Splines multidimensionnelles pénalisées pour modéliser le taux de survenue d'un événement : application au taux de mortalité en excès et à la survie nette en épidémiologie des maladies chroniques

Mathieu Fauvernier

N° d'ordre NNT : 2019LYSE1129

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale ED 341
Écosystèmes Évolution Modélisation Microbiologie (E2M2)

Spécialité de doctorat : Biostatistiques, Santé Publique

Soutenue publiquement le 24/09/2019, par :
Mathieu FAUVERNIER

Direction : **Nadine BOSSARD** Co-direction : **Laurent REMONTET**

Splines multidimensionnelles pénalisées pour modéliser le taux de survenue d'un événement. Application au taux de mortalité en excès et à la survie nette en épidémiologie des maladies chroniques

---

Devant le jury composé de :

RONDEAU Virginie, DR INSERM, Université de Bordeaux	Rapporteure
SAULEAU Erik-André, PU-PH, Université de Strasbourg	Rapporteur
ABRAHAMOWICZ Michal, PU, Université McGill de Montréal	Rapporteur
MAUCORT-BOULCH Delphine, PU-PH, Université de Lyon	Examinatrice
GIORGI Roch, PU-PH, Université Aix-Marseille	Examinateur
PLANCHET Frédéric, PU, Université de Lyon	Examinateur
MONNEREAU Alain, PH, Institut Bergonié, Bordeaux	Examinateur
BOSSARD Nadine, PH, Hospices Civils de Lyon	Directrice de thèse
ESTÈVE Jacques, PU-PH, Université de Lyon	Invité

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

À mon épouse Stéphanie
À mon fils Gaël

On fait la science avec des faits, comme on fait une maison avec des pierres ;
mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison.

Henri Poincaré - *La Science et l'hypothèse (1908)*

Résumé

L'étude du temps de survenue d'un événement représente un champ très important des statistiques. Lorsque l'événement étudié est le décès, on cherche à décrire la survie des individus ainsi que leur taux de mortalité, c'est-à-dire la « force de mortalité » qui s'applique à un instant donné.
Les patients atteints d'une maladie chronique présentent en général un excès de mortalité par rapport à une population ne présentant pas la maladie en question. En épidémiologie, l'étude du taux de mortalité en excès des patients, et notamment de l'impact des facteurs pronostiques sur celui-ci, représente donc un enjeu majeur de santé publique.

D'un point de vue statistique, la modélisation du taux de mortalité (en excès) implique de prendre en compte les effets potentiellement non-linéaires et dépendants du temps des facteurs pronostiques ainsi que les interactions. Les splines de régression, polynômes par morceaux paramétriques et flexibles, sont des outils particulièrement bien adaptés pour modéliser des effets d'une telle complexité.
Toutefois, la flexibilité des splines de régression comporte un risque de sur-ajustement. Pour éviter ce risque, les splines de régression pénalisées ont été proposées dans le cadre des modèles additifs généralisés (GAM). Leur principe est le suivant : à chaque spline peuvent être associés un ou plusieurs termes de pénalité contrôlés par des paramètres de lissage. Les paramètres de lissage représentent les degrés de pénalisation souhaités. En pratique, ils sont inconnus et doivent être estimés tout comme les paramètres de régression.

Dans le cadre de cette thèse, nous avons développé une méthode permettant de modéliser le taux de mortalité (en excès) à l'aide de splines de régression multidimensionnelles pénalisées. Des splines cubiques restreintes ont été utilisées comme splines unidimensionnelles ou bien comme bases marginales afin de former des splines multidimensionnelles par produits tensoriels. Le processus d'optimisation s'appuie sur deux algorithmes de Newton-Raphson emboîtés. L'estimation des paramètres de lissage est effectuée en optimisant un critère de validation croisée ou bien la vraisemblance marginale des paramètres de lissage par un algorithme de Newton-Raphson dit externe. À paramètres de lissage fixés, les paramètres de régression sont estimés par maximisation de la vraisemblance pénalisée par un algorithme de Newton-Raphson dit interne. Les bonnes propriétés de cette approche en termes de performances statistiques et de stabilité numérique ont ensuite été démontrées par simulation.
La méthode a ensuite été implémentée au sein du package R *survPen*.
Enfin, la méthode a été appliquée sur des données réelles afin de répondre aux deux questions épidémiologiques suivantes : l'impact de la défavorisation sociale sur la mortalité en excès des patients atteints d'un cancer du col de l'utérus et l'impact de l'âge courant sur la mortalité en excès des patients atteints de sclérose en plaques.

Mots clés : Splines pénalisées, survie, taux de mortalité, taux en excès, survie nette, épidémiologie, maladies chroniques

Abstract

Time-to-event analysis is a very important field in statistics. When the event under study is death, the analysis focuses on the probability of survival of the subjects as well as on their mortality hazard, that is, on the "force of mortality" that applies at any given moment.
Patients with a chronic disease usually have an excess mortality compared to a population that does not have the disease. Studying the excess mortality hazard associated with a disease and investigating the impact of prognostic factors on this hazard are important public health issues in epidemiology.

From a statistical point of view, modelling the (excess) mortality hazard involves taking into account potentially non-linear and time-dependent effects of prognostic factors as well as their interactions.
Regression splines (i.e., parametric and flexible piecewise polynomials) are ideal for dealing with such a complexity. They make it possible to build easily non-linear effects and, regarding interactions between continuous variables, make it easy to form a multidimensional spline from two or more marginal one-dimensional splines. However, the flexibility of regression splines presents a risk of overfitting. To avoid this risk, penalized regression splines have been proposed as part of generalized additive models (GAM). Their principle is to associate each spline with one or more penalty terms controlled by smoothing parameters. The smoothing parameters represent the desired degrees of penalization. In practice, these parameters are unknown and have to be estimated just like the regression parameters.

