.$$

Besides, one has

$$\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} \leq 2 \max_{j,l} \omega_{j,l} \|\beta_{j,\bullet}\|_1$$

Hence, we get

$$\begin{aligned}
\|\beta\|_1 &\leq \frac{8 \max_j (d_j + 1) \max_{j,l} \omega_{j,l}}{\min_{j,l} \omega_{j,l}} \sum_{j=1}^{p} \|(\beta_{j,\bullet})_{L_j}\|_1 \\
&\leq \frac{8 \max_j (d_j + 1) \max_{j,l} \omega_{j,l}}{\min_{j,l} \omega_{j,l}} \|\beta_L\|_1 \\
&\leq \sqrt{|L|} \frac{8 \max_j (d_j + 1) \max_{j,l} \omega_{j,l}}{\min_{j,l} \omega_{j,l}} \|\beta_L\|_2.
\end{aligned}$$

□

Conclusion

Les travaux présentés dans cette thèse portent sur l'introduction de nouvelles méthodes interprétables de machine learning dans un contexte de grande dimension. Différents modèles sont alors proposés, leur propriétés théoriques étudiées, et leurs performances pratiques évaluées et comparées avec l'état de l'art sur des données synthétiques et réelles.

Le Chapitre 2 présente une approche pour étudier et visualiser des variables longitudinales dans un contexte d'étude clinique. Cela a permis d'identifier des biomarqueurs pertinents à monitorer dans l'étude de cas considérée.

Le Chapitre 4 introduit un modèle de mélange de durées, le C-mix, qui surpasse l'état de l'art à la fois dans l'étude menée en simulation ainsi que dans l'application sur des données génétiques de grande dimension en cancérologie. La méthode détecte automatiquement des sous-groupes de patients suivant leur risque d'une apparition rapide de l'événement temporel étudié – offrant une interprétation immédiate et puissante des résultats – et en tire profit pour améliorer les prédictions de survie. Un algorithme efficace est proposé ainsi qu'une preuve de sa convergence. De plus, les gènes sélectionnés par le modèle dans l'application font sens cliniquement.

Le Chapitre 3 propose de comparer différents modèles en terme de performances prédictives et en sélection de variables, dans un contexte où l'évènement d'intérêt à prédire est le délai de réadmission à l'hôpital dans une étude de cas en grande dimension. Les modèles sont issus de deux cadres distincts : la prédiction binaire où il s'agit de prévoir si la réadmission va se produire ou non avant un certain seuil fixé *a priori*, et l'analyse de survie qui se défait de ce choix *a priori*. L'étude met l'accent sur l'importance de considérer un large éventail de méthodes face à une application pratique donnée : les conclusions de chaque méthode, en terme de performance prédictive et de sélection de variables, sont complémentaires et à tirer en fonction des hypothèses de la méthode. Il semble aussi qu'entraîner un modèle de survie, qui considère l'information temporelle complète de la sortie, pour ensuite prédire la survie à un seuil donné surpasse les performances prédictives des modèles entraînés directement dans le cadre binaire. Le C-mix présenté dans le chapitre précédent obtient les meilleurs résultats et fournit des interprétations intéressantes.

Le Chapitre 5 introduit une pénalité, appelée binarsity, qui s'applique sur l'en-

codage “one-hot” de covariables continues. Il s'agit d'une combinaison entre une pénalité par variation totale pondérée et une contrainte linéaire par bloc. Une inégalité oracle non-asymptotique à vitesse rapide est proposée dans le cadre des modèles linéaires généralisés. Une étude comparative de la méthode avec l'état de l'art est conduite sur de nombreux jeux de données standards démontrant ses bonnes performances pratiques en terme de prédiction et de temps de calcul requis. L'interprétation qui découle de la méthode est remarquable : en plus de l'aptitude en sélection de variables, des seuils pertinents pour la tâche de prédiction sous-jacente sont automatiquement détectés dans les covariables continues initiales, ce qui offre une compréhension profonde et précise du phénomène étudié.

Ce pouvoir d'interprétation attrayant est repris dans le Chapitre 6, cette fois dans un contexte d'analyse de survie. L'idée de la méthode proposée, appelée binacox, est alors d'utiliser la pénalité binarsity dans un modèle de Cox afin de détecter de multiples seuils dans les covariables continues de façon multivariée, ce qui est un problème récurrent dans de nombreuses applications médicales, et jusque là sans méthode adaptée à la grande dimension pour y faire face. Une inégalité oracle non-asymptotique à vitesse rapide est établie, illustrant les bonnes performances théoriques de la méthode. En outre, les résultats obtenus en simulation d'une part, et sur données génétiques en cancérologie d'autre part, attestent de ses bonnes performances pratiques et de son pouvoir d'interprétation accru.

Ces travaux, considérés dans différents cadres généraux d'apprentissage statistique – en particulier l'analyse de survie et le cadre des processus de comptage – ont été motivés par des questions pratiques principalement orientées sur la prédiction de risque et le pouvoir d'interprétabilité en grande dimension. En cela, les méthodes proposées, ainsi que les résultats théoriques et pratiques obtenus, apportent des réponses aux questions posées. Les études ménées au cours de cette thèse posent des bases pour le développement de méthodes puissantes permettant de fournir des soins médicaux personnalisés. Il reste différentes pistes de recherche à envisager, dont certaines font déjà l'objet de projets entâmés, afin de poursuivre l'étude commencée et de l'étendre à d'autres types de problèmes. Certaines pistes sont évoquées dans la section qui suit.

Directions of future research

In the following, we briefly present some ideas for future research associated with the three main methods introduced in this manuscript : the C-mix, binarsity and binacox. Some of these ideas are already work in progress, while others are at the stage of thought.

Extensions of the C-mix

Let us mention some ideas of extension for the C-mix model.

Longitudinal modeling. Regarding the good performances of the C-mix on the data considered in Chapter 3, and the fact that the time-dependent features such as the average cinetic during the last 48 hours of the stay (slope) or Gaussian Processes kernels parameters appear to have significant importances on the predictions, a natural extension to be considered is to model the longitudinal aspect, instead of using aggregated values over the longitudinal features.

We then consider joint modeling for multivariate longitudinal (in addition to fixed features) and survival data using the C-mix model. This is an on progress work in collaboration with Antoine Barbieri. Let us give the main ideas without going into the equations and the required cumbersome notations.

Concerning the longitudinal model, we consider a linear mixed model that allows to describe the longitudinal data through the average evolution with the fixed effects and the subject-specific evolution with random effects [Verbeke, 1997]. Two classical approaches are found in the literature : Joint Latent Class Models (JLCMs) and Shared Mixed Effect Models (SMEMs).

JLCMs assume that the population is heterogeneous and that there exist homogeneous subpopulations characterized by both the longitudinal profile and the risk to develop the event. The latent structure is a latent discrete variable defining the class membership [Lin et al., 2002, Proust-Lima et al., 2014]. By contrast, the latent structure of SMEMs is a function of the mixed effects which is included as feature for the survival model [Henderson et al., 2000, Rizopoulos, 2012].

Joint modeling has already been extended to multivariate longitudinal outcomes for both JLCMs [Proust-Lima et al., 2016] and SMEMs [Andrinopoulou et al., 2014]. Very recently, Andrinopoulou et al. [2018] propose to integrate latent classes in the shared parameter joint model in a fully Bayesian approach for univariate longitudinal data. We then consider this approach in the more general setting of high-dimensional multivariate longitudinal data for each individual, with a comparison of regularization techniques such as spike and slab [Ishwaran et al., 2005] or the Bayesian elastic net [Li et al., 2010]. Nonparametric Bayesian methods may also be considered regarding the latent class allocation aspect.

We wish to apply this extension on a high dimensional dataset on clinical and longitudinal data including chemotherapy doses and toxicities for patients diagnosed with colorectal cancer.

Nonparametric survival. In Section 4.3.3, we stressed the fact that on the one hand, the choice of the mixture densities f_k could be crucial regarding the prediction performances. Thus, a nonparametric approach may be considered for further investigations on this matter.

On the other hand, we observed the importance of imposing an order between the survival functions of the subpopulations. This actually sounds quite natural, since in the C-mix framework, we suppose that the survival data arise from K heterogeneous subpopulations in terms of their risk relatively to early death. Hence, we could try to assume directly within the model that the life lengths in the distinct subpopulations are stochastically ordered with each other. Indeed, survival estimates, if determined separately from the subpopulations for instance, may not be consistent with this prior assumption, because of inherent statistical variability in the observations.

This problem has been considered in a number of papers. The concept of stochastic ordering of distributions was first introduced in Lehmann [1955]. A real random variable X_1 with a distribution function F_1 is stochastically larger than another random variable X_2 with a distribution function F_2 if $F_1(x) \leq F_2(x)$ for all $x \in \mathbb{R}$. Statistical inference under stochastic ordering for the two-sample case has a rich history : see for instance Brunk et al. [1966], Dykstra [1982], and Dykstra and Feltz [1989] that consider the estimation of survival functions under an arbitrary partial stochastic ordering of the underlying populations. We also refer to Rojo and Samaniego [1993] and Mukerjee [1996] that give consistent estimators, Arjas and Gasbarra [1996] that consider a Bayesian approach, and El Barmi and Mukerjee [2005] for the case of K subpopulations.

Finally, S_1 and S_2 being two survival functions, we say that S_1 is uniformly stochastically smaller than S_2 if $S_1(x)/S_2(x)$ is nonincreasing on $\{x : S_2(x) > 0\}$,

which is a stronger version than simple stochastic ordering. We refer to Dykstra et al. [1991] that consider nonparametric maximum likelihood estimation for K subpopulation under uniform stochastic ordering and derives closed-form estimates with right-censored data. This kind of nonparametric estimators may also be considered for the C-mix.

Other penalties. Another idea of extension is to try different penalties, other than the elastic net one, and to study both theoretical properties as well as investigate and compare practical performances.

A first natural penalty we could think of is the group lasso [Yuan and Lin, 2006], where we recall that it penalizes the coefficients vector $\beta \in \mathbb{R}^d$ by

$$\lambda \sum_{k=1}^{K} \sqrt{p_k} \|\beta_{G_k}\|_2$$

with

$$\|\beta_{G_k}\|_2 = \sqrt{\sum_{j \in G_k} \beta_j^2},$$

where $\lambda \geq 0$ is the regularization parameter to be tuned and $(G_k)_{k=1,\dots,K}$ a partition of $\{1, \dots, d\}$ such that each group G_k is composed by $|G_k| = p_k$ features. This penalty generalizes the lasso by involving a per-group sparsity, which is of particular interest in the context of prognosis study relating to the C-mix, where families of features are often considered (for instance related to a given shared biological concept). The group fused lasso could also be considered [Alaíz et al., 2013].

Another penalty that would be interesting to consider is the sorted-ℓ_1 penalization (SLOPE) recently introduced in Bogdan et al. [2015]. It penalizes the coefficients vector $\beta \in \mathbb{R}^d$ by

$$\sum_{j=1}^{d} \lambda_j |\beta|_{(j)}$$

where $\lambda = (\lambda_1, \dots, \lambda_d)^\top \in \mathbb{R}_+^d$ with non-negative and non-increasing coefficients, and $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(d)}$ are the decreasing absolute values of β. Hence, this regularizer penalizes the regression coefficients according to their rank, following the Benjamini-Hochberg (BH) procedure idea that compares more significant p-values with more stringent thresholds [Benjamini and Hochberg, 1995]. Very strong statistical properties have been shown for the SLOPE penalty [Bogdan et al., 2015]. In particular, under orthogonal designs and chosing the BH related sequence $\lambda_j = z(1 - jq/2d)$ with $q \in (0, 1)$ and $z(\alpha)$ the quantile of order α of a standard normal distribution, SLOPE allows to recover the support of the regression coefficients with

a control on the False Discovery Rate (FDR), and retains good properties under more general designs.

Let us precise that in a classical multivariate regression model, considering the multi-test problem with null-hypothese $H_{0,j}$ corresponding to assuming that the $j-$th true coefficient equals 0, and denoting V (resp. R) the total number of false rejections (resp. total number of rejections), the FDR for estimator $\hat{\beta}$ is defined as

$$\mathrm{FDR}(\hat{\beta}) = \mathbb{E}\left[\frac{V}{\max\{R,1\}}\right],$$

being the expected proportion of irrelevant features among all selected ones.

Hence, it seems promising to consider the recent SLOPE penalty in other settings. For instance, it is used in Virouleau et al. [2017] where a procedure is introduced in a context of high-dimensional robust regression with guaranteed FDR and statistical power control for outliers detection under the mean-shift model. It is noteworthy to precise that to our knowledge, SLOPE has not yet been considered in a Cox model framework, which is also of high interest.

Nonparametric Bayesian methods. In Bayesian statistics, we model the parameter as a random variable : the value of the parameter is unknown and all forms of uncertainty is expressed as randomness. A nonparametric Bayesian model is a Bayesian model whose parameter space has infinite dimension. It's a way of getting flexible models that can automatically infer an adequate model size/complexity from the data.

In a nonparametric Bayesian mixture model, it is not necessary a priori to limit the number of components to be finite, and the problem of finding the "right" number of mixture components vanishes [Rasmussen, 2000]. Dirichlet processes play an important role in this setting [Antoniak, 1974], the latter being for instance at the very heart of the well know generative probabilistic model called Latent Dirichlet Allocation [Blei et al., 2003].

With the aim of considering the C-mix model on practical problems where the number of subpopulations is not fixed to 2, and to avoid invoking model selection methods to determine a suitable number of subpopulations (which is costly in terms of computing time, especially in a high dimensional context), we wish to consider and take advantage of the aforementioned appealing assets.

Extensions of binarsity

The context of the proposed idea is the Reinforcement Learning (RL) one, which is briefly introduced in the following paragraph.

Framework. A Markov Decision Process (MDP) is a 5-tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, with $\mathcal{S}$ the state space, $\mathcal{A}$ the action space, P a Markovian transition function such that $P(s'|s,a)$ denotes the probability density of a transition to state s' when taking action a in state s, R a reward function such that $R(s,a,s')$ denotes the expected reward for taking action a in state s and transitioning to state s', and $\gamma \in [0,1)$ a discount factor for future rewards. A policy π for an MDP is a mapping $\pi : \mathcal{S} \to \mathcal{A}$ from states to actions, such that $\pi(s)$ denotes the action choice in state s. The value function $V^\pi(s)$ evaluated in state s under a policy π is defined as the expected total discounted reward when the process begins in state s and all decisions are made according to policy π, that is

$$V^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t) | s_0 = s, \pi]$$

for all $s \in \mathcal{S}$. Given a fixed policy π, its value function V^π satisfies the Bellman equation

$$V^\pi(s) = R(s) + \gamma \int_{\mathcal{S}} P(s'|s,a) V^\pi(s') \mathrm{d}s'$$

for all $s \in \mathcal{S}$. We also define the optimal value function according to

$$V^\star(s) = \max_\pi V^\pi(s)$$

for all $s \in \mathcal{S}$, that is the best possible expected sum of discounted rewards that can be attained using any policy. The Bellman equation for the optimal value function is then given by

$$V^\star(s) = R(s) + \max_{a \in \mathcal{A}} \gamma \int_{\mathcal{S}} P(s'|s,a) V^\star(s') \mathrm{d}s'$$

for all $s \in \mathcal{S}$, and we define the optimal policy as follows

$$\pi^\star(s) = \mathrm{argmax}_{a \in \mathcal{A}} \int_{\mathcal{S}} P(s'|s,a) V^\star(s') \mathrm{d}s'$$

for all $s \in \mathcal{S}$, that straightforwardly verifies $V^\star(s) = V^{\pi^\star}(s)$ for all $s \in \mathcal{S}$.

In RL, a learner interacts with a stochastic process modeled as an MDP and typically observes the state and immediate reward at every step ; however, the transition model P and the reward function R are not accessible and need to be estimated. The

goal is to learn an optimal policy using the experience collected through interaction with the process. At each step of interaction, the learner observes the current state s, chooses an action a, and observes the resulting next state s' as well as the reward received r.

Problem. In many practical problems, data is acquired from dynamic physical systems that naturally involve continuous-valued state and action spaces. Then, a common solution procedure when learning to act in such continuous environments involves discretizing over the state and action dimensions to reduce the problem to a discrete MDP. Indeed, for finite state and action spaces MDP (that is when $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$), many efficient algorithms exist, are well studied and quicky converge to retrieve $\pi^\star$. For instance, one can use the fact that Bellman's equation can be used to efficiently solve for V^π in a finite-state MDP, since we can write down one such equation for $V^\pi(s)$ for every state s, which gives a set of $|\mathcal{S}|$ linear equations in $|\mathcal{S}|$ variables that can be efficiently solved.

However, simple discretization methods quickly run into the curse of dimensionality (for instance, the total number of discrete states grows exponentially with a uniform discretization), and chosing a "good" discretization is difficult to avoid a naive representation of V^π. Nevertheless, the discretization approach can work very well for many problems, but with current discretization methods, the dimension of the problem needs to stay relatively low (especially for $|\mathcal{S}|$). Let us precise that alternative methods also exist to overcome the problem of continuous state and action spaces, based for example on value function approximation, and we refer to Sutton et al. [1998] for more details on the basic concepts in RL we just introduced, as well as the main algorithms to solve MDP. But our concern here is on the choice of the discretization strategy, since a uniform choice over all dimensions is almost always considered.

Let us first point out a few studies. In Weinstein and Littman [2012] for instance, the MDP setting is the classical multi-armed bandit one, where the learner makes a decision as to which of the finite number arms to pull at each time step, attempting to maximize reward. In this context, the UCT algorithm (that only applies in discrete domains, see Kocsis and Szepesvári [2006]) is compared with a proposed algorithm called HOLOP, which is based on Hierarchical Optimistic Optimization (HOO) extended to sequential decision making, HOO being a bandit algorithm that optimizes the regret (*i.e.* maximizing immediate reward), with an assumption that the set of arms forms a general topological space (see Bubeck et al. [2009] for details). Conclusions are that in spite of a large number of possible discretizations for UCT, HOLOP yields significantly and systematically better performances. But the considered discretizations are restricted to uniform grids, with same steps for both state and action dimensions, which sounds unfair and suboptimal since the discretization

choice must significantly impact the performances of the RL algorithm trained on top.