This thesis describes the development of a method to model the (excess) hazard using multidimensional penalized regression splines. Restricted cubic splines were used as one-dimensional splines or marginal bases to form multidimensional splines by tensor products. The optimization process relies on two nested Newton-Raphson algorithms. Smoothing parameter estimation is performed by optimizing a cross-validation criterion or the marginal likelihood of the smoothing parameters with an outer Newton-Raphson algorithm. At fixed smoothing parameters, the regression parameters are estimated by maximizing the penalized likelihood by an inner Newton-Raphson algorithm. The good properties of this approach in terms of statistical performance and numerical stability were then demonstrated through simulation.
The described method was then implemented within the R package *survPen*.
Finally, the method was applied to real data to investigate two epidemiological issues : the impact of social deprivation on the excess mortality in cervical cancer patients and the impact of current age on the excess mortality in multiple sclerosis patients.

Keywords : Penalized splines, survival, mortality hazard, excess hazard, net survival, epidemiology, chronic diseases

Remerciements

Tout d'abord je tiens à remercier Nadine Bossard, ma directrice de thèse, pour ses conseils, sa patience et sa bienveillance mais également pour ses inépuisables ressources en fromages de Savoie.

Je remercie ensuite Laurent Remontet, mon co-directeur de thèse, pour son implication de tous les instants et son intransigeance salvatrice. Je tiens également à saluer son sens du partage, que ce soit au niveau de son amour inconditionnel pour Roger Federer ou de sa passion débordante pour les égouttoirs vides.

Je souhaiterais ajouter un remerciement collectif à mes deux directeurs de thèse pour leur aide et leur patience (qualité qui me fait parfois défaut).

Merci à Laurent Roche qui, par sa contribution essentielle à ce travail, mérite sans aucun doute le titre honorifique de co-co-directeur de thèse. Je le remercie également pour son flegme légendaire face à mes irruptions incessantes dans son bureau et pour son étonnante aptitude à jeûner un jour sur deux.

Merci à Zoé Uhry pour ses précieux conseils, sa capacité d'écoute et sa joie de vivre communicative. Merci également à elle pour sa capacité à subir les irruptions mentionnées ci-dessus et pour ses délicieuses salades à l'avocat.

Merci à Benjamin Riche pour les nombreux échanges constructifs qu'il a suscités durant ces trois années et pour tous ceux qu'il continuera à susciter.

Merci à Joris Giai pour les innombrables délires partagés, pour notre article en cardiologie, pour les MOOC avec Jean-Michel Sature, pour ses T-shirts dont il a le secret, pour les plats de son épouse Nannan et pour le poste d'AHU qu'il me cédera en novembre prochain.

Merci à Emmanuelle Dantony pour son aide statistique et informatique. Merci également à elle pour ses qualités d'arbitre de paris sérieux comme moins sérieux, notamment lorsque j'ai le plaisir de les remporter.

Merci à Fabien Subtil pour sa disponibilité et son assistance.

Je remercie Aurélien Belot pour sa spontanéité rafraichissante et son humour.

Merci à Hadrien Charvat pour sa haute technicité, sa gentillesse et sa générosité.

Merci à Mad-Hélénie Elsensohn pour son aide et ses traits d'humour.

Merci à mes compagnonnes (oui c'est moche mais il me semble qu'il s'agit du terme approprié) de bureau et d'infortune, Rose, Stéphanie et Ceren (≪ Dgérène ≫) pour leur bonne humeur constante. Je remercie tout particulièrement Ceren de m'avoir supporté aussi longtemps malgré mes remarques tant incessantes que désobligeantes sur ses fautes de français.

Merci à Yoann Blangero, camarade de classe en M2 B3S et compagnon d'aventure de doctorat (jusqu'en Australie), pour son aide et son humour. J'en profite pour le remercier pour ses qualités insoupçonnées de rapporteur de poster et de livreur de robot-cuiseur.

Je remercie René Ecochard et Delphine Maucort-Boulch pour leur accueil au sein du service de biostatistique des HCL ainsi que Pascal Roy et Muriel Rabilloud pour leur accueil au sein du laboratoire LBBE.

J'en profite pour adresser un remerciement particulier à Delphine Maucort-Boulch pour son soutien et sa confiance.

Un grand merci à Stéphanie Robert, Mariethé Chaumeil, Michèle Canova et Paola Damaso pour m'avoir grandement simplifié la vie pendant ces trois ans. Un merci tout particulier pour l'hôtel avec piscine à Corte et pour le financement du voyage à Melbourne à l'occasion de l'ISCB 2018.

Merci à Jean Iwaz pour ces relectures approfondies ainsi que ces excellents plats libanais.

Je remercie l'ensemble du service de biostatistique des HCL pour l'ambiance et le cadre de travail dans lequel il m'a été possible de réaliser ce travail.

Je remercie le projet CENSUR, l'ensemble de ses membres et notamment Roch Giorgi, son initiateur, pour cette belle aventure humaine et scientifique.

Merci aux équipes partenaires du service : Santé Publique France, le réseau FRANCIM (avec notamment Pascale Grosclaude et Alain Monnereau), l'unité ANTICIPE de Caen (avec notamment Laure Tron, Guy Launoy et Olivier Dejardin), l'OFSEP, l'Ecole des Hautes Etudes en Santé Publique de Rennes (et notamment à Emmanuelle Leray), le Centre de Recherche en Neurosciences de Lyon (et notamment à Fabien Rollot).

Merci au Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation d'avoir financé ce travail.

Je remercie également l'Agence Nationale de la Recherche, l'Institut National du Cancer, Santé Publique France et l'Institut de Recherche en Santé Publique pour leur soutien financier à des projets parallèles auxquels j'ai eu la chance de participer.

Merci aux membres de mon comité de suivi, Sam Meyer, Fabien Subtil, Frédéric Planchet, Hadrien Charvat, Nadine Bossard, Laurent Remontet, Zoé Uhry et Laurent Roche pour leurs conseils et leur soutien.

Je remercie Virginie Rondeau, Erik-André Sauleau et Michal Abrahamowicz d'avoir accepté d'être les rapporteurs de cette thèse. Je remercie Delphine Maucort-Boulch, Roch Giorgi, Alain Monnereau et Frédéric Planchet d'avoir accepté d'êtres examinateurs.

Merci également à Jacques Estève, pour son écoute, ses conseils avisés et pour avoir accepté de faire partie du jury en tant que membre invité.

Merci aux différents stagiaires pour leur enthousiasme et à tous ceux qui ont fortement contribué à la bonne ambiance du service : Mariéta, Elodie, Astrid, Hadi, Axel, Anne, Magali, Saliou, Alice, Clara, Marie, Maxime et Julien (et d'autres que je n'ai pas cités). Une pensée toute particulière pour les stagiaires de l'année 2019 grâce à qui j'aurai mangé un nombre inqualifiable de gâteaux.

Mille mercis à Aurélie Boldrini pour toutes les thèses auxquelles elle a secrètement apporté son soutien.

Un grand merci à mes amis et à ma famille. Je remercie mon épouse Stéphanie pour son soutien de chaque instant et mon fils Gaël pour avoir enchanté et dynamisé ma dernière année de thèse.

Je remercie enfin l'année 2015 qui aura été témoin de ma réorientation professionnelle puis de ma rencontre avec mon épouse.

Table des matières

Remerciements v

Table des figures **xv**

Liste des tableaux **xvii**

Liste des symboles **xviii**

Introduction générale **1**

I Éléments théoriques nécessaires aux objectifs de la thèse **7**

1 Modélisation de l'effet du temps et des covariables par des splines **9**

1.1 Régression polynomiale 9

1.2 Splines de régression 9

1.2.1 Base des puissances tronquées 12

1.2.2 B-splines 12

1.2.3 Splines cubiques restreintes 13

1.3 Nombre et position des nœuds 17

2 Modèles de survie **19**

2.1 Généralités 19

2.2 Formalisme et indicateurs 19

2.3 Estimateurs non-paramétriques 20

2.3.1 Estimateur de Kaplan-Meier 20

2.3.2 Estimateur de Nelson-Aalen 22

2.4 Interprétation et intérêt du taux 22

2.4.1 Intérêt du taux par rapport à la survie 22

2.4.2 Le taux comme densité conditionnelle 24

2.4.3 L'unité du taux 24

2.4.4 La dynamique du taux 25

2.5 Mélange de taux . . . 26
2.5.1 Mélange de deux taux constants . . . 26
2.5.2 Mélange de deux taux croissants . . . 27
2.5.3 Mélange d'un taux constant et d'un taux croissant . . . 28
2.6 Vraisemblance . . . 29
2.7 Modèles de taux . . . 30
2.7.1 Spécification du modèle . . . 30
2.7.2 Vraisemblance du modèle . . . 31
2.7.3 Approximation du taux cumulé . . . 31
2.8 Approche de Poisson pour les modèles de survie . . . 31
2.8.1 Modèle de Poisson classique . . . 32
2.8.2 Équivalence des vraisemblances entre modèle de Poisson et modèle de survie . 33
2.9 Survie nette et modèles de taux en excès . . . 34
2.9.1 Décomposition du taux observé . . . 34
2.9.2 Estimateur de la survie nette de Pohar-Perme . . . 35
2.9.3 Vraisemblance du modèle de taux en excès . . . 36
2.9.4 Exemples de modèles de taux en excès . . . 37
2.9.5 Approche de Poisson pour le taux en excès . . . 37
2.10 Validation des modèles de taux . . . 38
2.10.1 Survie populationnelle et comparaison avec un estimateur non-paramétrique . . 38
2.10.2 Taux populationnel et comparaison à un modèle de taux constant par intervalles 39
2.10.3 Autres outils diagnostiques . . . 40

3 Sélection de modèle 41
3.1 Particularité de la sélection de modèle en survie . . . 41
3.2 Exemples de procédures de sélection de modèle . . . 41
3.2.1 Divergence de Kullback-Liebler . . . 42
3.2.2 Critère d'information d'Akaike (AIC) . . . 42
3.2.3 Choix parmi un ensemble de modèles candidats . . . 43
3.2.4 Procédures *backward* et *forward* . . . 43
3.2.5 Limites . . . 44

4 Pénalisation 45
4.1 Interpolation et lissage . . . 45
4.2 Splines de régression pénalisées . . . 46
4.2.1 Pénalité sur la dérivée seconde . . . 47
4.2.2 Matrice de pénalisation . . . 48
4.3 Pénaliser plusieurs splines unidimensionnelles . . . 51

4.3.1 Construction d'un modèle additif . . . 51
4.3.2 Contrainte de centrage . . . 52
4.4 Pénaliser les interactions . . . 53
4.4.1 Interactions entre une spline pénalisée et une variable continue . . . 53
4.4.2 Interactions entre une spline pénalisée et une variable catégorielle . . . 54
4.4.3 Splines multidimensionnelles pénalisées . . . 55

5 Inférence **57**
5.1 Estimateur du maximum de vraisemblance . . . 57
5.2 Estimateur du maximum de vraisemblance pénalisée . . . 58
5.2.1 Approche fréquentiste via les M-estimateurs . . . 59
5.2.2 Approche Bayésienne . . . 59
5.3 Variance des estimateurs . . . 60

6 Estimation des paramètres de lissage **63**
6.1 Complexité et degrés de liberté effectifs . . . 63
6.2 Likelihood Cross Validation (LCV) . . . 64
6.3 *Laplace approximate marginal likelihood* (LAML) . . . 67
6.4 Lien entre paramètre de lissage et variance d'un effet aléatoire . . . 68
6.5 Incertitude sur les paramètres de lissage . . . 68
6.5.1 Variance corrigée . . . 68
6.5.2 AIC corrigé . . . 69