The same kind of comparisons – and consequently, the same conclusions – are obtained in Mansley et al. [2011] between UTC using uniform fixed discretization of state and action dimensions, and another introduced algorithm called HOOT (HOO applied to Trees).

In Feng et al. [2004], the idea of using an adaptive and more clever discretization is considered in the following sense. At each step of the dynamic programming and by exploiting the structure in the problem, the state space is dynamically partitioned into regions where the value function is the same throughout the region, while reserving a fine discretization for the regions of the state space where it is the most useful. The procedure is here based on algorithms such as SPI [Boutilier et al., 2000] and SPUDD [Hoey et al., 1999] that identify regions of the state-space having the same value under the optimal policy.

Idea. The main idea is to propose another clever discretization strategy using an ad hoc regression model penalized by the binarsity on the discretized state space, with the supervision made on the reward, either using past data collected before the next action to be taken (after a few starting steps using a classical uniform grid for instance, to get a "reasonable" amount of data collected, after what the discretization could be dynamically updated), or under the common assumption that we have access to a generative model of the environment to produce data.

Extensions of binacox

We mention here two extensions around the binacox method. The first one is an ongoing work for trying to prove consistency in cut-points detection. The second is an extension devoted to combining binacox with a high-dimensional sparse second order interaction model.

Consistency of cut-points detection

In this subsection, we just give the form of the theorems we would like to prove after introducing some useful quantities.

Approximation of $f^\star$. Let us recall that our estimator of $f^\star$ is by construction given by $\hat{f} = f_{\hat{\beta}}$. Since $\beta^\star \in \mathbb{R}^{p+K^\star}$ and $\hat{\beta} \in \mathbb{R}^{p+d}$, we define in this section an approximation of $f^\star$ denoted $f_{b^\star}$ with $b^\star \in \mathbb{R}^{p+d}$. For each single j-th block, we

associate to $\beta^\star_{j,\bullet}$ the $\mu^\star_{j,\bullet}$-piecewise constant function

$$f^\star_{j,\bullet}(\cdot) : x \mapsto \sum_{k=1}^{K^\star_j+1} \beta^\star_{j,k} \mathbb{1}(x \in I^\star_{j,k})$$

defined for all $x \in [0,1]$. Note that the association is isometric. We then define the vector $b^\star = (b^{\star\top}_{1,\bullet}, \dots, b^{\star\top}_{p,\bullet})^\top$ such that $b^\star_{j,\bullet}$ is associated to the μ-piecewise constant function

$$f_{b^\star_{j,\bullet}}(\cdot) : x \mapsto \sum_{k=1}^{K^\star_j+1} \beta^\star_{j,k} \sum_{l=l^\star_{j,k-1}+1}^{l^\star_{j,k}} \mathbb{1}(x \in I_{j,l}), \tag{6.27}$$

with the convention $l^\star_{j,0} = 0$ and $l^\star_{j,K^\star_j+1} = d_j + 1$, for all $j \in \{1, \dots, p\}$. With this definition, $f_{b^\star_{j,\bullet}}$ has the same number and amplitude of jumps as $f^\star_{j,\bullet}$. The only difference between those two functions is the location of the jumps : $f^\star_{j,\bullet}$ jumps once for each cut-point $\mu^\star_{j,k}$ for all $k \in \{1, \dots, K^\star_j + 1\}$, while $f_{b^\star_{j,\bullet}}$ jumps once for each $\mu_{j,l}$ the closest (on the right hand side) to cut-point $\mu^\star_{j,k}$ for all $k \in \{1, \dots, K^\star_j + 1\}$. Figure 6.1 gives a clearer view of the different quantities involved in the estimation procedure on a toy example.

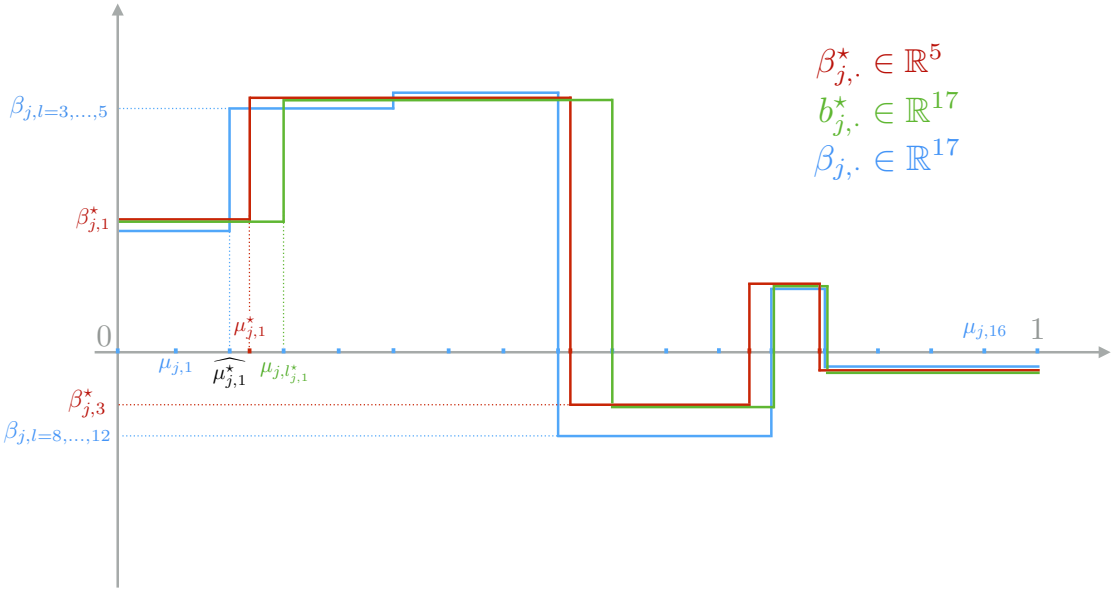


FIGURE 6.1 Illustration of vectors for a given block j with $d_j = 17$. In this scenario, the algorithm detects an extra cut-points and $\widehat{K^\star}_j = 5 = s_j$ while $K^\star_j = 4$.

Remark 6.2.1 *It may be tempting to define $b^\star$ such that*

$$f_{b^\star_{j,\bullet}}(\cdot) \in \operatorname{argmin}_{f_{\beta_{j,\bullet}}(\cdot) \in \mathcal{P}^{\mu_{j,\bullet}}} \|f_{\beta_{j,\bullet}}(\cdot) - f^\star_{j,\bullet}(\cdot)\|_{\mathcal{Q}}$$

for all $j \in \{1, \dots, p\}$, *with* $\mathcal{P}^{\mu_{j,\bullet}}$ *the set of* $\mu_{j,\bullet}$*-piecewise constant functions defined on* $[0,1]$, *and* $\mathcal{Q}$ *denoting either the Hilbert space over* $[0,1]$ *endowed by the norm* $\|f\|^2 = \int_0^1 f^2(x)dx$, *or the complete normed vector space of real integrable functions in the Lebesgue sense. In the first case* ($\mathcal{Q} = L^2([0,1])$), $f_{b^\star_{j,\bullet}}(\cdot)$ *could be viewed as an orthogonal projection. But the resulting approximated vector* $b^\star$ *would almost surely have a support set relative to the total variation penalization doubled in size compared to* $\beta^\star$*'s one, which goes against intuition. In the second case* ($\mathcal{Q} = L^1([0,1])$), *both* $\beta^\star$ *and* $b^\star$ *have the same cardinality of their respective support set relatively to the total variation penalization. But for a given cut-point* $\mu^\star_{j,k}$, *the corresponding* $b^\star$ *cut-point is* $\mu_{j,l^\star_{j,k}-1}$ *if* $\mu^\star_{j,k}$ *is closer to* $\mu_{j,l^\star_{j,k}-1}$ *than to* $\mu_{j,l^\star_{j,k}}$ *and vice versa, which would make the writing more cumbersome. To get around this difficulty, we simply add the constraint that the corresponding* $b^\star$ *cut-point is always the right bound of* $I_{j,l^\star_{j,k}}$, *that is* $\mu_{j,l^\star_{j,k}}$, *to obtain the definition given in* (6.27).

Consistency. For trying to prove the consistency for the cut-points detection, we treat the problem in each block separately. Towards this end, we need a set of assumptions that quantifies the asymptotic interplay between the following quantities :

- $I^\star_{j,\min} = \min_{1\le k\le K^\star_j} |\mu_{j,l^\star_{j,k+1}} - \mu_{j,l^\star_{j,k}}|$ the minimum distance between two consecutive cut-points of $f_{b^\star}$,
- $J^\star_{j,\min} = \min_{1\le k\le K^\star_j} |b^\star_{j,l^\star_{j,k+1}} - b^\star_{j,l^\star_{j,k}}|$ the smallest jump amplitude of $f_{\beta^\star}$,
- $J^\star_{j,\max} = \max_{1\le k\le K^\star_j} |b^\star_{j,l^\star_{j,k+1}} - b^\star_{j,l^\star_{j,k}}|$ the biggest jump amplitude of $f_{\beta^\star}$, and
- $(\varepsilon_n)_{n\ge 1}$ a non-increasing and positive sequence that goes to 0 as $n \to \infty$.

Assumption 5 *Some assumptions on* $I^\star_{j,\min}$, $J^\star_{j,\min}$, $J^\star_{j,\max}$ *and* $(\varepsilon_n)_{n\ge1}$ *as* $n \to \infty$ *for all* $j \in \{1, \dots, p\}$.

Thus, Assumption 5 remains to be precised, and we now give the form of the first theorem we want to prove.

Theorem 6.2.7 *Assume that Assumptions 3 and 5 hold. If* $\widehat{K}_j = K^\star_j$, *the estimated cut-points* $\{\hat{\mu}_{j,1}, \dots, \hat{\mu}_{j,\widehat{K}_j}\}$ *defined by* (6.8) *satisfy*

$$\mathbb{P}\Big[\max_{1\le k\le K^\star_j} |\widehat{\mu}_{j,k} - \mu^\star_{j,k}| \le \varepsilon_n\Big] \to 1 \text{ as } n \to \infty$$

for all $j \in \{1, \dots, p\}$.

In Theorem 6.2.7, the number of estimated cut-points is assumed to be the true number of cut-points. Since this information is not available in general, we need to relax the statement of this result. We propose to evaluate the distance between the

set $\widehat{\mathcal{M}}_j = \{\hat{\mu}_{j,1}, \ldots, \hat{\mu}_{j,\widehat{K}_j}\}$ of estimated cut-points and the set of true cut-points $\mathcal{M}_j^\star = \{\mu_{j,1}^\star, \ldots, \mu_{j,K_j^\star}^\star\}$ by using the two quantities $\mathcal{E}(\widehat{\mathcal{M}}_j||\mathcal{M}_j^\star)$ and $\mathcal{E}(\mathcal{M}_j^\star||\widehat{\mathcal{M}}_j)$ already defined in Section 6.4.2. When $\widehat{K}_j = K_j^\star$, Theorem 6.2.7 implies that $\mathcal{E}(\widehat{\mathcal{M}}_j||\mathcal{M}_j^\star) \leq \varepsilon_n$ and $\mathcal{E}(\mathcal{M}_j^\star)||\widehat{\mathcal{M}}_j) \leq \varepsilon_n$ with probability that tends to 1 as $n \to \infty$. In the case where $\widehat{K}_j > K_j^\star$, let us give the form of Theorem 6.2.8 that claims that $\mathcal{E}(\widehat{\mathcal{M}}_j||\mathcal{M}_j^\star) \leq \varepsilon_n$ with a probability tending to 1 as $n \to \infty$. This means that the cut-points consistency holds for our procedure whenever the estimated number of cut-points in not less than the true one.

Theorem 6.2.8 *Assume that Assumptions 3 and 5 hold. If $\widehat{K}_j \geq K_j^\star$, the estimated cut-points $\{\hat{\mu}_{j,1}, \ldots, \hat{\mu}_{j,\widehat{K}_j}\}$ defined by (6.8) satisfy*

$$\mathbb{P}[\mathcal{E}(\widehat{\mathcal{M}}_j||\mathcal{M}_j^\star) \leq \varepsilon_n] \to 1 \text{ as } n \to \infty$$

for all $j \in \{1, \ldots, p\}$.

Theorem 6.2.8 ensures that even when the number of cut-points is over-estimated, each true cut-point is close to the estimated one.

The main efforts for the proof of Theorems 6.2.7 and 6.2.8 lie on obtaining a control of $\|(\hat{\beta} - b^\star)_{\mathcal{A}(b^\star)}\|_1$, where the idea is first to bound it by a quantity involving $\|f_{b^\star} - f_{\hat{\beta}}\|_n$ thanks to Theorem 6.3.1 and inequality (6.17) in Lemma 6.B.3. This is still a work under progress.

Second order interaction

The binacox model allows to identify significant ranges of values, say intervals, for continuous features in a prognosis study. A natural question arising from this is to evaluate the impact of two features being simultaneously within two given intervals. This is the problem of introducing two-way interaction features, meaning the entry-wise multiplication between two features. This is a well known problem [Cox, 1984], being relevant for instance in genomics to detect possible epistasis between genes. So one can think of considering this question on top of the results obtained in Chapter 6 on the TCGA data (described in Section A.2.2).

But this question is hardly discussed for genomic data in the literature, partly because of the explosion in the number of interactions to consider, especially for high dimensional problems such as genetic related ones. Yet some studies try different stategies to select only a few interactions to be included in the model, for instance if at least one of the two corresponding main features is selected by the model, but

still in relatively low dimensional settings [Bickel et al., 2010, Bien et al., 2013, Lim and Hastie, 2015, Haris et al., 2016].

However, a lot of new methods have recently been developped to accelerate solvers for sparsity constrained optimization problems, based on the sparsity of the solution to detect inactive features, that can be set aside from the coefficients to be updated during optimization. The improvement in the scalability opens up new horizons for the aforementioned questions that can be reconsidered.

Among those new methods may be found safe screening rules related algorithms [Ghaoui et al., 2010, Xiang et al., 2011, Fercoq et al., 2015] with in particular the Safe Pattern Pruning (SPP) method for binary features [Nakagawa et al., 2016] ; and working set algorithms that are less conservative by iteratively solving subproblems restricted to a subset of features in the primal or to a subset of constraints in the dual [Johnson and Guestrin, 2015], with in particular the recent WHInter algorithm for binary features that allows significant speed up [Morvan and Vert, 2018].

The idea here would be to extend the SPP and/or WHInter methods to be applied on top of the binacox and its generated binary features, which involves to consider those questions within a survival analysis framework.

Annexe A

Appendices

Sommaire

A.1 Quelques rappels

Nous donnons dans cette section quelques détails et outils utiles dans le manuscrit, sur lesquels nous sommes passés rapidement et qui peuvent nécessiter quelques rappels. Précisions qu'il ne s'agit pas de présenter de façon exhaustive les notions évoquées mais d'introduire les notations et propriétés utilisées dans les chapitres précédents. Pour plus de détails, nous renvoyons le lecteur aux livres Andersen et al. [1993], Aalen et al. [2008], Therneau and Grambsch [2013].

A.1.1 Les processus de comptage

Modèle stochastique à temps continu. On modélise un phénomène aléatoire qui évolue dans le temps par un *processus stochastique* et une *filtration*. Les phénomènes considérés dans ce manuscrit se déroulent en temps continu, le processus stochastique noté $X = (X_t)_{t\in\mathcal{T}}$ est alors représenté par une famille de variables aléatoires sur l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ où $\mathcal{A}$ est une σ-algèbre et $\mathbb{P}$ une mesure de probabilité sur $\mathcal{A}$, indexées par un intervalle $\mathcal{T} = [0, \tau[\subset \mathbb{R}^+$. Une trajectoire de X est une fonction $t \mapsto X_t(\omega)$, pour un certain $\omega \in \Omega$. Puis, une filtration exprime l'information détenue à chaque instant. Comme l'information croît avec le temps, une filtration est représentée par une suite croissante de sous-tribus. Elle est continue à droite (càd), ce qui exprime le fait que l'information détenue à chaque instant est exactement l'information du futur immédiat. On note $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathcal{T}}$ la filtration et on la définie alors formellement par

$$\text{(i) } \mathcal{F}_s \subseteq \mathcal{F}_t \text{ pour tout } s \leq t \text{ dans } \mathcal{T} \text{ (croissante)},$$
$$\text{(ii) } \mathcal{F}_s = \bigcap_{t>s} \mathcal{F}_t \text{ pour tout } s \in \mathcal{T} \text{ (càd)}.$$

La filtration contient plus d'information que le processus, ce qui se traduit par le fait qu'à chaque instant $t \in \mathcal{T}$, la variable aléatoire X_t est $\mathcal{F}_t$-mesurable. On dit alors que le processus X est *adapté* à la filtration $\mathcal{F}$.

Le processus X est dit *càdlàg* (continu à droite, limite à gauche) si $\mathbb{P}$-presque toutes ses trajectoires appartiennent à l'espace des fonctions càdlàg sur $\mathcal{T}$; et *prévisible* si en tant que fonction de $(t, \omega) \in \mathcal{T} \times \Omega$, il est mesurable par rapport à la tribu sur $\mathcal{T} \times \Omega$ engendrée par les processus càg et adaptés. Lorsque seul le processus stochastique X est observé, l'information détenue est minimale et à chaque instant $t \in \mathcal{T}$, on ne dispose de l'information sur X que jusqu'à l'instant t. Autrement dit, à tout processus X càdlàg, on peut associer une filtration dite "historique" telle que pour tout $t \in \mathcal{T}$,

$$\mathcal{F}_t = \sigma\{X(s), s \leq t \in \mathcal{T}\}$$

est la sous-tribu engendrée par le passé à t de X. Par construction, le processus est adapté à la filtration historique. La filtration $\mathcal{F}$ est supposée *complétée*, c'est-à-

dire constituée de tribus qui contiennent tous les événement de probabilité nulle (en ajoutant les événement négligeables, elle conserve sa propriété de filtration).

Un temps d'arrêt pour la filtration $\mathcal{F}$ est une variable aléatoire T à valeurs dans $[0, \tau]$ telle que pour tout t $\mathcal{T}$, $\{T \leq t\} \in \mathcal{F}_t$.

Martingale à temps continu. On dit que M est une martingale (resp. une sur-martingale, resp. une sous-martingale) relativement à la filtration $\mathcal{F}$ si M est un processus stochastique càdlàg adapté à $\mathcal{F}$ et si, pour tous $0 \leq s \leq t$, $M_t \in \mathbb{L}^1$ et

$$\mathbb{E}[M_t|\mathcal{F}_s] = M_s \text{ (resp. } \leq, \text{ resp. } \geq).$$

Une martingale M est dite de carré intégrable si

$$\sup_{t \in \mathcal{T}} \mathbb{E}[M_t^2] < +\infty.$$

En général, le concept de martingale est restrictif et on préfére la notion plus générale de martingale *locale*, où un processus M càdlàg et adapté est une martingale locale s'il existe une suite de temps d'arrêts $(T_n)_{n\in\mathbb{N}}$ telle que

(i) $(T_n)_{n\in\mathbb{N}}$ est croissante et vérifie $\lim_{n\to+\infty} T_n = +\infty$ *p.s.*,

(ii) M^{T_n} est une martingale pour tout $n \in \mathbb{N}$.

On dit que $(T_n)_{n\in\mathbb{N}}$ est une suite localisante pour la martingale locale M. On dit de plus que M est une martingale locale de carré intégrable si $(T_n)_{n\in\mathbb{N}}$ est telle que

$$M^{T_n} \mathbb{1}_{\{T_n>0\}}$$

est de carré intégrable. Énonçons alors le théorème de décomposition de Doob-Meyer utilisé au Chapitre 6.

Theorem A.1.1 *(Décomposition de Doob-Meyer) Soit X une sous-martingale positive continue à droite et adaptée à la filtration $\mathcal{F}$. Il existe alors un unique couple (M, Λ) tel qu'on ait*

$$X_t = M_t + \Lambda_t \text{ p.s. pour tout } t \in \mathcal{T},$$

avec M une martingale càd et Λ un processus prévisible croissant, càd et intégrable.

En particulier, si M est une martingale locale de carré intégrable, il existe un unique processus prévisible noté $\langle M \rangle$ et appelé *variation prévisible* de M, qui soit càdlàg, à variation finie, nul en zéro et tel que $M^2 - \langle M \rangle$ soit une martingale locale. On définit aussi le *processus de variation optionnel* de M, noté $[M]$, par

$$[M]_t = \sum_{s \leq t} \Delta M_s^2,$$

où on note ΔX_t le "saut" en t du processus càdlàg X défini par $\Delta X_t = X_t - X_{t^-}$ avec

$$X_{t^-} = \lim_{\substack{s<t \\ s\to t}} X_s.$$

Processus de comptage. Dans de nombreux cas d'application, on est amené à introduire un processus stochastique qui n'évolue que par sauts d'amplitude 1 qui correspondent à des instants aléatoires où se produisent certains événements spécifiques, comme c'est le cas dans le cadre de travail de l'analyse de survie. La modélisation adaptée est alors celle des processus de comptage. On dira alors que N est un processus de comptage s'il est adapté, si ses trajectoires sont càd, constantes par morceau, nulles à l'instant 0 et croissantes par sauts d'amplitude 1 (donc constantes par morceau). Le processus de comptage N peut être représenté par la suite de ses instants de sauts $(T_n)_{n\in\mathbb{N}^*}$ tels que $0 < N_1 < N_2 < \cdots$ $p.s.$ et pour tout $t \in \mathcal{T}$, on a la représentation

$$N_t = \sum_{k\geq 1} \mathbb{1}_{\{T_k \leq t\}},$$

ainsi que

$$N_t \geq n \;\Leftrightarrow\; T_n \leq t.$$

Il existe alors Λ_i croissants tels que $M_i = N_i - \Lambda_i$ soit une martingale locale de carré intégrable et il existe un processus prévisible f_i, appelée intensité, telle que

$$\Lambda_i = \int_0^t f_i(s)\mathrm{d}s.$$

On parle alors de processus de comptage à intensité pour N_i et on a $[M_i] = N_i$ et $\langle M_i \rangle = \Lambda_i$. Pour les processus de comptage marqués qui sont plus généraux, on observe en plus une variable X_k (appelée marque) à chaque T_k, et (T_k, X_k) est appelé processus ponctuel marqué.

A.1.2 Plus sur le modèle de Cox

Cette section est dédiée à quelques questions relatives au modèle de Cox (présenté dans la Section 1.2.3) non traitées en détail dans le manuscrit, mais jugées importantes pour comprendre en profondeur certains développements.

a) La vraisemblance partielle de Cox

La première version de la log-vraisemblance partielle de Cox que nous avons introduite est la suivante [Cox, 1972]

$$\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \Big(x_i^\top \beta - \log \sum_{i' : z_{i'} \geq z_i} \exp(x_{i'}^\top \beta) \Big). \tag{A.1}$$

Commençons par donner une justification intuitive de cette écriture.

Construction intuitive. L'idée est de considérer qu'aucune information ne peut être donnée sur $\beta^\star$ par les intervalles pendant lesquels aucun événement n'a eu lieu, et de supposer que les instants où se produisent les censures n'apportent pas d'information sur $\beta^\star$. On travaille alors conditionnellement à l'ensemble des instants où un décès a lieu.

Supposons qu'il n'y a qu'un seul décès pour chaque temps d'événement en se basant sur un raisonnement en temps continu. Les modifications à apporter en cas d'événements simultanés sont par ailleurs évoquées dans un paragraphe à venir. Si on note D le nombre total de décès observés sur la période de l'étude parmi les n individus, $t_1 < \cdots < t_D$ les temps de décès distincts et ordonnés, alors la probabilité p_i que l'individu i décède au temps t_i sachant qu'un décès a effectivement eu lieu à cet instant est naturellement donnée par

$$p_i = \frac{\lambda_0^\star(t_i)\exp(x_i^\top\beta^\star)}{\sum_{j\in R(t_i)}\lambda_0^\star(t_i)\exp(x_j^\top\beta^\star)} = \frac{\exp(x_i^\top\beta^\star)}{\sum_{j\in R(t_i)}\exp(x_j^\top\beta^\star)} \tag{A.2}$$

où $R(t_i)$ est l'ensemble des individus toujours à risque juste avant t_i. Comme il y a des contributions à la vraisemblance à chaque temps de décès, la vraisemblance partielle de Cox est définie comme le produit sur tous les temps de décès, soit

$$L_n(\beta) = \prod_{i=1}^{D} p_i = \prod_{i=1}^{n} \frac{\delta_i \exp(x_i^\top\beta)}{\sum_{j:z_j\geq z_i}\exp(x_j^\top\beta)},$$

en remarquant simplement que $R(t_i) = \{j : z_j \geq z_i\}$ et qu'en ajoutant la puissance δ_i, on ne fait intervenir que les temps de décès en écartant les temps censurés. En prenant le logarithme de l'expression précédente et en multipliant par n^{-1} (version "normalisée" de la log-vraisemblance qui est couramment utilisée), on retrouve bien l'écriture (A.1).

Autre écriture. Nous avons également donné une seconde écriture de cette log-vraisemblance partielle (qui s'exprime avec l'échantillon aléatoire) qui est la suivante

$$\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{X_i^T\beta - \log S_n(t,\beta)\}\mathrm{d}N_i(t). \tag{A.3}$$

Vérifions alors l'égalité entre l'expressions (A.1) exprimée avec l'échantillon aléatoire et l'expression (A.3). Il suffit pour cela de montrer que

$$\Delta_i = \int_0^\tau \mathrm{d}N_i(t) \tag{A.4}$$

et

$$\Delta_i \log \sum_{j:Z_j\geq Z_i}\exp(X_j^\top\beta) = \int_0^\tau \log\Big(\sum_{j=1}^{n} Y_j(t)\exp(X_j^\top\beta)\Big)\mathrm{d}N_i(t) \tag{A.5}$$

pour tout $i \in \{1, \dots, n\}$ et tout $\beta \in \mathbb{R}^d$. En rappelant que par définition

$$N_i(t) = \mathbb{1}_{\{Z_i \leq t, \Delta_i = 1\}},$$

on a

$$\begin{aligned}
\int_0^\tau \mathrm{d}N_i(t) &= \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_0^\tau \Big(N_i(t+\varepsilon) - N_i(t) \Big) \mathrm{d}t \\
&= \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_0^\tau \Delta_i \Big(\mathbb{1}_{\{Z_i \leq t+\varepsilon\}} - \mathbb{1}_{\{Z_i \leq t\}} \Big) \mathrm{d}t \\
&= \Delta_i \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \underbrace{\int_{Z_i - \varepsilon}^{Z_i} \mathrm{d}t}_{\varepsilon} \\
&= \Delta_i,
\end{aligned}$$

ce qui achève la preuve de (A.4). Ensuite, on a

$$\begin{aligned}
\int_0^\tau \log \Big(\sum_{j=1}^n \underbrace{Y_j(t)}_{\mathbb{1}_{\{Z_j \geq t\}}} \exp(X_j^\top \beta) \Big) \mathrm{d}N_i(t) &= \int_0^\tau \log \sum_{j : Z_j \geq t} \exp(X_j^\top \beta) \mathrm{d}N_i(t) \\
&= \Delta_i \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_0^\tau \log \sum_{j : Z_j \geq t} \exp(X_j^\top \beta) \Big(\mathbb{1}_{\{Z_i \leq t+\varepsilon\}} \\
&\qquad - \mathbb{1}_{\{Z_i \leq t\}} \Big) \mathrm{d}t \\
&= \Delta_i \lim_{\varepsilon \to 0} \underbrace{\frac{1}{\varepsilon} \int_{Z_i - \varepsilon}^{Z_i} \mathrm{d}t}_{=1} \log \sum_{j : Z_j \geq Z_i + \varepsilon} \exp(X_j^\top \beta) \\
&= \Delta_i \log \sum_{j : Z_j \geq Z_i} \exp(X_j^\top \beta),
\end{aligned}$$

ce qui achève la preuve de (A.5). □

Justification théorique. Nous donnons enfin une justification théorique de l'écriture (A.3) à partir de la définition de la log-vraisemblance pour les processus de comptage. On se place dans le cadre de la Section 1.2.2 en reprenant les mêmes notations, avec la fonction de risque du modèle de Cox définie dans la Section 1.2.3. D'après le Théorème A.1.1 de décomposition de Doob-Meyer, on a

$$\mathrm{d}N_i(t) = \lambda_0^\star(t) \exp(X_i^\top \beta^\star) Y_i(t) \mathrm{d}t + \mathrm{d}M_i(t),$$

où M_i est une martingale locale de carré intégrable. D'après la définition de la log-vraisemblance pour les processus de comptage définie dans Andersen et al. [1993],

on a la log-vraissemblance (normalisée par n^{-1}) suivante

$$\mathcal{L}(\lambda, \beta; \mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^{n} \Big\{ \int_0^\tau \log \Big(\lambda(t) \exp(X_i^\top \beta) \Big) \mathrm{d}N_i(t) - \int_0^\tau \lambda(t) \exp(X_i^\top \beta) Y_i(t) \mathrm{d}t \Big\},$$

que l'on peut ré-écrire

$$\begin{aligned} \mathcal{L}(\lambda, \beta; \mathcal{D}_n) = & \underbrace{\frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \log \Big(\frac{\exp(X_i^\top \beta)}{S_n(t, \beta)} \Big) \mathrm{d}N_i(t)}_{\ell_n(\beta)} + \int_0^\tau \log \Big(\lambda(t) S_n(t, \beta) \Big) \mathrm{d}\bar{N}(t) \\ & - \int_0^\tau \lambda(t) S_n(t, \beta) \mathrm{d}t, \end{aligned}$$

où $S_n(t, \beta)$ est défini dans l'Équation (1.15) et où on voit apparaître la vraissemblance partielle de Cox $\ell_n(\beta)$, dont la fonction de perte naturellement associée est donnée par

$$\begin{aligned} \ell(\beta, \beta^\star) &= \mathbb{E}[\ell_n(\beta^\star) - \ell_n(\beta)] \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big[\int_0^\tau \log \Big(\frac{\exp(X_i^\top \beta)}{S_n(t, \beta)} \frac{S_n(t, \beta^\star)}{\exp(X_i^\top \beta^\star)} \Big) \lambda_0^\star(t) \exp(X_i^\top \beta^\star) Y_i(t) \mathrm{d}t \Big] \\ &\quad \underbrace{- \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big[\int_0^\tau \log \Big(\frac{\exp(X_i^\top \beta)}{S_n(t, \beta)} \frac{S_n(t, \beta^\star)}{\exp(X_i^\top \beta^\star)} \Big) \mathrm{d}M_i(t) \Big]}_{=0}. \end{aligned}$$

On retrouve alors la divergence de Kullback-Leibler introduite dans Senoussi [1990], sa version empirique étant introduite dans l'équation (6.9) pour obtenir les résultats théoriques du Chapitre 6.

b) Détails supplémentaires

Nous discutons ci-après deux points mis de côté ou vaguement évoqués jusqu'alors.

Événements simultanés. Les raisonnements précédents supposent que les temps d'événements sont distincts. Mais dans le cas de données réelles, cette hypothèse n'est pas toujours vérifiée. En présence d'un nombre de décès $d_i \geq 1$ au temps t_i, on admet simplement que les décès se produisent les uns à la suite des autres. Cependant, on ne connaît pas l'ordre des décès et il faut donc considérer toutes les possibilités. En notant D_i les indices des d_i individus qui décèdent en t_i, l'expression (A.2) de p_i, qui est ici la probabilité que les d_i individus qui décèdent soient effectivement ceux de D_i, devient alors ici

$$p_i = \frac{\prod_{j \in D_i} \exp(x_j^\top \beta^\star)}{\sum_{P_i} \prod_{j \in P_i} \exp(x_j^\top \beta^\star)},$$

où la somme du dénominateur est sur toutes les permutations P_i de $R(t_i)$ de taille d_i (façons de choisir d_i indices parmi $R(t_i)$). Le problème est que le temps de calcul devient long lorsqu'il y a beaucoup de décès simultanés, puisque le nombre de permutations croît très rapidement avec d_i. L'approximation de Breslow utilisée en pratique consiste à remplacer le dénominateur par

$$\Big[\sum_{j\in R(t_i)} \exp(x_j^\top \beta^\star)\Big]^{d_i}$$

pour obtenir finalement

$$L_n(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta^\top \sum_{j\in D_i} x_j)}{\Big[\sum_{j\in R(t_i)} \exp(x_j^\top \beta)\Big]^{d_i}}.$$

Nous précisons que d'autres approximations plus fines existent et renvoyons à Therneau and Grambsch [2000] pour plus de détails à ce sujet.

Tests. De façon classique, il existe différentes approches pour tester des hypothèses sur $\beta^\star$. Les tests sont déduits de propriétés asymptotiques de $\hat{\beta}$ et portent en général sur une hypothèse du type

$$H_0 : \beta^\star = \beta_0.$$

Les principales statistiques de test utilisées sont :

— la statistique du rapport de vraisemblance, et sous H_0 on a

$$2[\ell_n(\hat{\beta}) - \ell_n(\beta_0)] \underset{n\to+\infty}{\longrightarrow} \chi^2(d),$$

— la statistique de Wald, et sous H_0 on a

$$(\hat{\beta} - \beta_0)^\top \nabla^2 \ell_n(\hat{\beta})(\hat{\beta} - \beta_0) \underset{n\to+\infty}{\longrightarrow} \chi^2(d),$$

— la statistique du score, et sous H_0 on a

$$\nabla \ell_n(\beta_0)^\top \Big(\nabla^2 \ell_n(\hat{\beta})\Big)^{-1} \nabla \ell_n(\beta_0) \underset{n\to+\infty}{\longrightarrow} \chi^2(d).$$

On peut bien entendu déduire de ces statistiques des tests partiels permettant de tester des hypothèses concernant certaines coordonnées de $\beta^\star$, comme par exemple

$$H_0 : \beta_j^\star = 0.$$

On peut aussi montrer que tester $H_0 : \beta^\star = 0$ est équivalent à faire un test du logrank où l'hypothèse nulle serait l'égalité des fonctions de survie de K sous-groupes d'individus.

Il existe différentes méthodes pour tester l'adéquation au modèle, notamment pour vérifier l'hypothèse des risques proportionnels. Des méthodes graphiques ont été proposées mais elles sont peu puissantes. Une autre approche consiste à considérer une covariable qui dépend du temps. Si on veut par exemple tester si une covariable X vérifie l'hypothèse de risques proportionnels, on introduit un terme d'interaction entre le temps et la covariable : par exemple $\beta_1 X + \beta_2 X \log(t)$. Il s'agit ensuite de tester si β_2 est significativement différent de 0, si c'est le cas alors l'hypothèse de proportionnalité n'est pas respectée. Une autre méthode consiste à partir de l'expression

$$S(t|X = x) = \exp\{-\Lambda(t|X = x)\}$$

de l'Équation (1.13) et de considérer la variable aléatoire $V = \Lambda(T|X)$. On remarque alors que

$$\mathbb{P}[V > v|X = x] = \mathbb{P}[T > \Lambda^{-1}(v|X = x)] = \exp(-v)$$

et donc V suit une loi exponentielle $\mathcal{E}(1)$. Une procédure de vérification de l'adéquation du modèle consiste alors à comparer l'estimateur du risque cumulé de V à la droite d'équation $v = t$. Nous renvoyons à Therneau and Grambsch [2013] pour plus de détails concernant ces différentes procédures de test.