II Développement d'un modèle de taux pénalisé **71**

7 Aperçu des modèles de survie pénalisés **73**
7.1 Approche de Poisson . . . 73
7.2 Modèles mixtes . . . 74
7.3 Inférence Bayésienne . . . 74
7.4 Inférence Fréquentiste . . . 74
7.5 Modèles de taux en excès pénalisés . . . 75

8 Proposition d'un modèle de taux pénalisé **77**
8.1 Le modèle . . . 77
8.1.1 Exemple de spécification de modèles avec des variables continues . . . 78
8.1.2 Exemple de spécification de modèle avec des variables catégorielles . . . 78
8.2 Calcul de la log-vraisemblance pénalisée . . . 79
8.3 Estimation des paramètres de régression à paramètres de lissage fixés (*inner algorithm*) 80

8.4 Les critères LCV et LAML pour l'estimation des paramètres de lissage (*outer algorithm*) 81
8.4.1 Laplace approximate marginal likelihood criterion (LAML) 81
8.4.2 Likelihood cross-validation criterion (LCV) 83
8.4.3 *Outer algorithm* 84
8.5 Comparaison de LCV et LAML 84
8.6 Variance des estimateurs 87
8.7 Effets aléatoires 87
8.8 Algorithme et techniques numériques 88
8.8.1 Schéma du double Newton-Raphson 88
8.8.2 Critères de convergence 88
8.8.3 Complexité algorithmique 89
8.8.4 Perturbation de la hessienne 91
8.8.5 Contrôle du pas 91
8.8.6 Inversion de la matrice de pénalisation 91
8.8.7 Paramètres d'échelle 92

9 Proposition d'un modèle de taux en excès pénalisé 93
9.1 Le modèle 93
9.2 Calcul de la vraisemblance et de ses dérivées 94
9.3 Modèle de taux en excès pénalisé à effets mixtes 95
9.4 Illustration de la pénalisation sur l'échelle du rapport de taux en excès 96

10 Étude par simulation des propriétés statistiques de la méthode proposée 99
10.1 Design 99
10.2 Description des tendances théoriques 100
10.3 Modèles ajustés 100
10.4 Évaluation de la performance de l'approche 101
10.5 Résultats 102

11 Le package *survPen* 107
11.1 Motivation 107
11.2 datCancer 107
11.3 Débuter avec *survPen* 108
11.3.1 Taux constant 108
11.3.2 Taux constant par intervalles 108
11.3.3 Taux log-linéaire 109
11.3.4 Splines cubiques restreintes 109
11.3.5 Splines cubiques restreintes pénalisées 109

11.4 Prédictions et sorties du modèle . . . 110
11.4.1 Prédictions standards . . . 110
11.4.2 Prédictions à la carte . . . 111
11.4.3 Résumé du modèle . . . 112
11.4.4 Sélection de modèle . . . 113
11.5 Estimation des paramètres de lissage . . . 114
11.6 Position des noeuds . . . 116
11.7 Taux en excès . . . 117
11.8 Produit tensoriel . . . 119
11.8.1 Deux dimensions . . . 119
11.8.2 Trois dimensions . . . 125
11.9 Interactions entre des splines pénalisées et des termes paramétriques . . . 127
11.9.1 Spécification avec des variables continues . . . 127
11.9.2 Illustration des variables *by* continues . . . 127
11.9.3 Spécification avec des variables catégorielles . . . 128
11.9.4 Illustration des variables *by* catégorielles . . . 129
11.10 Effets aléatoires . . . 134
11.11 Troncature à gauche . . . 137
11.12 Autres fonctionnalités utiles . . . 138
11.12.1 lambda . . . 138
11.12.2 beta.ini et rho.ini . . . 139
11.12.3 detail.rho et detail.beta . . . 140

12 Comparaison avec des approches existantes 149
12.1 Temps d'exécution . . . 149
12.2 Estimation de splines pénalisées : comparaison à *rstpm2* . . . 152
12.3 Estimation d'effets aléatoires : comparaison à *mexhaz* . . . 154
12.3.1 Simulation . . . 154
12.3.2 Modèles ajustés . . . 154
12.3.3 Résultats . . . 155

III Applications épidémiologiques 157

13 Effet de la défavorisation sociale sur la mortalité en excès des patients atteints de cancer 159
13.1 Contexte et objectifs . . . 159
13.2 Données . . . 159
13.3 Méthode . . . 160

13.4 Résultats 160
13.5 Conclusion 162

14 Étude de la mortalité en excès des patients atteints de sclérose en plaques 163
14.1 Contexte 163
14.2 Objectifs 164
14.3 Données 164
14.4 Méthode 164
14.5 Résultats 164
14.6 Conclusion 167

IV Conclusion 169

Résumé de la contribution originale de la thèse 171

Discussion 173

Annexes 177
A Intégration numérique 178
A.1 Cavalieri-Simpson 178
A.2 Gauss-Legendre 178
A.3 Calibration du nombre de nœuds pour la quadrature de Gauss-Legendre 179

Valorisation scientifique 183
B Liste des articles et des communications scientifiques 183
C Article méthodologique publié dans JRSSC 186
D Article *survPen* publié dans JOSS 221

Bibliographie 225

Table des figures

1.1 Données simulées et vraie fonction 11
1.2 Comparaison entre régression polynomiale et spline de régression 11
1.3 Spline cubique et spline cubique restreinte à 10 nœuds ajustées sur le jeu de données *mcycle* du package R *MASS* 13
1.4 Exemple de bases associées à une *cubic regression spline* 16
1.5 Construction d'une *cubic regression spline* 17