A.1.3 Inégalités de concentration

Les inégalités de concentration fournissent des bornes sur la probabilité qu'une variable aléatoire dévie d'une certaine valeur, généralement l'espérance de cette variable aléatoire. Énonçons une première inégalité très utile en pratique, l'inégalité de Hoeffding, qui permet d'évaluer la déviation d'un processus empirique par rapport à sa moyenne.

Theorem A.1.2 *(Inégalité de Hoeffding). Soient $X_1, \ldots, X_n$ des variables aléatoires réelles indépendantes telles que $a_i \leq X_i \leq b_i$ presque sûrement. On note $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ la moyenne empirique. On a alors*

$$\mathbb{P}\Big[\bar{X}_n - \mathbb{E}[\bar{X}_n] \geq x\Big] \leq \exp \frac{-2n^2x^2}{\sum_{i=1}^n (b_i - a_i)^2}$$

pour tout $x > 0$.

Nous utilisons une version de ce résultat dans le Chapitre 6.

Le gros défaut de l'inégalité précédente et qu'elle suppose les variables X_i bornées. Énonçons simplement l'inégalité de Bernstein qui relâche cette hypothèse mais qui fait en contre partie des hypothèses de moments et qui perd un peu au niveau de la borne.

Theorem A.1.3 *(Inégalité de Bernstein). Soient $X_1, \ldots, X_n$ des variables aléatoires réelles indépendantes telles qu'il existe des nombres positifs v et c vérifiant $\sum_{i=1}^n \mathbb{E}[X_i^2] < v$ et*

$$\sum_{i=1}^{n} \mathbb{E}[(X_i)_+^q] \leq \frac{q!}{2} v c^{q-2}$$

pour $q \geq 3$, avec $x_+ = \max(0, x)$. On a alors

$$\mathbb{P}\Big[\bar{X}_n - \mathbb{E}[\bar{X}_n] \geq x\Big] \leq \exp \frac{-n^2 x^2}{2(v + cnx)}$$

pour tout $x > 0$.

Précisons enfin qu'il existe de nombreuses autres inégalités de concentration, avec des versions pour les martingales par exemple [Shorack and Wellner, 2009]. Pour plus de détails, nous renvoyons le lecteur au livre Massart [2007].

A.2 Quelques détails supplémentaires

Nous donnons ici quelques détails supplémentaires non indispensables à une première lecture, mais jugés utiles pour approfondir certains aspects rapidemment évoqués.

A.2.1 Structuration des données des Chapitres 2 et 3

Les données utilisées dans les Chapitres 2 et 3 proviennent donc de l'entrepôt de données utilisant une plateforme I2B2 de l'HEGP, I2B2 pour "Informatics for Integrating Biology and the Bedside". Cet entrepôt regroupe l'ensemble des informations contenues dans le système d'information hospitalier en une seule base pauci-relationnelle permettant de faire facilement des extractions utilisant des critères complexes sur plusieurs sources de données. L'objectif de cet entrepôt est de générer de nouvelles hypothèses à partir de stratégies de fouille de données, de réaliser des études épidémiologiques, des études sur les services en santé et de faciliter la recherche clinique en identifiant rapidement les patients éligibles. La Figure A.1 renseigne sur l'alimentation de cet entrepôt à partir du système d'information hospitalier, et la Figure A.2 fournit le diagramme de classes de l'entrepôt.

Les données des patients disponibles dans l'entrepôt sont multi-dimensionnelles, évoluent dans le temps et à granularité sémantique fine, ce qui rend leur exploitation directe difficile. Une autre particularité réside dans le fait que la trajectoire de chaque patient au sein de l'hôpital est unique : si on prend deux patients quelconques, il auront nécessairement subi un certain nombre de tests biologiques différents ; et

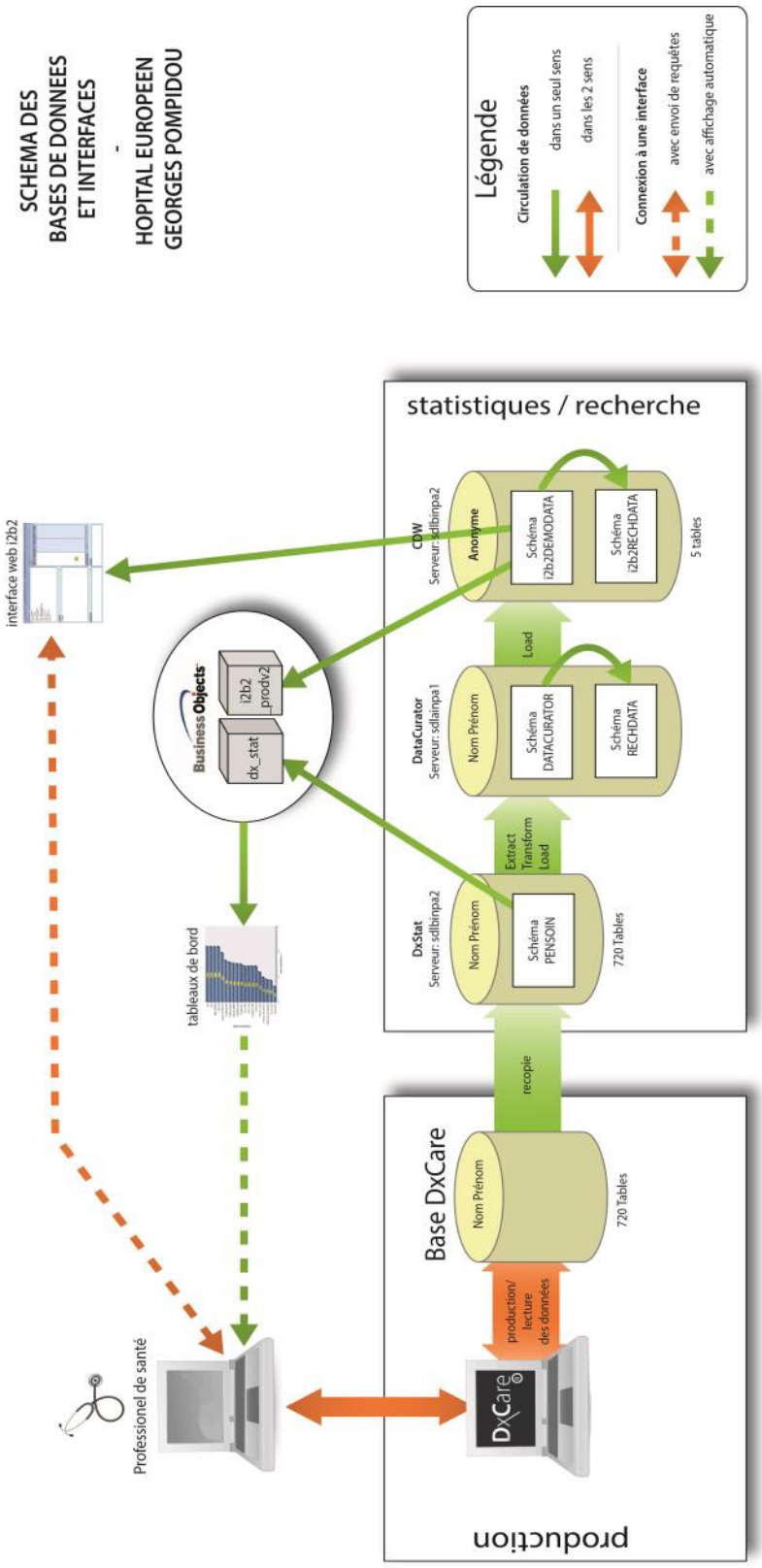


FIGURE A.1 Organisation des données à l'HEGP.

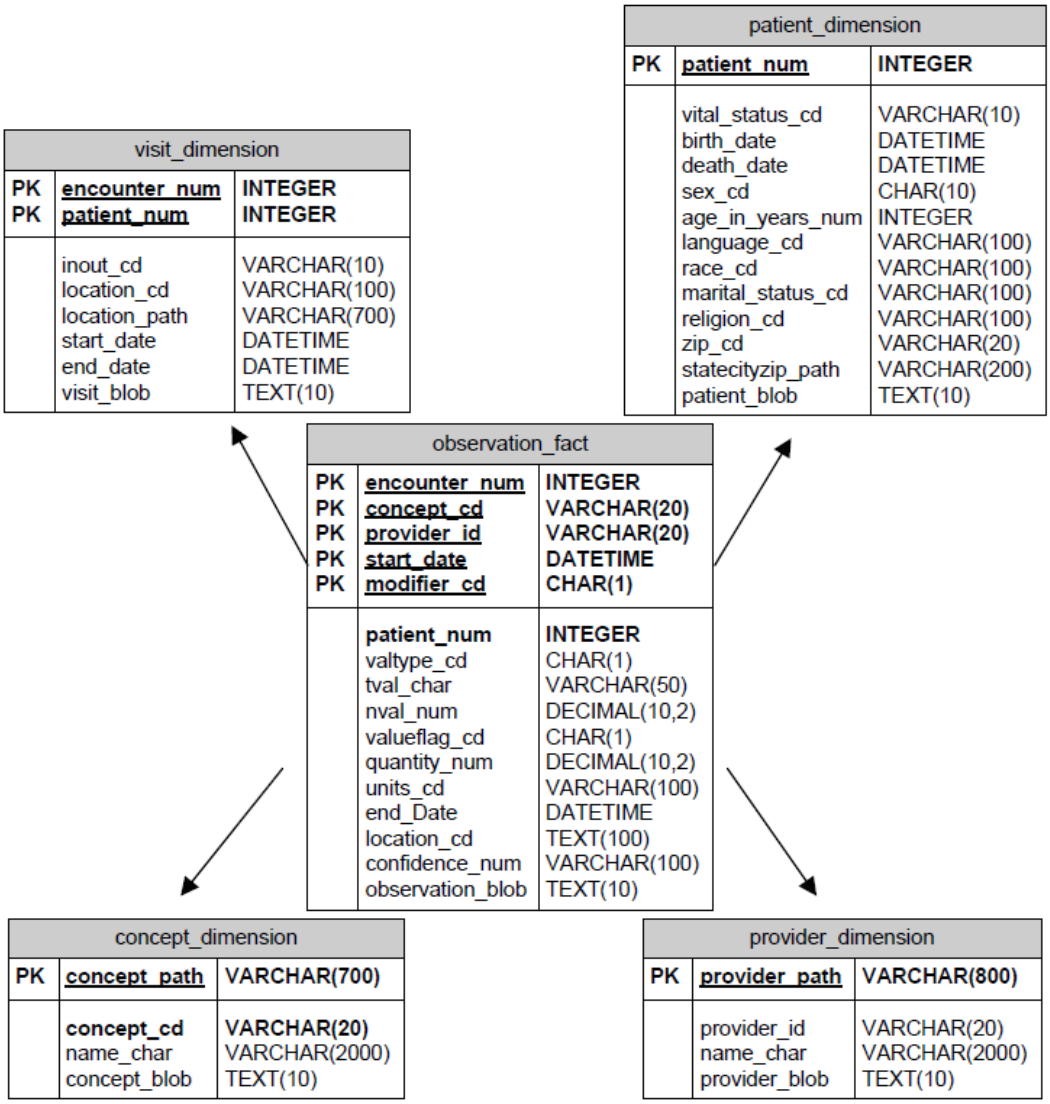


FIGURE A.2 Diagramme de classes de l'entrepôt I2B2.

pour les concepts enregistrés identiques, les instants d'enregistrement des variables considérées ne seront pas les mêmes, ni leur nombre qui dépendra de l'état du patient et de la durée de son séjour. Cela implique une quantité énorme de données manquantes si on considère une structuration classique, où on souhaite créer des covariables renseignées pour chacun des exemples de notre ensemble d'apprentissage.

Dans ce projet portant sur les crises vaso-occlusives chez les patients drépanocytaires, de nombreuses extractions ont été nécessaires pour aboutir à un jeu de données final comportant l'ensemble des informations désirées. Pour être plus clair, les données ont été organisées selon trois grandes catégories : les caractéristiques des patients, les données biologiques et enfin les données concernant les paramètres vitaux. Une description précise des variables extraites est donnée dans la Section 3.A. L'ensemble des données concernent la cohorte de drépanocytaires de l'HEGP entre 2009 et 2015 (avant 2009, les traitements et les protocoles étaient différents).

Les choix dans la récupération de toutes ces données ont été fait avec l'aide des cliniciens spécialistes. De nombreuses difficultés sont apparues après les premières extractions. Pour n'en citer que quelques unes, l'information concernant les paramètres vitaux a par exemple dû être regroupée car présente dans différentes bases : soit codée selon des concepts de paramètres vitaux, soit codée dans une base appelée "pancarte" et correspondant à des questionnaires remplis par les médecins ou infirmiers. Certaines variables n'étaient alors présentes que dans l'une ou l'autre des bases "pancarte" ou "paramètres vitaux" et d'autres dans les deux à la fois, à des intervalles de temps identiques ou non et avec des unités identiques ou non. Un autre exemple serait celui des tris nécessaires pour regrouper les variables biologiques reflétant le même concept mais codées différemment, correspondant par exemple à des quantités physiques identiques mais provenant de tests biologiques différents (l'un sanguin et l'autre urinaire, etc.).

Finalement, les données brutes après la première phase de tri et de nettoyage peuvent se résumer par la Figure A.3 qui renseigne aussi sur le poids du jeu de données brutes final.

Le fichier sejour_drepano.csv contient les informations de base des séjours comme les dates d'entrée et de sortie de l'hôpital, et correspond à la table Visit_drepano dans le diagramme des classes de la Figure A.4. Le fichier demography.csv correspond quant à lui aux données dites "basiques" précédemment évoquées. Les fichiers parametres_vitaux.csv et pancarte.csv contenant le même type d'information ont été réorganisés dans une table unique appelée Vital_parameters dans la Figure A.4.

Après ce premier travail d'extraction et de nettoyage des données, la création d'une structure organisant au mieux les informations par patient et par séjour était

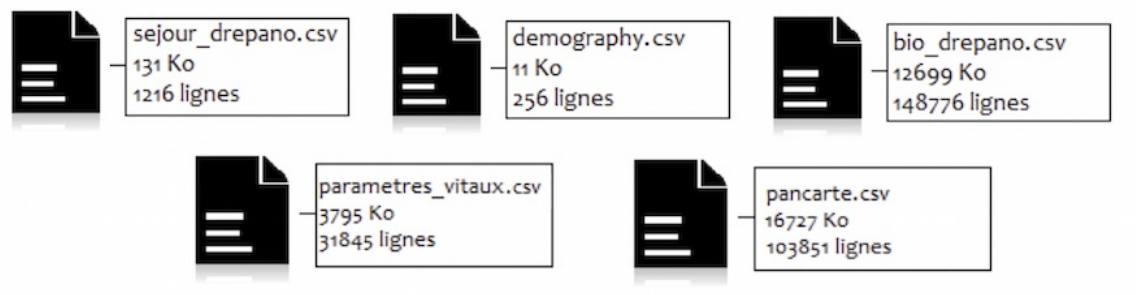


FIGURE A.3 Illustration des fichiers initiaux d'extraction des données brutes.

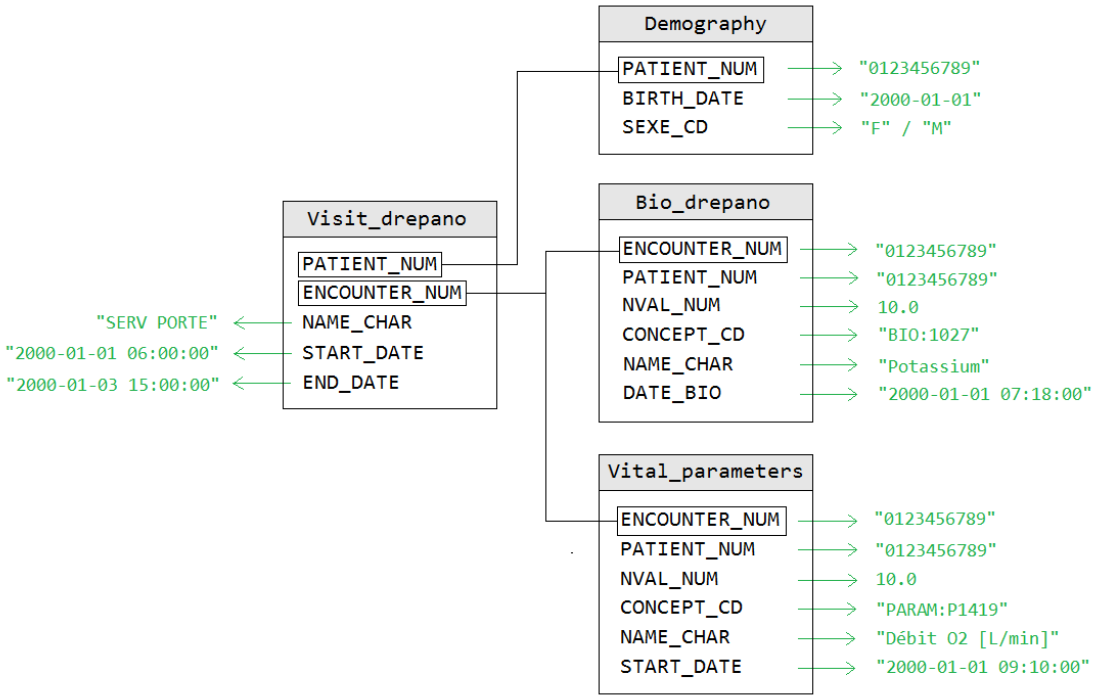


FIGURE A.4 Diagramme des classes après le premier nettoyage avec en vert des exemples représentatifs pour chacun des attributs.

indispensable avant de pouvoir utiliser et interroger les données de façon efficace. Les données ont alors été organisées dans un fichier JSON dont la structure est explicitée sur la Figure A.5. Toutes les informations sont ainsi triées par patient puis par séjour. L'avantage du format JSON est qu'il est facilement compréhensible, la syntaxe n'utilise que quelques marques de ponctuation et il ne dépend d'aucun langage. Comme ce format est très ouvert, il est pris en charge par de nombreux langages et permet de stocker des données de différents types : chaînes de caractères,

nombres, tableaux, objets, ou encore booléens ; donc il est tout à fait adapté à nos données. Sa structure en arborescence et sa syntaxe simple lui permet de rester très “léger” et efficace.

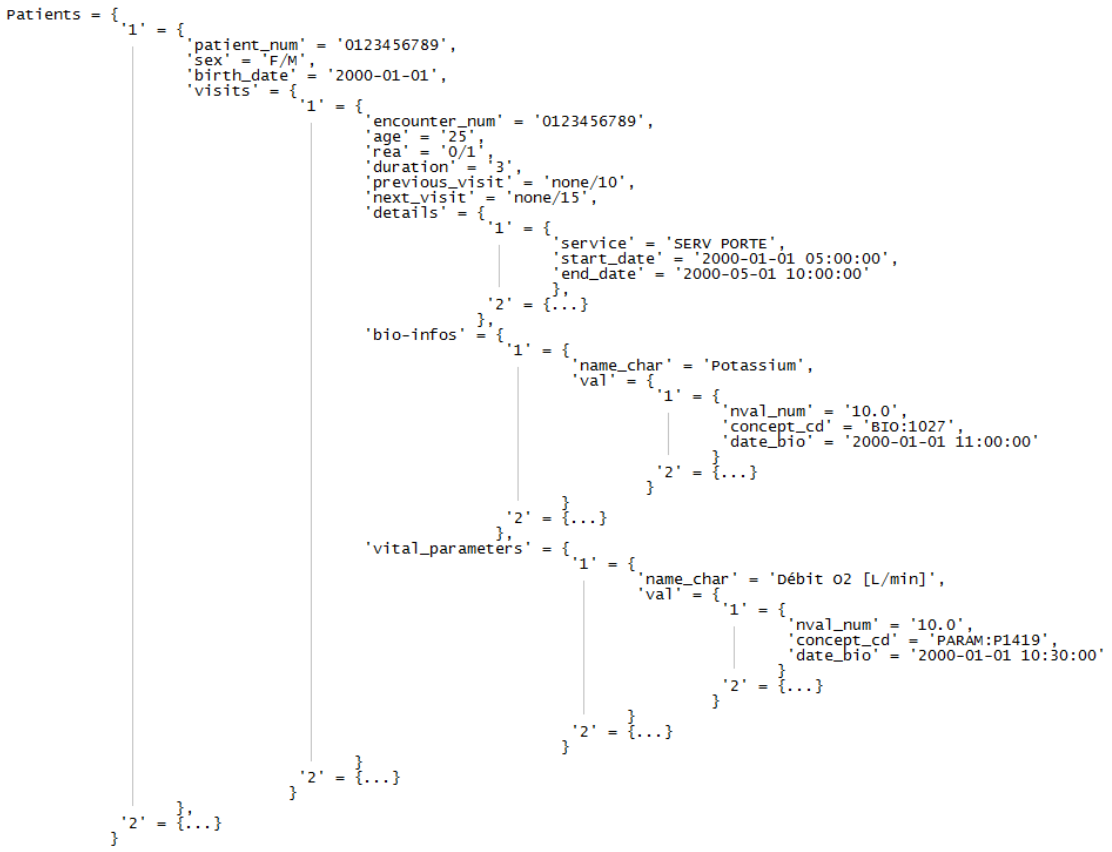


FIGURE A.5 Structure du fichier JSON.

A.2.2 Les données du TCGA

Dans les Chapitres 4 et 6, nous avons utilisé les données issues du TCGA pour le cas d’usage. Quelques explications s’imposent pour mieux comprendre ces données et leur richesse, ainsi que pour saluer le projet TCGA et les avancées scientifiques qu’il rend possible.

La génomique du cancer. L’ADN contient toute l’information génétique, appelée génome, permettant le développement, le fonctionnement et la reproduction des

êtres vivants. Dans les cellules, l'ADN est organisé en structures appelées chromosomes, ce qui est illustré par la Figure A.6. Ces chromosomes sont composés de six milliards de lettres – avec un alphabet de quatre lettres : A, C, G et T – et participent notamment à la régulation de l'expression génétique en déterminant quelles parties de l'ADN doivent être transcrites en ARN. L'ARN est une molécule biologique très proche chimiquement de l'ADN qui est en général synthétisée dans les cellules à partir de l'ADN dont il est une copie. Les cellules utilisent ensuite l'ARN comme un support intermédiaire des gènes pour synthétiser les protéines dont elles ont besoin. La génomique est alors l'étude de la séquence des lettres de l'ADN, notamment pour comprendre les informations transmises aux cellules.

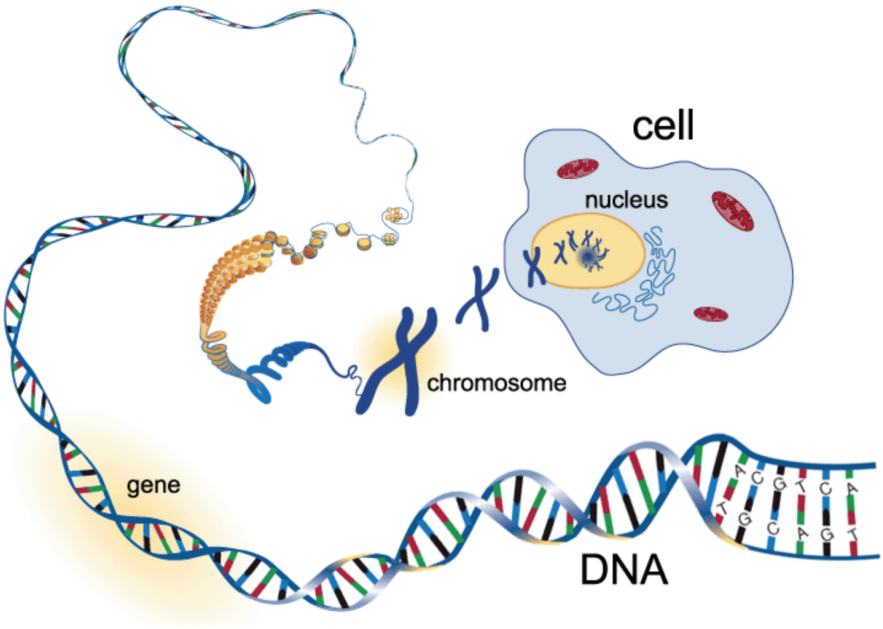


FIGURE A.6 Illustration de la présence d'ADN dans chaque cellule. [1]

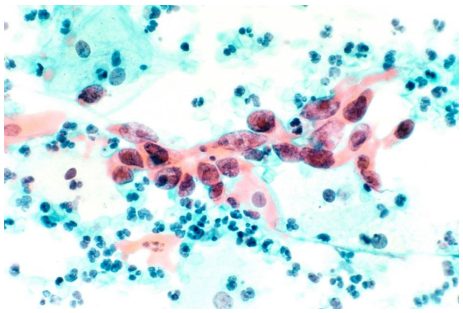
Dans les cellules cancéreuses, une altération génétique (c'est-à-dire un changement de lettre, provoquée par exemple par mutations de l'ADN) peut amener la cellule à fabriquer une protéine qui ne permet pas à la cellule de fonctionner comme elle le devrait, ce qui est illustré par la Figure A.7. Ces protéines peuvent entraîner une croissance incontrôlable des cellules et une malignité pouvant endommager les cellules voisines. Les altérations génétiques peuvent être héritées des parents, causées par des facteurs environnementaux, ou se produire lors de processus naturels tels que la division cellulaire. Les changements qui s'accumulent au cours de la vie

1. Source de la Figure A.6 : https://www.genome.gov/.

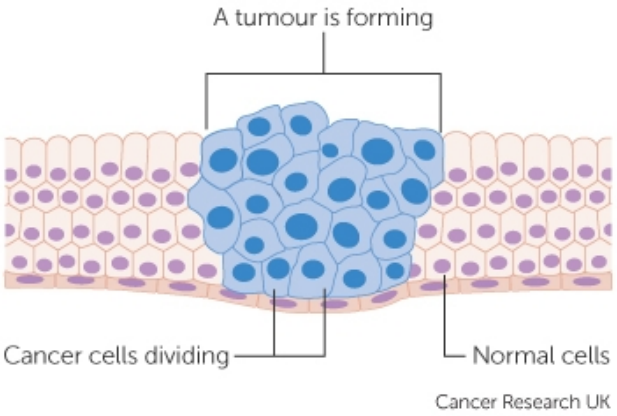
sont appelés changements acquis ou somatiques et représentent 90 à 95% de tous les cas de cancer.

En étudiant le génome du cancer, l'idée est d'essayer de découvrir quels changements de lettre sont à l'origine du cancer. Le génome d'une cellule cancéreuse peut également être utilisé pour distinguer un type de cancer d'un autre. Dans certains cas, l'étude du génome d'un cancer peut aider à identifier un sous-type de cancer, un exemple étant le cancer du sein HER2+.

Comprendre le génome du cancer peut donc également contribuer à la médecine de précision en définissant les types de cancer et les sous-types en fonction de leur génétique. L'idée est alors d'être capable de classifier les patients en sous-groupes disctincts – ce qui est l'objet du Chapitre 4 – pour tenter de fournir aux patients un diagnostic plus précis, et donc une stratégie de traitement personnalisée.



(a) Cellules cancéreuses des cervicales.



(b) Apparition d'une tumeur.

FIGURE A.7 Illustration de cellules cancéreuses entourées de cellules saines. [2]

En séquençant l'ADN et l'ARN des cellules cancéreuses, on peut mesurer l'activité des gènes codés dans l'ADN afin de comprendre quelles protéines sont anormalement actives ou rendues silencieuses dans les cellules cancéreuses, contribuant ainsi à leur croissance incontrôlée.

Ainsi, réunir de vastes ensembles de données génomiques et les partager permet d'identifier les changements génétiques sous-jacents au cancer, déterminer leur rôle

2. Source de la Figure A.7a : http://www.abc.net.au/news/2018-03-06/cervical-cancer-cells/9515806.
Source de la Figure A.7b : https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts/cancer-cells.

dans le développement des tumeurs et exploiter ces résultats pour lutter contre le cancer. C'est l'objet du projet TCGA.

The Cancer Genome Atlas (TCGA). Il s'agit d'un projet lancé en 2005, supervisé par le Centre de Génomique du Cancer de l'Institut National du Cancer et l'Institut National de Recherche sur le Génome Humain, dont le but est de cataloguer les mutations génétiques responsables du cancer en utilisant le séquençage génomique et la bio-informatique. Le portail fournit des données cliniques, des caractérisations génomiques et des analyses de séquences associées aux tumeurs de plus de 33 types de cancers chez l'humain, comme l'illustre la Figure A.8. Ces données représentent plus de 2 petabytes qui sont accessibles en libre accès [3] à la communité scientifique dans le but d'améliorer la prévention, le diagnostic et le traitement du cancer, grâce par exemple à une meilleure compréhension de la base génétique du cancer.

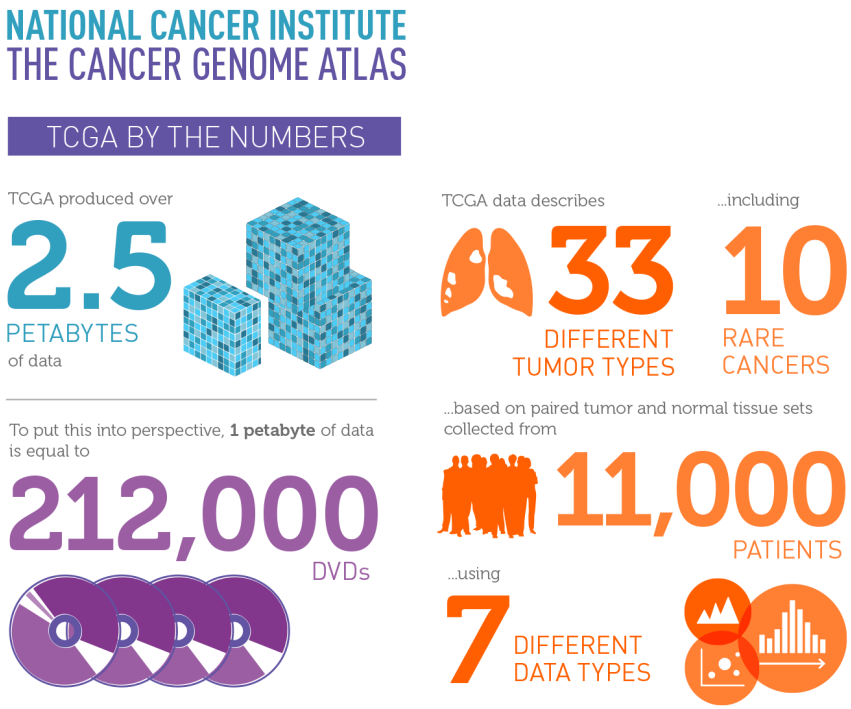


FIGURE A.8 Le TCGA en quelques chiffres. [4]

En particulier, les données que nous avons utilisées dans les Chapitres 4 et 6 sont des données d'expression génétique de l'ARN, c'est-à-dire des covariables à valeurs

3. Les données du TCGA sont accessibles à l'adresse http://cancergenome.nih.gov.
4. Source de la Figure A.8 : https://cancergenome.nih.gov/abouttcga.

réelles correspondant à l'expressions de $p = 20531$ gènes, concernent les trois cancers suivant : le cancer du sein "breast invasive carcinoma" (BRCA) avec un échantillon de $n = 1211$ patients, le cancer du cerveau "glioblastoma multiforme" (GBM) avec un échantillon de $n = 168$ patients et le cancer du rein "kidney renal clear cell carcinoma" (KIRC) avec un échantillon de $n = 605$ patients. Précisons que nous sommes dans un cas typique de grande dimension où $p \gg n$. La Figure A.9 donne alors une visualisation de ce type de données.

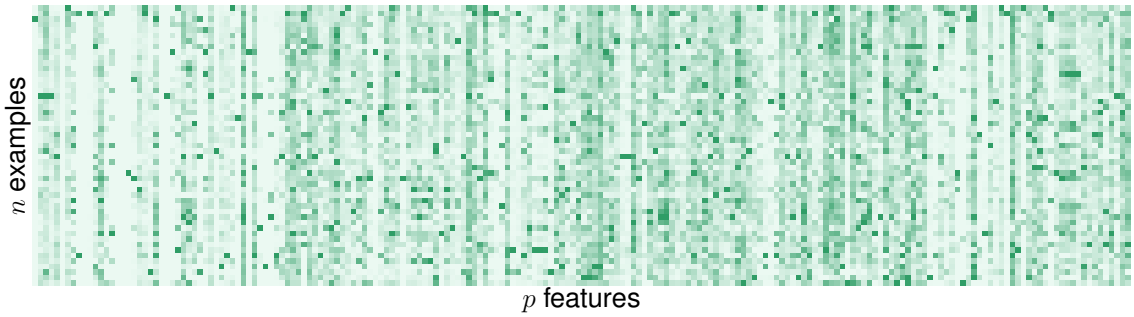


FIGURE A.9 Visualisation d'un sous échantillon des données relatives au cancer BRCA, en prenant ici les $n = 50$ premiers patients et les $p = 200$ premiers gènes, où les valeurs d'expression sont normalisées dans le segment $[0, 1]$. Une valeur de 0 est représentée par un carré de couleur blanche, couleur qui tend vers le vert foncé à mesure que la valeur est proche de 1.

A.2.3 Les métriques en pratique

Nous donnons ici quelques détails sur l'utilisation pratique des deux principales métriques d'analyse de survie utilisées dans ce manuscrit.

AUC(t). Pour un seuil c donné, le taux cumulé de vrais positifs (TPR) et le taux cumulé de faux positifs (FPR) sont deux fonctions du temps respectivement définies par

$$TPR^{\boldsymbol{C}}(c, t) = \mathbb{P}[M > c | T \leq t]$$

et

$$FPR^{\boldsymbol{D}}(c, t) = \mathbb{P}[M > c | T > t],$$

où M est le marqueur étudié. Puis, en utilisant la définition suivante pour l'AUC cumulée dynamique ainsi que le théorème de Bayes, on obtient

$$\begin{aligned} AUC^{\boldsymbol{C},\boldsymbol{D}}(t) &= \int_{-\infty}^{\infty} TPR^{\boldsymbol{C}}(c,t) \left| \frac{\partial FPR^{\boldsymbol{D}}(c,t)}{\partial c} \right| \mathrm{d}c \\ &= \int_{-\infty}^{\infty} \int_{c}^{\infty} \frac{\mathbb{P}[T \leq t | M = m]\mathbb{P}[T > t | M = c]}{\mathbb{P}[T \leq t]\mathbb{P}[T > t]} h(m)h(c)\mathrm{d}m\mathrm{d}c. \end{aligned}$$

En remarquant que

$$\mathbb{P}[T > t] = \int_{-\infty}^{\infty} \mathbb{P}[T > t | M = m]h(m)\mathrm{d}m,$$

un estimateur $\hat{S}_n(t|m)$ de la fonction de survie conditionnelle $\mathbb{P}[T > t|M = m]$ mène à l'estimateur suivant

$$\hat{AUC}^{\boldsymbol{C},\boldsymbol{D}}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{S}_n(t|M_j)[1 - \hat{S}_n(t|M_i)]\mathbb{1}_{\{M_i > M_j\}}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{S}_n(t|M_j)[1 - \hat{S}_n(t|M_i)]}$$

On utilise un estimateur de Kaplan-Meier de la fonction de survie conditionnelle $\mathbb{P}[T > t|M = m]$, comme cela est proposé dans Blanche et al. [2013], ce qui est déjà implémenté dans le package `timeROC` de `R`.