2.1 Fonction de survie obtenue par l'estimateur de Kaplan-Meier 21
2.2 Comparaison de courbes de taux et de courbes de survie. La première ligne correspond à une loi exponentielle de paramètre $\lambda = 0,2$. La deuxième ligne à une loi Weibull de paramètres $\lambda = 0,2$ et $\gamma = 0,9$. La troisième ligne à une loi Weibull de paramètres $\lambda = 0,2$ et $\gamma = 1,1$. 23
2.3 Mélange de deux taux constants 26
2.4 Mélange de deux taux constants, proportion du groupe1 27
2.5 Mélange de deux taux croissants 28
2.6 Mélange d'un taux constant et d'un taux croissant 28
2.7 Comparaison entre survie brute et survie nette (Nelson-Aalen vs Pohar-Perme) 36
2.8 Comparaison entre taux populationnel et taux constant par intervalles 39

4.1 Lien entre régularité et valeur de l'intégrale de la dérivée seconde au carré pour quatre fonctions définies sur $[0;1]$ 48

8.1 Taux et survie théoriques associés aux scénarios constant et non-linéaire 85
8.2 Allures des critères LCV et LAML dans les deux scénarios 86

9.1 Comparaison des taux en excès Hommes/Femmes chez les patients atteints de cancer lèvre-bouche-pharynx 97
9.2 Rapport de taux en excès Hommes/Femmes chez les patients atteints de cancer lèvre-bouche-pharynx 98

10.1 Simulation : tendances théoriques 101
10.2 Boxplots de la RMISE (multipliés par 100) sur le taux en excès pour chaque scénario et taille d'échantillon (45 combinaisons pour œsophage et 60 pour col). 102

10.3 Boxplots du biais et de la RMSE (multipliés par 100) sur la survie nette pour chaque scénario et taille d'échantillon (270 combinaisons pour œsophage et 360 pour col). . . 103

11.1 *survPen* - Comparaison de différents modèles 110
11.2 *survPen* - Prédictions avec intervalles de confiance 111
11.3 *survPen* - LCV vs LAML avec 10 paramètres de régression 115
11.4 *survPen* - LCV et LAML comme fonctions des paramètres de lissage 116
11.5 *survPen* - Taux brut vs taux en excès 118
11.6 *survPen* - Surfaces du taux, tensor vs tint 121
11.7 *survPen* - Surfaces du taux, tensor vs tint, deuxième exemple 123
11.8 *survPen* - Coupes 2D, tensor vs tint 124
11.9 *survPen* - Surfaces prédites par le tensor à trois dimensions 126
11.10 *survPen* - taux et rapport de taux théoriques 130
11.11 *survPen* - comparaison des taux, variables *by* catégorielles 132
11.12 *survPen* - comparaison des rapports de taux, variables *by* catégorielles 133
11.13 *survPen* - troncature gauche 138
11.14 *survPen* - Effet du paramètre de lissage sur la prédiction 139

12.1 Comparaison *rstpm2*. Boxplots de la RMISE (multipliée par 100) sur le taux en excès pour chaque scénario (taille d'échantillon = 2 000). 153
12.2 Comparaison *rstpm2*. Boxplots du biais et de la RMSE (multipliés par 100) sur la survie nette pour chaque scénario (taille d'échantillon = 2 000). 153

13.1 Résultats principaux de l'étude sur l'EDI 161

14.1 Résultats principaux de l'étude sur la SEP 165
14.2 Adéquation du modèle chez les rémittents 166
14.3 Adéquation du modèle chez les progressifs 167
A.1 Calibration Gauss-Legendre : écarts sur les prédictions du modèle univarié 180
A.2 Calibration Gauss-Legendre : écarts sur les prédictions de taux du modèle trivarié . . 180
A.3 Calibration Gauss-Legendre : écarts sur les prédictions de survie du modèle trivarié . 181

Liste des tableaux

2.1 Estimateur de Kaplan-Meier 21
2.2 Estimateur de Nelson-Aalen 22
2.3 Estimateur de Pohar-Perme 36

10.1 Médianes des probabilités de couverture (en %) dans l'estimation du taux en excès en fonction du type d'estimation de variance 104
10.2 Médianes des probabilités de couverture (en %) dans l'estimation de la survie nette en fonction du type d'estimation de variance 104
10.3 Médianes des degrés de liberté effectifs parmi tous les modèles ajustés 105

11.1 *survPen* - jeu de données datCancer 108

12.1 Fonctionnalités de différentes approches en termes de splines pénalisées pour les modèles de survie 150
12.2 Temps d'exécution (en secondes) pour N = 2 000 151
12.3 Temps d'exécution (en secondes) pour N = 20 000 151
12.4 Comparaison *mexhaz*. Biais, probabilités de couverture et RMSE selon les quatre scénarios et les trois modèles proposés. 155
12.5 Comparaison *mexhaz*. Temps de calcul médian pour ajuster un modèle selon les quatre scénarios et les trois modèles proposés. 156

Liste des symboles

n	Taille de la population étudiée
$\boldsymbol{X}$	Matrice de design du modèle
$\boldsymbol{\beta}$	Vecteur des paramètres de régression
p	Longueur du vecteur $\boldsymbol{\beta}$
T	Variable aléatoire de durée jusqu'à la survenue de l'événement étudié
C	Variable aléatoire de censure à droite
δ	Indicatrice d'événement
h	Taux de mortalité observé
h_P	Taux de mortalité attendu
h_E	Taux de mortalité en excès
H	Taux de mortalité cumulé observé
H_E	Taux de mortalité cumulé en excès
S	Fonction de survie toutes causes
SN	Fonction de survie nette
$\boldsymbol{GL}^k$	k^e matrice de design associée à la quadrature de Gauss-Legendre
q	Nombre de points associés à la quadrature de Gauss-Legendre
$\boldsymbol{\lambda}$	Vecteur des paramètres de lissage
$\boldsymbol{\rho}$	Vecteur des logarithmes des paramètres de lissage
M	Longueur des vecteurs $\boldsymbol{\lambda}$ et $\boldsymbol{\rho}$
$\boldsymbol{S}^m$	Matrice de pénalisation associée au m^e paramètre de lissage
$\boldsymbol{S}^\lambda$	Matrice de pénalisation associée au m^e paramètre de lissage
l	Log-vraisemblance
$\mathcal{L}$	Log-vraisemblance pénalisée
$\boldsymbol{H}$	Hessienne de l'opposé de la log-vraisemblance
$\boldsymbol{\mathcal{H}}$	Hessienne de l'opposé de la log-vraisemblance pénalisée
$\boldsymbol{\mathcal{I}}$	Information de Fisher
$\boldsymbol{\mathcal{J}}$	Information de Fisher observée