C-Index. La mesure communément utilisée pour palier au problème de la dépendance temporelle de la métrique $AUC(t)$ est le C-index [Harrell et al., 1996] défini par

$$\mathcal{C} = \mathbb{P}[M_i > M_j | T_i < T_j],$$

avec $i \neq j$ deux individus indépendants. Dans ce manuscrit, T est sujet à une censure à droite, on considère alors une version modifiée du C-index, à savoir $\mathcal{C}_\tau$ [Heagerty and Zheng, 2005] défini par

$$\mathcal{C}_\tau = \mathbb{P}[M_i > M_j | Z_i < Z_j, Z_i < \tau],$$

avec τ la durée de l'étude. $\mathcal{C}_\tau$ peut être approximé par l'estimateur non-paramétrique, consistant et non biaisé suivant

$$\hat{\mathcal{C}}_\tau = \frac{\sum_{i,j=1}^{n} \delta_i \{\hat{G}(Z_i)\}^{-2} \mathbb{1}_{\{Z_i < Z_j, Z_i < \tau\}} \mathbb{1}_{\{M_i > M_j\}}}{\sum_{i,j=1}^{n} \delta_i \{\hat{G}(Z_i)\}^{-2} \mathbb{1}_{\{Z_i < Z_j, Z_i < \tau\}}},$$

avec $\hat{G}$ l'estimateur de Kaplan-Meier de la distribution de la censure $G(t) = \mathbb{P}[C > t]$, ce qui mène à un estimateur non-paramétrique et consistant de $\mathcal{C}_\tau$ [Uno et al., 2011], implémenté dans le package `survival` de `R`. De plus, on peut obtenir des intervales de confiance sur $\hat{\mathcal{C}}_\tau$ puisque $\sqrt{n}(\hat{\mathcal{C}}_\tau - \mathcal{C}_\tau)$ suit asymptotiquement une loi normale centrée.

A.2.4 Choix du niveau de censure en simulation du C-mix

Nous donnons ici le détail du calcul de la Section 4.D. La question est de déterminer α_c pour un taux de censure r_c souhaité, avec les paramètres α_0, α_1 et π_0 fixés. On a alors

$$
\begin{aligned}
1 - r_c &= \mathbb{E}[\Delta] \\
&= \mathbb{P}[C - T \geq 0] \\
&= \sum_{k=0}^{+\infty} \mathbb{P}[C - T = k] \\
&= \sum_{k=0}^{+\infty} \mathbb{P}[\bigcup_{j=1}^{\infty} \{T = j, C = j + k\}] \\
&= \sum_{k=0}^{+\infty} \sum_{j=1}^{+\infty} \mathbb{P}[T = j]\mathbb{P}[C = j + k] \\
&= \sum_{k=0}^{+\infty} \sum_{j=1}^{+\infty} \Big[\mathbb{P}[T = j|Z = 0]\mathbb{P}[Z = 0] + \mathbb{P}[T = j|Z = 1]\mathbb{P}[Z = 1]\Big]\mathbb{P}[C = j + k] \\
&= \sum_{k=0}^{+\infty} \sum_{j=1}^{+\infty} \Big[\alpha_0(1 - \alpha_0)^{j-1}\pi_0 + \alpha_1(1 - \alpha_1)^{j-1}(1 - \pi_0)\Big]\alpha_c(1 - \alpha_c)^{j+k-1} \\
&= \alpha_0\alpha_c\pi_0 \sum_{k=0}^{+\infty}(1 - \alpha_c)^k \sum_{j=0}^{+\infty} \Big[(1 - \alpha_0)(1 - \alpha_c)\Big]^j \\
&\quad + \alpha_1\alpha_c(1 - \pi_0) \sum_{k=0}^{+\infty}(1 - \alpha_c)^k \sum_{j=0}^{+\infty} \Big[(1 - \alpha_1)(1 - \alpha_c)\Big]^j \\
&= \frac{\alpha_0\pi_0\Big[1 - (1 - \alpha_1)(1 - \alpha_c)\Big] + \alpha_1(1 - \pi_0)\Big[1 - (1 - \alpha_0)(1 - \alpha_c)\Big]}{\Big[1 - (1 - \alpha_0)(1 - \alpha_c)\Big]\Big[1 - (1 - \alpha_1)(1 - \alpha_c)\Big]}
\end{aligned}
$$

Et nous retrouvons bien le résultat obtenu à la Section 4.D.

Table des figures

Liste des tableaux

Bibliographie

Les 131 centres de référencebanque nationale de données maladies rares. http://www.bndmr.fr/le-projet/nos-partenaires/les-131-centres-de-reference/. Accessed : 2014-09-30. (Cited on pages 50 and 78.)

O. Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978. (Cited on pages 24 and 212.)

O. Aalen. A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pages 1–25. Springer, 1980. (Cited on page 17.)

O. Aalen, O. Borgan, and H. Gjessing. *Survival and event history analysis : a process point of view*. Springer Science & Business Media, 2008. (Cited on page 244.)

A. Agresti. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, 2015. (Cited on page 148.)

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics*, pages 610–624. Springer, 1992. (Cited on page 5.)

C. M. Alaíz, A. Barbero, and J. R. Dorronsoro. Group fused lasso. In *International Conference on Artificial Neural Networks*, pages 66–73. Springer, 2013. (Cited on pages 9 and 233.)

M. Z. Alaya, S. Gaïffas, and A. Guilloux. Learning the intensity of time events with change-points. *Information Theory, IEEE Transactions on*, 61(9) :5148–5171, 2015. (Cited on page 162.)

M. Z. Alaya, S. Bussy, S. Gaïffas, and A. Guilloux. Binarsity : a penalization for one-hot encoded features. *preprint*, 2017. (Cited on pages 183, 191, 205, and 225.)

A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, et al. Distinct types of diffuse large b-cell

lymphoma identified by gene expression profiling. *Nature*, 403(6769) :503–511, 2000. (Cited on page 101.)

D. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *JNCI : Journal of the National Cancer Institute*, 86(11) :829–835, 1994. (Cited on page 195.)

P. K. Andersen and R. D. Gill. Cox's regression model for counting processes : a large sample study. *The annals of statistics*, pages 1100–1120, 1982. (Cited on page 24.)

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. ISBN 0-387-97872-0. (Cited on pages 17, 182, 244, and 248.)

G. Andrew and J. Gao. Scalable training of l1-regularized log-linear models. In *International Conference on Machine Learning*, pages 33–40. ACM, 2007. (Cited on page 107.)

E.-R. Andrinopoulou, D. Rizopoulos, J. J. Takkenberg, and E. Lesaffre. Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in medicine*, 33(18) :3167–3178, 2014. (Cited on page 232.)

E.-R. Andrinopoulou, K. Nasserinejad, R. Szczesniak, and D. Rizopoulos. Integrating latent classes in the bayesian shared parameter joint model of longitudinal and survival outcomes. *arXiv preprint arXiv :1802.10015*, 2018. (Cited on page 232.)

C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974. (Cited on page 234.)

A. Antonov. Bioprofiling. de : analytical web portal for high-throughput cell biology. *Nucleic acids research*, 39(suppl_2) :W323–W327, 2011. (Cited on page 207.)

A. Antonov, M. Krestyaninova, R. Knight, I. Rodchenkov, G. Melino, and N. Barlev. Ppisurv : a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome. *Oncogene*, 33(13) :1621, 2014. (Cited on page 207.)

E. Arjas and D. Gasbarra. Bayesian inference of survival probabilities, under stochastic ordering constraints. *Journal of the American Statistical Association*, 91 (435) :1101–1109, 1996. (Cited on page 232.)

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4 :384–414, 2010. (Cited on pages 152, 168, 213, and 221.)

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1) :1–106, 2012. (Cited on page 150.)

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(Jun) :1179–1225, 2008. (Cited on page 9.)

E. Bacry, M. Bompaire, S. Gaïffas, and S. Poulsen. tick : a Python library for statistical learning, with a particular emphasis on time-dependent modeling. *ArXiv e-prints*, July 2017. (Cited on pages 155 and 191.)

S. Badve, D. Turbin, M. Thorat, A. Morimiya, T. Nielsen, C. Perou, S. Dunn, D. Huntsman, and H. Nakshatri. Foxa1 expression in breast cancer—correlation with luminal subtype a and survival. *Clinical cancer research*, 13(15) :4415–4421, 2007. (Cited on page 207.)

V. Bagdonavicius and M. Nikulin. *Accelerated life models : modeling and statistical analysis*. Chapman and Hall/CRC, 2001. (Cited on page 21.)

P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5, 2014. (Cited on page 155.)

P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson. Parameterized neural networks for high-energy physics. *The European Physical Journal C*, 76 (5) :1–7, Apr 2016. (Cited on page 155.)

S. K. Ballas and E. Smith. Red blood cell changes during the evolution of the sickle cell painful crisis. *Blood*, 79(8) :2154–2163, 1992. (Cited on pages 67 and 68.)

J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993. (Cited on page 101.)

E. M. Bargoma, J. K. Mitsuyoshi, S. K. Larkin, L. A. Styles, F. A. Kuypers, and S. T. Test. Serum c-reactive protein parallels secretory phospholipase a2 in sickle cell disease patients with vasoocclusive crisis or acute chest syndrome. *Blood*, 105 (8) :3384–3385, 2005. (Cited on pages 67 and 68.)

H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. (Cited on page 150.)

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1) :183–202, 2009. (Cited on page 11.)

D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine*, 8(8) : 816–824, 2002. (Cited on pages 31 and 101.)

E. I. Benchimol, L. Smeeth, A. Guttmann, K. Harron, D. Moher, I. Petersen, H. T. Sørensen, E. von Elm, S. M. Langan, R. W. Committee, et al. The reporting of studies conducted using observational routinely-collected health data (record) statement. *PLoS medicine*, 12(10) :e1001885, 2015. (Cited on page 51.)

R. Bender and U. Grouven. Logistic regression models used in medical research are poorly presented. *BMJ : British Medical Journal*, 313(7057) :628, 1996. (Cited on page 76.)

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995. (Cited on page 233.)

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb) :281–305, 2012. (Cited on page 81.)

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995. (Cited on pages 10 and 126.)

A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24) :13790–13795, 2001. (Cited on pages 31 and 101.)

P. J. Bickel, B. Li, A. B. Tsybakov, S. A. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, and A. van der Vaart. Regularization in statistics. *Test*, 15(2) :271–344, 2006. (Cited on page 5.)

P. J. Bickel, Y. Ritov, A. B. Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4) :1705–1732, 2009. (Cited on pages 7, 14, 15, and 143.)

P. J. Bickel, Y. Ritov, A. B. Tsybakov, et al. Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing strength : theory powering applications–a Festschrift for Lawrence D. Brown*, pages 56–69. Institute of Mathematical Statistics, 2010. (Cited on page 241.)

J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3) :1111, 2013. (Cited on page 241.)

BioProfiling. Hbs1l ppisurv, 2009. URL http://www.bioprofiling.de/cgi-bin/GEO/DRUGSURV/display_GENE_GEO.pl?ID=GSE2034&affy=209314_S_AT&ncbi=10767&geneA=HBS1L. (Cited on page 207.)

L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3) :203–268, 2001. (Cited on page 10.)

L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2) :33–73, 2007. (Cited on page 10.)

J. A. Blackard and D. J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3) :131–151, 1999. (Cited on page 155.)

P. Blanche, A. Latouche, and V. Viallon. Time-dependent auc with right-censored data : A survey. *Risk Assessment and Evaluation of Predictions*, 215 :239–251, 2013. (Cited on pages 114 and 262.)

J. M. Bland and D. G. Altman. The logrank test. *Bmj*, 328(7447) :1073, 2004. (Cited on page 19.)

K. Bleakley and J. P. Vert. The group fused Lasso for multiple change-point detection. June 2011. (Cited on pages 10 and 182.)

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022, 2003. (Cited on page 234.)

M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès. Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3) : 1103, 2015. (Cited on page 233.)

C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni*, pages 13–60, 1935. (Cited on pages 85 and 268.)

W. Boulding, S. W. Glickman, M. P. Manary, K. A. Schulman, and R. Staelin. Relationship between patient satisfaction with inpatient care and hospital readmission within 30 days. *The American journal of managed care*, 17(1) :41–48, 2011. (Cited on page 76.)

A.-L. Boulesteix and C. Strobl. Optimal classifier selection and negative bias in error rate estimation : an empirical study on high-dimensional prediction. *BMC medical research methodology*, 9(1) :85, 2009. (Cited on page 77.)

C. Boutilier, R. Dearden, and M. Goldszmidt. Stochastic dynamic programming with factored representations. *Artificial intelligence*, 121(1-2) :49–107, 2000. (Cited on page 237.)

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. (Cited on pages 125, 162, 165, 213, and 219.)

J. Bradic, J. Fan, and J. Jiang. Regularization for cox's proportional hazards model with np-dimensionality. *Annals of statistics*, 39(6) :3092, 2011. (Cited on page 25.)

A. 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. (Cited on pages 30 and 82.)

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4) :373–384, 1995. (Cited on pages 4 and 6.)

L. Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001. (Cited on pages 29, 38, 77, 81, 144, and 154.)

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984. (Cited on page 144.)

N. E. Breslow. Contribution to the discussion of the paper by dr cox. *Journal of the Royal Statistical Society, Series B*, 34(2) :216–217, 1972. (Cited on pages 24, 95, 111, and 120.)

D. C. Brousseau, P. L. Owens, A. L. Mosso, J. A. Panepinto, and C. A. Steiner. Acute care utilization and rehospitalizations for sickle cell disease. *Jama*, 303 (13) :1288–1294, 2010. (Cited on pages 27, 29, 50, 52, 77, and 80.)

H. Brunk, W. Franck, D. Hanson, and R. Hogg. Maximum likelihood estimation of the distributions of two stochastically ordered random variables. *Journal of the American Statistical Association*, 61(316) :1067–1080, 1966. (Cited on page 232.)

S. Bubeck, G. Stoltz, C. Szepesvári, and R. Munos. Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems*, pages 201–208, 2009. (Cited on page 236.)

J. Budczies, F. Klauschen, B. V. Sinn, B. Győrffy, W. D. Schmitt, S. Darb-Esfahani, and C. Denkert. Cutoff finder : a comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. *PloS one*, 7(12) :e51862, 2012. (Cited on pages 41 and 195.)

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data : methods, theory and applications*. Springer Science & Business Media, 2011. (Cited on pages 8, 143, and 191.)

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation and sparsity via ℓ_1 penalized least squares. In *International Conference on Computational Learning Theory*, pages 379–391. Springer, 2006. (Cited on page 7.)

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.*, 1 :169–194, 2007. (Cited on pages 7, 14, and 143.)

H. F. Bunn. Pathogenesis and treatment of sickle cell disease. *New England Journal of Medicine*, 337(11) :762–769, 1997. (Cited on pages 27, 49, and 77.)

S. Bussy, A. Guilloux, S. Gaïffas, and A.-S. 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. (Cited on pages 29, 77, 82, and 89.)

S. Bussy, R. Veil, V. Looten, A. Burgun, S. Gaïffas, A. Guilloux, R. Ranque, and A. Jannot. Comparison of methods for early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework. *ArXiv e-prints*, 2018. (Cited on page 67.)

R. L. Camp, M. Dolled-Filhart, and D. L. Rimm. X-tile : a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clinical cancer research*, 10(21) :7252–7259, 2004. (Cited on page 182.)

A. A. Canalli, N. Conran, A. Fattori, S. T. Saad, and F. F. Costa. Increased adhesive properties of eosinophils in sickle cell disease. *Experimental hematology*, 32(8) : 728–734, 2004. (Cited on pages 66 and 67.)

E. Candès and M. B. Wakin. An Introduction To Compressive Sampling. *Signal Processing Magazine, IEEE*, 25(2) :21–30, 2008. (Cited on page 143.)

E. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5) :877–905, 2008. (Cited on page 143.)

E. Candès, T. Tao, et al. The dantzig selector : Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6) :2313–2351, 2007. (Cited on page 7.)

E. Canu, M. Boccardi, R. Ghidoni, L. Benussi, S. Duchesne, C. Testa, G. Binetti, and G. B. Frisoni. Hoxa1 a218g polymorphism is associated with smaller cerebellar volume in healthy humans. *Journal of Neuroimaging*, 19(4) :353–358, 2009. (Cited on page 200.)

C. Chang, M. Hsieh, W. Chang, A. Chiang, and J. Chen. Determining the optimal number and location of cutoff points with application to data of cervical cancer. *PloS one*, 12(4) :e0176231, 2017. (Cited on page 195.)

M. C. U. Cheang, S. K. Chia, D. Voduc, D. Gao, S. Leung, J. Snider, M. Watson, S. Davies, P. S. Bernard, J. S. Parker, et al. Ki67 index, her2 status, and prognosis of patients with luminal b breast cancer. *JNCI : Journal of the National Cancer Institute*, 101(10) :736–750, 2009. (Cited on page 183.)

H.-C. Chen, R. L. Kodell, K. F. Cheng, and J. J. Chen. Assessment of performance of survival prediction models for cancer prognosis. *BMC medical research methodology*, 12(1) :102, 2012. (Cited on pages 76, 77, and 89.)

B. Chlebus and S. H. Nguyen. On finding optimal discretizations for two attributes. In L. Polkowski and A. Skowron, editors, *Rough Sets and Current Trends in Computing*, volume 1424 of *Lecture Notes in Computer Science*, pages 537–544. Springer Berlin Heidelberg, 1998. (Cited on page 144.)

H. Cho and P. Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 77(2) :475–507, 2015. (Cited on page 182.)

L. Condat. A Direct Algorithm for 1D Total Variation Denoising. *IEEE Signal Processing Letters*, 20(11) :1054–1057, 2013. (Cited on pages 12, 151, and 205.)

C. Contal and J. O'Quigley. An application of changepoint methods in studying the effect of age on survival in breast cancer. *Computational statistics & data analysis*, 30(3) :253–270, 1999. (Cited on page 182.)

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2) :187–220, 1972. (Cited on pages 22, 24, 29, 32, 33, 77, 82, 102, 110, 183, and 246.)