$\boldsymbol{V}_{\hat{\beta}}$ Matrice de variance fréquentiste

$\boldsymbol{V}_{\beta}$ Matrice de variance Bayésienne

$\boldsymbol{V}'_{\beta}$ Matrice de variance Bayésienne corrigée pour l'incertitude sur les paramètres de lissage

Introduction

Analyse de survie

L'étude du délai de survenue d'un événement représente un champ très important des statistiques. Souvent appelée « analyse de survie » (le décès étant un événement fréquemment étudié), elle permet d'étudier des événements de toutes natures : défaillance d'un système électronique, apparition d'une récidive de cancer, arrêt maladie d'un salarié, invalidité d'un individu, formation d'un tsunami ou d'un tremblement de terre, etc.

L'indicateur le plus souvent restitué est la fonction de survie. Cette dernière donne, à un instant t, la probabilité de ne pas avoir encore présenté l'événement considéré. La fonction de survie représente ainsi une vision cumulée du processus conduisant à l'événement. Par exemple, si la survie à 5 ans vaut 60%, cela veut dire que, au bout de 5 années de suivi, 60% des individus présents au début de l'étude n'ont toujours pas présenté l'événement. Si l'on calcule la survie à différents temps, on obtient alors ce que l'on appelle une courbe de survie. Les différentes probabilités données par la courbe de survie ne sont toutefois pas suffisantes pour comprendre le phénomène de survenue de l'événement. En effet, la pente de la courbe, c'est-à-dire la variation de la survie, est également informative. Typiquement, une forte pente sur un intervalle de temps donné implique une probabilité importante de présenter l'événement durant cet intervalle.

La survie est intimement liée à un autre indicateur : le taux de survenue de l'événement (parfois appelé taux de défaillance). Étant donné que notre événement d'intérêt sera le décès, nous parlerons ici de taux de mortalité. Le taux de mortalité représente une vision instantanée du processus de survenue d'événement. En effet, le taux est intimement lié à la notion de pente de la courbe de survie évoquée plus haut et il constitue la « force de mortalité » qui s'applique à un instant donné. Il est ainsi particulièrement utile de décrire la **dynamique du taux de mortalité**, c'est-à-dire l'évolution du taux au cours du temps, afin d'identifier les moments où cette force est la plus intense. Cette vision dynamique du processus entre le début du suivi et un temps donné vient donc utilement compléter la survie qui ne fait que restituer les effets cumulés de cette dynamique au temps considéré.

Épidémiologie

En épidémiologie, l'étude de la survie des patients atteints d'une certaine pathologie, et surtout de ses déterminants, représente un enjeu majeur de santé publique. En effet, la compréhension des phénomènes entraînant le décès des patients permet notamment d'orienter les médecins vers un traitement le plus adapté possible en fonction de l'avancement dans la maladie et des caractéristiques du patient. Pour suivre le raisonnement du paragraphe précédent, l'enjeu majeur devient donc de déterminer le taux de mortalité subi par les patients, et bien sûr ses déterminants. Un intérêt particulier porte sur le taux de mortalité supplémentaire subi par les patients en comparaison avec des

personnes non atteintes.

Deux indicateurs différents existent pour quantifier cette mortalité supplémentaire : la mortalité dite ≪ cause spécifique ≫ et la mortalité en excès par rapport à une mortalité dite ≪ attendue ≫ en l'absence de maladie. La première s'appuie sur la cause de décès de chaque individu afin d'identifier les décès dus à la pathologie étudiée et les décès dus aux autres causes. La seconde utilise le fait que, dans la grande majorité des cas, les patients atteints d'une maladie chronique présentent un excès de mortalité par rapport à une population ne présentant pas la maladie en question. Ainsi, la mortalité spécifique suppose que la cause de décès est disponible et fiable tandis que la mortalité en excès suppose que l'on puisse calculer les taux de mortalité attendus en l'absence de maladie pour des personnes dont les caractéristiques démographiques sont comparables.

Une des limites principales de la mortalité spécifique est qu'elle repose sur le concept de cause unique de décès. Quand bien même la cause serait toujours unique, celle-ci n'est pas toujours disponible ou fiable. En outre, on se retrouve très souvent confrontés non pas à une cause mais des causes et il devient très difficile d'identifier les décès qui sont liés à la pathologie étudiée, notamment à long terme. Par exemple, le décès d'un patient par suicide doit-il être considéré comme un décès dû à la maladie ? De même, comment considérer un patient atteint d'une maladie (neurologique par exemple) qui l'empêche de pratiquer une activité physique régulière et qui décède d'une maladie cardiovasculaire ?

La mortalité en excès ne présente pas une telle limite. En outre, la mortalité en excès permet de réaliser des comparaisons entre différentes périodes et différentes zones géographiques puisqu'elle prend en compte le fait que la mortalité autres causes diffère d'une année sur l'autre ou d'un pays à l'autre par exemple.

Dans cette thèse, nous nous concentrerons donc sur la mortalité en excès pour toutes les raisons évoquées plus haut.