D. R. Cox. Interaction. *International Statistical Review/Revue Internationale de Statistique*, pages 1–24, 1984. (Cited on page 240.)

M. Csikos, Z. Orosz, G. Bottlik, H. Szöcs, Z. Szalai, Z. Rozgonyi, J. Hársing, E. Török, L. Bruckner-Tuderman, A. Horváth, et al. Dystrophic epidermolysis bullosa complicated by cutaneous squamous cell carcinoma and pulmonary and

renal amyloidosis. *Clinical and experimental dermatology*, 28(2) :163–166, 2003. (Cited on page 207.)

S. A. Curtis, N. Danda, Z. Etzion, H. W. Cohen, and H. H. Billett. Elevated steady state wbc and platelet counts are associated with frequent emergency room use in adults with sickle cell anemia. *PLoS One*, 10(8) :e0133116, 2015. (Cited on page 67.)

J. J. Dai, L. Lieu, and D. Rocke. Dimension reduction for classification with gene expression microarray data. *Statistical applications in genetics and molecular biology*, 5(1), 2006. (Cited on page 77.)

A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1) :552–581, 2017. (Cited on pages 166 and 167.)

A. Dancau, L. Wuth, M. Waschow, F. Holst, A. Krohn, M. Choschzick, L. Terracciano, S. Politis, S. Kurtz, A. Lebeau, et al. Ppfia1 and ccnd1 are frequently coamplified in breast cancer. *Genes, Chromosomes and Cancer*, 49(1) :1–8, 2010. (Cited on page 207.)

D. S. Darbari, Z. Wang, M. Kwak, M. Hildesheim, J. Nichols, D. Allen, C. Seamon, M. Peters-Lawrence, A. Conrey, M. K. Hall, et al. Severe painful vaso-occlusive crises and mortality in a contemporary adult sickle cell anemia cohort study. *PloS one*, 8(11) :e79923, 2013. (Cited on page 49.)

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics : A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11) :1413–1457, 2004. (Cited on page 11.)

R. De Angelis, R. Capocaccia, T. Hakulinen, B. Soderman, and A. Verdecchia. Mixture models for cancer survival analysis : application to population-based data with covariates. *Statistics in medicine*, 18(4) :441–454, 1999. (Cited on pages 32, 102, and 104.)

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977. (Cited on pages 33, 105, and 108.)

L. Diggs. Sickle cell crises : Ward burdick award contribution. *American Journal of Clinical Pathology*, 44(1) :1–19, 1965. (Cited on page 49.)

D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. In *PROC. NATL ACAD. SCI. USA 100 2197–202*, 2002. (Cited on page 143.)

D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on*, 47(7) :2845–2862, 2001. (Cited on page 143.)

D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1) :6–18, 2006. (Cited on page 7.)

J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *Machine Learning Proceedings 1995*, pages 194–202. Elsevier, 1995. (Cited on page 36.)

L. Duchateau and P. Janssen. *The frailty model*. Springer Science & Business Media, 2007. (Cited on page 23.)

S. Dudoit and M. J. Van Der Laan. *Multiple testing procedures with applications to genomics*. Springer Science & Business Media, 2007. (Cited on page 196.)

R. Dykstra, S. Kochar, T. Robertson, et al. Statistical inference for uniform stochastic ordering in several populations. *The Annals of Statistics*, 19(2) :870–888, 1991. (Cited on page 233.)

R. L. Dykstra. Maximum likelihood estimation of the survival functions of stochastically ordered random variables. *Journal of the American Statistical Association*, 77(379) :621–628, 1982. (Cited on page 232.)

R. L. Dykstra and C. J. Feltz. Nonparametric maximum likelihood estimation of survival functions with a general stochastic ordering and its dual. *Biometrika*, 76 (2) :331–341, 1989. (Cited on page 232.)

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2) :407–499, 2004. (Cited on page 7.)

H. El Barmi and H. Mukerjee. Inferences under a stochastic ordering constraint : the k-sample case. *Journal of the American Statistical Association*, 100(469) : 252–261, 2005. (Cited on page 232.)

J.-B. Escudié, A.-S. Jannot, E. Zapletal, S. Cohen, G. Malamut, A. Burgun, and B. Rance. Reviewing 741 patients records in two hours with fastvisu. In *AMIA Annual Symposium Proceedings*, volume 2015, page 553. American Medical Informatics Association, 2015. (Cited on pages 52 and 78.)

J. Fan, Y. Feng, Y. Wu, et al. High-dimensional variable selection for cox's proportional hazards model. In *Borrowing Strength : Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 70–86. Institute of Mathematical Statistics, 2010. (Cited on pages 111 and 129.)

D. Faraggi and R. Simon. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Statistics in medicine*, 15(20) : 2203–2213, 1996. (Cited on page 182.)

V. T. Farewell. The use of mixture models for the analysis of sureval data with long-term survivors. *Biometrics*, 38(4) :1041–1046, 1982. (Cited on pages 29, 32, 34, 77, 82, 102, and 111.)

Z. Feng, R. Dearden, N. Meuleau, and R. Washington. Dynamic programming for structured continuous markov decision problems. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 154–161. AUAI Press, 2004. (Cited on page 237.)

O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap : safer rules for the lasso. *arXiv preprint arXiv :1505.03410*, 2015. (Cited on page 241.)

M. J. Frei-Jones, J. J. Field, and M. R. DeBaun. Risk factors for hospital readmission within 30 days : a new quality measure for children with sickle cell disease. *Pediatric blood & cancer*, 52(4) :481–485, 2009. (Cited on pages 50, 52, and 80.)

B. Friedman and J. Basu. The rate and cost of hospital readmissions for preventable conditions. *Medical Care Research and Review*, 61(2) :225–240, 2004. (Cited on page 77.)

J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, et al. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2) :302–332, 2007. (Cited on pages 7, 10, and 144.)

J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4) :367–378, 2002. (Cited on pages 29, 38, 77, 81, and 154.)

S. Gaïffas and A. Guilloux. High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics*, 6 :522–546, 2012. (Cited on pages 25 and 219.)

T. S. Garadah, A. A. Jaradat, M. E. AlAlawi, A. B. Hassan, and R. P. Sequeira. Pain frequency, severity and qt dispersion in adult patients with sickle cell anemia : correlation with inflammatory markers. *Journal of blood medicine*, 7 :255, 2016. (Cited on pages 67 and 68.)

S. Garcia, J. Luengo, J. A. Saez, V. Lopez, and F. Herrera. A survey of discretization techniques : Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4) :734–750, 2013. (Cited on pages 144 and 146.)

L. E. Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv :1009.4219*, 2010. (Cited on page 241.)

X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011. (Cited on page 81.)

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *science*, 286(5439) :531–537, 1999. (Cited on page 101.)

P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models : a roughness penalty approach*. Chapman and Hall, London, 1994. (Cited on pages 36 and 147.)

E. Greenshtein, Y. Ritov, et al. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6) :971–988, 2004. (Cited on page 7.)

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar) :1157–1182, 2003. (Cited on page 76.)

Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.*, 105(492) :1480–1493, 2010. (Cited on pages 10, 182, and 195.)

A. Haris, D. Witten, and N. Simon. Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25(4) :981–1004, 2016. (Cited on page 241.)

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. (Cited on pages 33, 114, and 262.)

D. P. Harrington and T. R. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3) :553–566, 1982. (Cited on page 84.)

J. M. Harvey, G. M. Clark, C. K. Osborne, D. C. Allred, et al. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. *Journal of clinical oncology*, 17(5) :1474–1481, 1999. (Cited on page 183.)

T. Hastie and R. Tibshirani. *Generalized additive models*. Wiley Online Library, 1990. (Cited on pages 22, 38, and 154.)

T. Hastie, R. Tibshirani, D. Botstein, and P. Brown. Supervised harvesting of expression trees. *Genome Biology*, 2(1) :research0003–1, 2001a. (Cited on pages 31 and 101.)

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001b. (Cited on page 143.)

D. M. Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1) :1–12, 2004. (Cited on page 77.)

P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1) :92–105, 2005. (Cited on pages 82, 114, 201, and 262.)

P. J. Heagerty, T. Lumley, and M. S. Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2) :337–344, 2000. (Cited on pages 33 and 114.)

M. Hebiri. *Quelques questions de sélection de variables autour de l'estimateur LASSO*. PhD thesis, Université Paris-Diderot-Paris VII, 2009. (Cited on page 7.)

R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4) :465–480, 2000. (Cited on page 231.)

H. Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4) :984–1006, 2010. (Cited on page 10.)

A. E. Hoerl and R. W. Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67, 1970. (Cited on page 9.)

J. Hoey, R. St-Aubin, A. Hu, and C. Boutilier. Spudd : Stochastic planning using decision diagrams. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 279–288. Morgan Kaufmann Publishers Inc., 1999. (Cited on page 237.)

D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013. (Cited on pages 29, 77, and 80.)

J. Huang, T. Sun, Z. Ying, Y. Yu, and C. H. Zhang. Oracle inequalities for the lasso in the cox model. *The Annals of Statistics*, 41(3) :1142–1165, 06 2013. (Cited on pages 25, 222, 223, and 225.)

N. Huang, S. Cheng, X. Mi, Q. Tian, Q. Huang, F. Wang, Z. Xu, Z. Xie, J. Chen, and Y. Cheng. Downregulation of nitrogen permease regulator like-2 activates pdk1-akt1 and contributes to the malignant growth of glioma cells. *Molecular carcinogenesis*, 55(11) :1613–1626, 2016. (Cited on page 207.)

H. Ishwaran, J. S. Rao, et al. Spike and slab variable selection : frequentist and bayesian strategies. *The Annals of Statistics*, 33(2) :730–773, 2005. (Cited on page 232.)

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008. (Cited on pages 183 and 202.)

S. Ivanoff, F. Picard, and V. Rivoirard. Adaptive lasso and group-lasso for functional poisson regression. *The Journal of Machine Learning Research*, 17(1) :1903–1948, 2016. (Cited on pages 152 and 191.)

M. A. James, Y. Lu, Y. Liu, H. G. Vikis, and M. You. Rgs17, an overexpressed gene in human lung and prostate cancer, induces tumor cell proliferation through the cyclic amp-pka-creb pathway. *Cancer research*, 69(5) :2108–2116, 2009. (Cited on page 207.)

T. Johnson and C. Guestrin. Blitz : A principled meta-algorithm for scaling sparse optimization. In *International Conference on Machine Learning*, pages 1171–1179, 2015. (Cited on page 241.)

A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Annals of Statistics*, pages 681–712, 2000. (Cited on page 7.)

J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011. (Cited on pages 23 and 34.)

A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms : a study on high-dimensional spaces. *Knowledge and information systems*, 12(1) : 95–116, 2007. (Cited on page 83.)

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282) :457–481, 1958. (Cited on pages 19, 95, and 120.)

M. H. R. Khan and J. E. H. Shaw. Variable selection for survival data with a class of adaptive elastic net techniques. *Statistics and Computing*, 26(3) :725–741, 2016. (Cited on page 131.)

J. P. Klein. Small sample moments of some estimators of the variance of the kaplan-meier and nelson-aalen estimators. *Scandinavian Journal of Statistics*, pages 333–340, 1991. (Cited on page 19.)

J. P. Klein and M. L. Moeschberger. *Survival analysis : techniques for censored and truncated data*. Springer Science & Business Media, 2005. (Cited on pages 18, 33, 104, 109, 183, and 193.)

J. P. Klein and J. Wu. Discretizing a continuous covariate in survival studies. *Handbook of Statistics*, 23 :27–42, 2003. (Cited on page 182.)

D. G. Kleinbaum and M. Klein. *Survival analysis*, volume 3. Springer, 2010. (Cited on page 76.)

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000. (Cited on pages 7 and 143.)

R. P. Kocher and E. Y. Adashi. Hospital readmissions and the affordable care act : paying for coordinated quality care. *Jama*, 306(16) :1794–1795, 2011. (Cited on page 77.)

L. Kocsis and C. Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006. (Cited on page 236.)

R. Kohavi. Scaling up the accuracy of naive-Bayes classifiers : A decision-tree hybrid. In *KDD*, volume 96, pages 202–207, 1996. (Cited on page 155.)

R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995. (Cited on page 81.)

S. Kong and B. Nan. Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Statistica Sinica*, 24(1) :25, 2014. (Cited on page 25.)

S. Krishnan, Y. Setty, S. G. Betal, V. Vijender, K. Rao, C. Dampier, and M. Stuart. Increased levels of the inflammatory biomarker c-reactive protein at baseline are associated with childhood sickle cell vasocclusive crises. *British Journal of Haematology*, 148(5) :797–804, 2010. (Cited on page 67.)

A. Y. Kuk and C.-H. Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3) :531–541, 1992. (Cited on pages 77 and 102.)

P. Kulkarni, T. Shiraishi, K. Rajagopalan, R. Kim, S. M. Mooney, and R. H. Getzenberg. Cancer/testis antigens and urological malignancies. *Nature Reviews Urology*, 9(7) :386, 2012. (Cited on page 207.)

L. Kuo and F. Peng. A mixture-model approach to the analysis of survival data. *Biostatistics-Basel-*, 5 :255–272, 2000. (Cited on pages 32, 33, 102, and 104.)

S. S. Kutateladze. *Fundamentals of functional analysis*, volume 12. Springer Science & Business Media, 2013. (Cited on page 211.)

B. Lausen and M. Schumacher. Maximally selected rank statistics. *Biometrics*, pages 73–85, 1992. (Cited on pages 182 and 196.)

M. LeBlanc and J. Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422) :457–467, 1993. (Cited on page 182.)

E. L. Lehmann. Ordered families of distributions. *The Annals of Mathematical Statistics*, pages 399–419, 1955. (Cited on page 232.)

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer texts in statistics. Springer, New York, 1998. (Cited on page 178.)

H. Li and Y. Luan. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, 21(10) : 2403–2409, 2005. (Cited on pages 183 and 202.)

Q. Li, N. Lin, et al. The bayesian elastic net. *Bayesian Analysis*, 5(1) :151–170, 2010. (Cited on page 232.)

M. Lichman. UCI Machine Learning Repository, 2013. (Cited on pages 38, 154, and 155.)

M. Lim and T. Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3) :627–654, 2015. (Cited on page 241.)

H. Lin, B. W. Turnbull, C. E. McCulloch, and E. H. Slate. Latent class models for joint analysis of longitudinal biomarker and event process data : application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97(457) :53–65, 2002. (Cited on page 231.)

J. Little, J. P. Higgins, J. P. Ioannidis, D. Moher, F. Gagnon, E. Von Elm, M. J. Khoury, B. Cohen, G. Davey-Smith, J. Grimshaw, et al. Strengthening the reporting of genetic association studies (strega) : an extension of the strobe statement. *Human genetics*, 125(2) :131–151, 2009. (Cited on page 76.)

M. A. Little and N. S. Jones. Generalized methods and solvers for noise removal from piecewise constant signals. i. background theory. *Proceedings of the Royal Society A-Mathematical Physical and Engineering Sciences*, 467 :3088–3114, 2011. (Cited on page 10.)

H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization : an enabling technique. *Data Min. Knowl. Discov.*, 6(4) :393–423, 2002. (Cited on pages 143, 144, and 186.)

J. Liu, L. Yuan, and J. Ye. An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–332. ACM, 2010. (Cited on page 10.)

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, pages 30–55, 2004. (Cited on page 144.)

C. R. Mansley, A. Weinstein, and M. L. Littman. Sample-based planning for continuous action markov decision processes. In *ICAPS*, 2011. (Cited on page 237.)

T. Martinussen and T. H. Scheike. Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics*, 36(4) :602–619, 2009. (Cited on page 25.)

P. Massart. Concentration inequalities and model selection. 2007. (Cited on page 252.)

P. Massart and C. Meynet. The lasso as an ℓ_1-ball model selection procedure. *Electronic Journal of Statistics*, 5 :669–687, 2011. (Cited on page 14.)

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. (Cited on pages 9, 38, and 154.)

N. Meinshausen, P. Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3) :1436–1462, 2006. (Cited on page 7.)

N. Meinshausen, B. Yu, et al. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1) :246–270, 2009. (Cited on pages 7 and 14.)

B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10(1) :213, 2009. (Cited on page 81.)

R. T. Mikolajczyk, A. DiSilvesto, and J. Zhang. Evaluation of logistic regression reporting in current obstetrics and gynecology literature. *Obstetrics & Gynecology*, 111(2, Part 1) :413–419, 2008. (Cited on page 76.)

R. Mizutani, N. Imamachi, Y. Suzuki, H. Yoshida, N. Tochigi, T. Oonishi, and N. Akimitsu. Oncofetal protein igf2bp3 facilitates the activity of proto-oncogene protein eif4e through the destabilization of eif4e-bp2 mrna. *Oncogene*, 35(27) : 3495, 2016. (Cited on page 207.)

B. Modell and M. Darlison. Global epidemiology of haemoglobin disorders and derived service indicators. *Bulletin of the World Health Organization*, 86(6) : 480–487, 2008. (Cited on page 49.)

M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, pages 1097–1104, 2009. (Cited on page 13.)

D. S. Moore. Chi-square tests. Technical report, Purdue univ lafayette ind dept of statistics, 1976. (Cited on page 20.)

J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Ser. A Math.*, 255 :2897–2899, 1965. (Cited on page 11.)

S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62 :22–31, 2014. (Cited on page 155.)

M. L. Morvan and J.-P. Vert. Whinter : A working set algorithm for high-dimensional sparse second order interaction models. *arXiv preprint arXiv :1802.05980*, 2018. (Cited on page 241.)

R. J. Motzer, M. Mazumdar, J. Bacik, W. Berg, A. Amsterdam, and J. Ferrara. Survival and prognostic stratification of 670 patients with advanced renal cell carcinoma. *Journal of clinical oncology*, 17(8) :2530–2530, 1999. (Cited on page 182.)