Cancers

La mortalité en excès représente un enjeu crucial en épidémiologie des cancers. Le temps écoulé depuis le diagnostic et l'âge au diagnostic représentent des facteurs pronostiques majeurs de la survie des patients atteints de cancer, c'est à dire qu'ils influencent la mortalité des patients de manière non négligeable. Il est donc primordial de modéliser correctement les effets du temps et de l'âge au diagnostic sur la mortalité en excès. On doit notamment être en mesure de restituer la dynamique du taux de mortalité en excès et d'identifier comment cette dynamique varie en fonction de l'âge au diagnostic.

D'autres facteurs pronostiques revêtent une importance particulière en épidémiologie des cancers. Par exemple, les études dites de ≪ tendances ≫ s'attachent à décrire comment la dynamique du taux de mortalité en excès varie en fonction de l'année de diagnostic. D'un point de vue épidémiologique, l'année de diagnostic est une variable qui témoigne de l'évolution des pratiques médicales et l'analyse de son effet sur le taux est donc essentielle. De plus, il est naturel de se demander si une tendance favorable de dynamique du taux de mortalité en excès observée chez les sujets jeunes est également retrouvée chez des patients âgés : une évolution favorable des pratiques a-t-elle bénéficié à tous ? Nous touchons ici à la notion d'interaction (entre l'âge et l'année) qui occupera une place centrale dans ce

travail de thèse.

D'autres études s'intéressent à l'impact des inégalités sociales sur la survie des patients atteints de cancer. Ces études ont alors pour but de décrire l'effet d'un indice de défavorisation économique et sociale. Dans ce cadre, les épidémiologistes peuvent légitimement se demander si l'effet de cet indice est le même quel que soit le temps écoulé depuis le diagnostic. Nous touchons ici à une autre notion centrale de ce travail qui est l'interaction entre une covariable et le temps, aussi appelée non-proportionnalité.

Notons ici que le stade au diagnostic est un facteur pronostique majeur. Lorsque les études sont effectuées tous stades confondus, il est toujours important d'interpréter les effets de l'année de diagnostic ou des inégalités en fonction de cet élément. Par exemple, l'effet des inégalités peut passer par celui de stades plus avancés chez les personnes défavorisés, ou un effet favorable de l'année de diagnostic peut refléter une évolution temporelle de la distribution du stade au diagnostic, avec des stades moins avancés pour les années de diagnostic les plus récentes.

Sclérose en plaques

La sclérose en plaques (SEP) est une maladie neurologique chronique pour laquelle la dynamique du taux de mortalité en excès a très peu été étudiée. Le faible excès de mortalité associé à la SEP en fait pourtant un sujet d'étude particulièrement adapté à la méthode de la mortalité en excès. Il faut également noter que le suivi des patients peut atteindre une cinquantaine d'année (contre 10 ou 15 ans en général pour les données de cancer françaises). Pour les patients atteints de SEP, les épidémiologistes se demandent notamment si, à âge courant identique, le taux en excès est le même quel que soit l'âge d'entrée dans la maladie. Autrement dit, si l'on prend l'exemple de deux patients ayant 60 ans aujourd'hui, ont-ils le même taux de mortalité en excès sachant que l'un a été diagnostiqué à l'âge de 20 ans et l'autre à 50 ans ?

Questions statistiques posées par la problématique épidémiologique

Les problèmes épidémiologiques exposés plus haut impliquent un besoin important de développement statistique pour les résoudre. Ainsi, il est nécessaire de traduire ces problématiques épidémiologiques en problématiques statistiques.

Le premier élément de notre cahier des charges épidémiologique est l'accès à la dynamique du taux de mortalité en excès. Cet aspect implique une modélisation du taux de mortalité en excès qui inclut le taux de base (contrairement au très populaire modèle de Cox par exemple, qui considère le taux de base comme un paramètre de « nuisance »).

Ensuite, il est également nécessaire de modéliser les effets des facteurs pronostiques sur la mortalité en excès. Lorsque ces facteurs sont continus, il convient alors de ne pas les catégoriser afin de préserver un maximum d'information. Il faut par exemple éviter d'étudier l'âge en classes d'âges. En outre, il faut tenir compte de la potentielle non-linéarité des effets des covariables continues.

Par ailleurs, comme déjà évoqué ci-dessus, il faut également tenir compte des interactions entre les effets des différents facteurs pronostiques. On parle d'interaction lorsque l'effet d'un facteur pronostique dépend d'un autre facteur, typiquement lorsque l'effet de l'âge au diagnostic d'un patient dépend de

l'année durant laquelle il a été diagnostiqué. Si l'effet d'un facteur pronostique dépend du temps de suivi considéré, on dira qu'il s'agit d'un effet dépendant du temps, ou non-proportionnel.

La notion de non-linéarité des effets des covariables continues amène à considérer la notion de fonction flexible. Les splines de régression, qui sont des polynômes par morceaux paramétriques, constituent un outil de choix en termes de flexibilité. En effet, celles-ci sont notamment capables de restituer des effets non-linéaires très complexes. En ce qui concerne les interactions entre variables continues, il est assez facile de former une spline multidimensionnelle à partir de deux ou plusieurs splines unidimensionnelles marginales. Par exemple, en effectuant le produit de toutes les splines marginales, on construit une spline multidimensionnelle appelée « produit tensoriel ».

Toutefois, la flexibilité des splines de régression a un coût : le sur-ajustement. On parle de sur-ajustement lorsqu'un modèle contenant de nombreux paramètres est ajusté sur un échantillon de taille trop faible. Si un nouvel échantillon de cette population était utilisé, un modèle sur-ajusté risquerait de ne plus être valide. De plus, le sur-ajustement va à l'encontre de l'idée de modélisation qui est de proposer une vision simplifiée mais utile de la réalité.