J. W. Moul, L. Sun, J. M. Hotaling, N. J. Fitzsimons, T. J. Polascik, C. N. Robertson, P. Dahm, M. S. Anscher, V. Mouraviev, P. A. Pappas, et al. Age adjusted prostate specific antigen and prostate specific antigen velocity cut points in prostate cancer screening. *The Journal of urology*, 177(2) :499–504, 2007. (Cited on page 182.)

G. S. Mudholkar and G. D. Kollia. Generalized weibull family : a structural analysis. *Communications in statistics-theory and methods*, 23(4) :1149–1171, 1994. (Cited on page 23.)

H. Mukerjee. Estimation of survival functions under uniform stochastic ordering. *Journal of the American Statistical Association*, 91(436) :1684–1689, 1996. (Cited on page 232.)

B. N. Mukherjee and S. S. Maiti. On some properties of positive definite toeplitz matrices and their possible applications. *Linear algebra and its applications*, 102 : 211–240, 1988. (Cited on pages 112 and 192.)

S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. C. Chueh, S. Churchill, and I. Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2) :124–130, 2010. (Cited on page 50.)

K. Nakagawa, S. Suzumura, M. Karasuyama, K. Tsuda, and I. Takeuchi. Safe pattern pruning : An efficient approach for predictive pattern mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1785–1794. ACM, 2016. (Cited on page 241.)

Y. Nardi, A. Rinaldo, et al. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2 :605–633, 2008. (Cited on page 9.)

W. H. Organization. *International statistical classification of diseases and related health problems*, volume 1. World Health Organization, 2004. (Cited on page 78.)

T. Oskarsson, S. Acharyya, X. H. Zhang, S. Vanharanta, S. F. Tavazoie, P. G. Morris, R. J. Downey, K. Manova-Todorova, E. Brogi, and J. Massagué. Breast cancer cells produce tenascin c as a metastatic niche component to colonize the lungs. *Nature medicine*, 17(7) :867–874, 2011. (Cited on page 124.)

L. Pauling, H. A. Itano, S. J. Singer, and I. C. Wells. Sickle cell anemia, a molecular disease. *Science*, 110(2865) :543–548, 1949. (Cited on page 49.)

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011a. (Cited on page 155.)

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn : Machine learning in python. *Journal of machine learning research*, 12(Oct) :2825–2830, 2011b. (Cited on page 81.)

F. B. Piel, A. P. Patil, R. E. Howes, O. A. Nyangiri, P. W. Gething, M. Dewi, W. H. Temperley, T. N. Williams, D. J. Weatherall, and S. 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. (Cited on page 25.)

M. Pimentel, D. A. Clifton, L. Clifton, and L. 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. (Cited on pages 29 and 79.)

J. Pittman, E. Huang, H. Dressman, C.-F. Horng, S. H. Cheng, M.-H. Tsou, C.-M. Chen, A. Bild, E. S. Iversen, A. T. Huang, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22) :8431–8436, 2004. (Cited on page 76.)

O. S. Platt, B. D. Thorington, D. J. Brambilla, P. F. Milner, W. F. Rosse, E. Vichinsky, and T. R. Kinney. Pain in sickle cell disease : rates and risk factors. *New England Journal of Medicine*, 325(1) :11–16, 1991. (Cited on pages 27, 49, and 77.)

O. S. Platt, D. J. Brambilla, W. F. Rosse, P. F. Milner, O. Castro, M. H. Steinberg, and P. P. Klug. Mortality in sickle cell disease–life expectancy and risk factors for early death. *New England Journal of Medicine*, 330(23) :1639–1644, 1994. (Cited on page 49.)

R. Prasad, S. Hasan, O. Castro, E. Perlin, and K. Kim. Long-term outcomes in patients with sickle cell disease and frequent vaso-occlusive crises. *The American journal of the medical sciences*, 325(3) :107–109, 2003. (Cited on page 49.)

C. Proust-Lima, M. Séne, J. M. Taylor, and H. Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data : A review. *Statistical methods in medical research*, 23(1) :74–90, 2014. (Cited on page 231.)

C. Proust-Lima, J.-F. Dartigues, and H. Jacqmin-Gadda. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death : a latent process and latent class approach. *Statistics in medicine*, 35(3) :382–398, 2016. (Cited on page 232.)

P. E. Puddu and A. Menotti. Artificial neural networks versus proportional hazards cox models to predict 45-year all-cause mortality in the italian rural areas of the seven countries study. *BMC medical research methodology*, 12(1) :100, 2012. (Cited on page 81.)

J. R. Quinlan. *C4.5 : Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann, 1 edition, 1993. (Cited on page 144.)

P. Rajaraman, A. Hutchinson, N. Rothman, P. M. Black, H. A. Fine, J. S. Loeffler, R. G. Selker, W. R. Shapiro, M. S. Linet, and P. D. Inskip. Oxidative response

gene polymorphisms and risk of adult brain tumors. *Neuro-oncology*, 10(5) :709–715, 2008. (Cited on page 200.)

F. Rapaport, E. Barillot, and J. P. Vert. Classification of arraycgh data using fused SVM. *Bioinformatics*, 24(13) :i375–i382, 2008. (Cited on page 144.)

C. E. Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000. (Cited on page 234.)

D. C. Rees, A. D. Olujohungbe, N. E. Parker, A. D. Stephens, P. Telfer, and J. Wright. Guidelines for the management of the acute painful crisis in sickle cell disease. *British journal of haematology*, 120(5) :744–752, 2003. (Cited on pages 27 and 77.)

D. C. Rees, T. N. Williams, and M. T. Gladwin. Sickle-cell disease. *The Lancet*, 376(9757) :2018–2031, 2010. (Cited on page 49.)

M. W. Rich, V. Beckham, C. Wittenberg, C. L. Leven, K. E. Freedland, and R. M. Carney. A multidisciplinary intervention to prevent the readmission of elderly patients with congestive heart failure. *New England Journal of Medicine*, 333 (18) :1190–1195, 1995. (Cited on page 76.)

P. Rigollet. Kullback Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2) :639–665, 2012. (Cited on pages 147 and 178.)

A. Rinaldo et al. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B) :2922–2952, 2009. (Cited on page 10.)

D. Rizopoulos. *Joint models for longitudinal and time-to-event data : With applications in R*. Chapman and Hall/CRC, 2012. (Cited on page 231.)

R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970. (Cited on page 220.)

A. L. Rogovik, Y. Li, M. A. Kirby, J. N. Friedman, and R. D. Goldman. Admission and length of stay due to painful vasoocclusive crisis in children. *The American journal of emergency medicine*, 27(7) :797–801, 2009. (Cited on pages 67 and 68.)

J. Rojo and F. J. Samaniego. On estimating a survival curve subject to a uniform stochastic ordering constraint. *Journal of the American Statistical Association*, 88(422) :566–572, 1993. (Cited on page 232.)

A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse

large-b-cell lymphoma. *New England Journal of Medicine*, 346(25) :1937–1947, 2002. (Cited on pages 31 and 101.)

M. Rota, L. Antolini, and M. G. Valsecchi. Optimal cut-point definition in biomarkers : the case of censored failure time outcome. *BMC medical research methodology*, 15(1) :24, 2015. (Cited on page 182.)

K. J. Rothman. Estimation of confidence limits for the cumulative probability of survival in life table analysis. *Journal of chronic diseases*, 31(8) :557–560, 1978. (Cited on page 19.)

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D : nonlinear phenomena*, 60(1-4) :259–268, 1992. (Cited on page 10.)

M. A. Russell. *Mining the Social Web : Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media, 2013. (Cited on page 143.)

D. F. Saldana and Y. Feng. Sis : An r package for sure independence screening in ultrahigh dimensional statistical models. *Journal of Statistical Software*, 2016. (Cited on page 129.)

B. Schölkopf and A. J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT press, 2002. (Cited on pages 29, 38, 77, 80, and 154.)

G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6 (2) :461–464, 1978. (Cited on pages 5 and 6.)

R. Senoussi. Problème d'identification dans le modèle de cox. *Ann. Inst. Henri Poincaré*, 26 :45–64, 1990. (Cited on pages 188 and 249.)

M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, et al. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1) :68–74, 2002. (Cited on pages 31 and 101.)

Y. Shirota, J. Stoehlmacher, J. Brabender, Y. Xiong, H. Uetake, K. D. Danenberg, S. Groshen, D. D. Tsao-Wei, P. V. Danenberg, and H. J. Lenz. Ercc1 and thymidylate synthase mrna levels predict survival for colorectal cancer patients receiving combination oxaliplatin and fluorouracil chemotherapy. *Journal of Clinical Oncology*, 19(23) :4298–4304, 2001. (Cited on page 183.)

G. R. Shorack and J. A. Wellner. *Empirical processes with applications to statistics*, volume 59. Siam, 2009. (Cited on page 252.)

V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10(3) :262–266, 1989. (Cited on page 155.)

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5) :1, 2011. (Cited on pages 32, 82, 102, 111, 183, 206, and 272.)

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2) :231–245, 2013. (Cited on page 9.)

T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19) :10869–10874, 2001. (Cited on page 101.)

M. J. Stuart and R. L. Nagel. Sickle-cell disease. *The Lancet*, 364(9442) :1343–1360, 2004. (Cited on pages 25 and 49.)

G. T. Sueyoshi. Semiparametric proportional hazards estimation of competing risks models with time-varying covariates. *Journal of econometrics*, 51(1-2) :25–58, 1992. (Cited on page 23.)

R. S. Sutton, A. G. Barto, et al. *Reinforcement learning : An introduction*. MIT press, 1998. (Cited on page 236.)

T. M. Therneau and P. M. Grambsch. Multiple events per subject. In *Modeling survival data : extending the Cox Model*, pages 169–229. Springer, 2000. (Cited on page 250.)

T. M. Therneau and P. M. Grambsch. *Modeling survival data : extending the Cox model*. Springer Science & Business Media, 2013. (Cited on pages 23, 244, and 251.)

L. Tian, D. Zucker, and L. Wei. On the cox model with time-varying regression coefficients. *Journal of the American statistical Association*, 100(469) :172–183, 2005. (Cited on page 23.)

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. (Cited on pages 4, 7, 38, 143, and 154.)

R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4) :385–395, 1997. (Cited on page 25.)

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10) :6567–6572, 2002. (Cited on pages 31 and 101.)

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(1) :91–108, 2005. (Cited on pages 10, 143, and 144.)

R. J. Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7 :1456–1490, 2013. (Cited on page 8.)

A. Tikhonov and V. Y. Arsenin. *Methods for solving ill-posed problems*. John Wiley and Sons, Inc, 1977. (Cited on page 4.)

L. Tong, C. Erdmann, M. Daldalian, J. Li, and T. Esposito. Comparison of predictive modeling approaches for 30-day all-cause non-elective readmission risk. *BMC medical research methodology*, 16(1) :26, 2016. (Cited on page 76.)

B. Trombert-Paviot, A. Rector, R. Baud, P. Zanstra, C. Martin, E. van der Haring, L. Clavel, and J. M. Rodrigues. The development of ccam : the new french coding system of clinical procedures. *Health Information Management*, 31(1) :2–11, 2003. (Cited on page 78.)

J. A. Tropp. Greed is good : Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10) :2231–2242, 2004. (Cited on page 6.)

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3) :475–494, 2001. (Cited on pages 126 and 128.)

H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10) :1105–1117, 2011. (Cited on pages 114, 202, and 262.)

G. J. Upton. Fisher's exact test. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 395–402, 1992. (Cited on page 84.)

Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5) :552–556, 2011. (Cited on page 50.)

S. van de Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, 36(2) :614–645, 2008. (Cited on pages 37, 147, 152, and 166.)

S. van de Geer and J. Lederer. *The Lasso, correlated design, and improved oracle inequalities*, volume Volume 9 of *Collections*, pages 303–316. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2013. (Cited on page 166.)

S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.*, 3 :1360–1392, 2009. (Cited on pages 7 and 189.)

L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871) :530–536, 2002. (Cited on pages 31 and 101.)

V. Vapnik. Principles of risk minimization for learning theory. In *NIPS*, pages 831–838, 1991. (Cited on page 3.)

V. Vapnik. *Statistical learning theory. 1998*, volume 3. Wiley, New York, 1998. (Cited on page 3.)

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015. (Cited on page 13.)

G. Verbeke. Linear mixed models for longitudinal data. In *Linear mixed models in practice*, pages 63–153. Springer, 1997. (Cited on page 231.)

E. P. Vichinsky, L. D. Neumayr, A. N. Earles, R. Williams, E. T. Lennette, D. Dean, B. Nickerson, E. Orringer, V. McKie, R. Bellevue, et al. Causes and outcomes of the acute chest syndrome in sickle cell disease. *New England Journal of Medicine*, 342(25) :1855–1865, 2000. (Cited on page 49.)

J. M. Vinson, M. W. Rich, J. C. Sperry, A. S. Shah, and T. McNamara. Early readmission of elderly patients with congestive heart failure. *Journal of the American Geriatrics Society*, 38(12) :1290–1295, 1990. (Cited on page 76.)

A. Virouleau, A. Guilloux, S. Gaïffas, and M. Bogdan. High-dimensional robust regression and outliers detection with slope. *arXiv preprint arXiv :1712.02640*, 2017. (Cited on page 234.)

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5) :2183–2202, 2009. (Cited on page 7.)

L.-J. Wei. The accelerated failure time model : a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15) :1871–1879, 1992. (Cited on pages 111 and 131.)

A. Weinstein and M. L. Littman. Bandit-based planning and learning in continuous-action markov decision processes. In *ICAPS*, 2012. (Cited on page 236.)

P. H. Westfall, S. S. Young, and S. P. Wright. On adjusting p-values for multiplicity. *Biometrics*, 49(3) :941–945, 1993. (Cited on page 196.)

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1 (6) :80–83, 1945. (Cited on page 84.)

C. J. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11 :95–103, 1983. (Cited on page 108.)

J. Wu and S. Coggeshall. *Foundations of Predictive Analytics (Chapman & Hall/-CRC Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 1st edition, 2012. (Cited on pages 143, 144, 183, and 186.)

Z. J. Xiang, H. Xu, and P. J. Ramadge. Learning sparse representations of high dimensional data on large scale dictionaries. In *Advances in neural information processing systems*, pages 900–908, 2011. (Cited on page 241.)

B. Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009. (Cited on pages 29, 77, and 81.)

I. C. Yeh and C. H. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2) :2473–2480, 2009. (Cited on page 155.)

Y. L. Yu. On decomposing the proximal map. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 91–99. 2013. (Cited on page 161.)

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49–67, 2006. (Cited on pages 9 and 233.)

E. Zapletal, N. Rodon, N. Grabar, and P. Degoulet. Methodology of integration of a clinical data warehouse with a clinical information system : the hegp case. In *MedInfo*, pages 193–197, 2010. (Cited on pages 50 and 78.)

C.-H. Zhang, J. Huang, et al. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4) :1567–1594, 2008. (Cited on pages 7 and 14.)

H. H. Zhang and W. Lu. Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3) :691–703, 2007. (Cited on page 25.)

P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7 :2541–2563, 2006. (Cited on pages 7, 14, and 143.)

Y. Zhou, C. Yau, J. W. Gray, K. Chew, S. H. Dairkee, D. H. Moore, U. Eppenberger, S. Eppenberger-Castori, and C. C. Benz. Enhanced nfκb and ap-1 transcriptional activity associated with antiestrogen resistant breast cancer. *BMC cancer*, 7(1) : 1, 2007. (Cited on page 124.)

C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778 : L-bfgs-b : Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4) :550–560, 1997. (Cited on pages 33, 105, and 106.)

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476) :1418–1429, 2006. (Cited on pages 8 and 9.)

H. Zou. A note on path-based variable selection in the penalized proportional hazards model. *Biometrika*, 95(1) :241–247, 2008. (Cited on page 25.)

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67 (2) :301–320, 2005. (Cited on pages 9, 33, 76, 77, 80, 81, and 105.)

H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4) :1733, 2009. (Cited on page 9.)

Simon BUSSY **16/01/2019**

Subject : Introduction of high-dimensional interpretable machine learning models and their applications

Abstract: This dissertation focuses on the introduction of new interpretable machine learning methods in a high-dimensional setting. We developped first the C-mix, a mixture model of censored durations that automatically detects subgroups based on the risk that the event under study occurs early; then the binarsity penalty combining a weighted total variation penalty with a linear constraint per block, that applies on one-hot encoding of continuous features; and finally the binacox model that uses the binarsity penalty within a Cox model to automatically detect cut-points in the continuous features. For each method, theoretical properties are established: algorithm convergence, non-asymptotic oracle inequalities, and comparison studies with state-of-the-art methods are carried out on both simulated and real data. All proposed methods give good results in terms of prediction performances, computing time, as well as interpretability abilities.

Keywords : High-dimensional statistics; survival analysis; machine learning; non-asymptotic oracle inequalitie; counting processes; healthcare applications

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

Résumé : Dans ce manuscrit sont introduites de nouvelles méthodes interprétables de machine learning dans un contexte de grande dimension. Différentes procédures sont alors proposées: d'abord le C-mix, un modèle de mélange de durées qui détecte automatiquement des sous-groupes suivant le risque d'apparition rapide de l'événement temporel étudié; puis la pénalité binarsity, une combinaison entre variation totale pondérée et contrainte linéaire par bloc qui s'applique sur l'encodage "one-hot" de covariables continues ; et enfin la méthode binacox qui applique la pénalité précédente dans un modèle de Cox en tirant notamment parti de sa propriété de détection automatique de seuils dans les covariables continues. Pour chacune d'entre elles, les propriétés théoriques sont étudiées comme la convergence algorithmique ou l'établissement d'inégalités oracles non-asymptotiques, et une étude comparative avec l'état de l'art est menée sur des données simulées et réelles. Toutes les méthodes obtiennent de bons résultats prédictifs ainsi qu'en terme de complexité algorithmique, et chacune dispose d'atouts intéressants sur le plan de l'interprétabilité.

Mots clés : Statistiques en grande dimension; analyse de survie; apprentissage automatique; inégalités oracle non-asymptotiques; processus de comptage; applications en santé