Plus le nombre de paramètres de régression d'un modèle est important, plus le risque de sur-ajustement est élevé et plus les prédictions deviennent erratiques (c'est à dire que leur variance est élevée). Afin d'éviter cette instabilité tout en conservant la même flexibilité, il convient de lisser les prédictions obtenues. Ce lissage est obtenu en pénalisant les fonctions qui conduisent à des prédictions trop instables. Le lissage se justifie par la volonté d'obtenir une inférence robuste aux fluctuations d'échantillonnage. Cette inférence robuste passe par une stabilité des prédictions. Typiquement, pour prendre l'exemple de l'âge, toutes choses égales par ailleurs, la mortalité d'un individu de 60 ans n'est pas très différente de celle d'un individu de 61 ans.

Les splines de régression pénalisées ont été popularisées dans le cadre des modèles additifs généralisés (GAM). Leur principe est simple : au lieu d'estimer les paramètres de régression en maximisant la vraisemblance, on cherche à maximiser un compromis entre la fidélité aux données (représentée par la vraisemblance) et un terme de pénalité (représenté par la régularité de l'effet à estimer) : on parle de vraisemblance pénalisée. Techniquement, à chaque spline peuvent être associés un ou plusieurs termes de pénalité contrôlés par des paramètres de lissage. Les paramètres de lissage représentent les degrés de pénalisation souhaités. En pratique, ils sont inconnus et doivent être estimés tout comme les paramètres de régression.

Objectifs de la thèse

L'**objectif principal** de cette thèse est de développer une méthode permettant de modéliser le taux de mortalité (en excès) à l'aide de splines de régression multidimensionnelles pénalisées afin :

- d'avoir accès à la dynamique du taux de mortalité
- de modéliser les effets non-linéaires et non-proportionnels des covariables
- de modéliser les interactions entre covariables

grâce à une modélisation flexible mais produisant des estimations lisses.

Le **second objectif** est de proposer une implémentation logicielle de cette méthode.
Le **troisième objectif** de la thèse est d'appliquer la méthode développée afin de répondre aux problèmes épidémiologiques exposés ci-dessus, en cancérologie et dans le domaine de la SEP.

Structuration de la thèse

La première partie de cette thèse rappelle les différents éléments théoriques nécessaires à l'élaboration d'un modèle de taux pénalisé : splines de régression, modèles de survie, sélection de modèles, pénalisation et inférence. La deuxième partie détaille la méthode proposée dans cette thèse, ses propriétés statistiques, son implémentation logicielle puis propose des comparaisons avec des méthodes déjà existantes. La troisième partie décrit deux exemples d'application de la méthode développée : étude de l'effet de la défavorisation sociale sur la mortalité en excès dans le cancer du col de l'utérus ; et étude de l'effet de l'âge courant sur la mortalité due à la sclérose en plaques. Enfin, la quatrième et dernière partie fait office de conclusion et ouvre la voie à de nouvelles perspectives de recherche.

Première partie

Éléments théoriques nécessaires aux objectifs de la thèse

Chapitre 1

Modélisation de l'effet du temps et des covariables par des splines

1.1 Régression polynomiale

Lorsque l'on cherche à modéliser l'effet d'une variable explicative continue x (i.e le temps ou une covariable), il est usuel de considérer la forme fonctionnelle f telle que $f(x) = \beta_0 + \beta_1 x$ avec β_0 et β_1 les paramètres à estimer. Toutefois, si l'hypothèse de linéarité ne tient pas, il faut trouver une alternative à cette approche. On peut facilement généraliser la construction de la forme fonctionnelle f en considérant un polynôme de degré p :

$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p$$

Supposons par exemple que l'on pense que f est un polynôme de degré 3. Ainsi, l'espace des polynômes de degrés 3 et inférieurs contiennent la fonction f et celle-ci s'écrit :

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Le gros inconvénient des bases polynomiales est que la complexité de la fonction est directement liée au degré du polynôme. Ainsi, plus la forme à approcher sera complexe, plus le degré du polynôme devra être important. En outre, les polynômes ont de bonnes propriétés d'approximation **au voisinage d'un point** (le théorème de Taylor en est une illustration) mais étendre ces propriétés à tout l'ensemble de définition de la fonction est beaucoup plus problématique. En effet, les bases du polynôme ne sont pas définies de manière locale, c'est à dire que la valeur de la fonction en un point x dépend fortement de la valeur des données très éloignées de x.

1.2 Splines de régression

Les splines de régression sont des fonctions polynomiales définies par morceaux. Les points de jonction entre les différents morceaux sont appelés nœuds. De manière générale, une spline f à k paramètres s'écrit :

$$f(x) = \sum_{j=1}^{k} \beta_j b_j(x)$$

où les b_j sont les bases de la spline (connues mais à spécifier) et les β_j sont les paramètres à estimer.

L'idée des splines est de construire des polynômes par morceaux et ainsi de rendre les bases définies de manière locale. Le terme « splines de régression » correspond au fait que les k nœuds sont définis a priori et que k est inférieur au nombre de valeurs à approcher. C'est la raison pour laquelle, pour une complexité donnée, les splines demandent en général moins de paramètres que les polynômes pour approcher les données. Une spline est dite linéaire (respectivement quadratique, cubique, quartique) si elle est composée de polynômes dont le degré maximal vaut 1 (respectivement 2, 3, 4). Il existe de nombreuses bases, mais certaines ont des propriétés plus intéressantes que d'autres (De Boor, 1972).

Les splines de régression apportent de la flexibilité tout en conservant l'avantage de l'écriture linéaire par rapport au vecteur des paramètres de régression $\boldsymbol{\beta}$. En effet, si l'on se donne les données $x_1, \ldots, x_n$, nous pouvons écrire

$$\{f(x_i)\}_{i=1,\ldots,n} = \boldsymbol{X}\boldsymbol{\beta}$$

où $\boldsymbol{X} = \begin{bmatrix} b_1(x_1) & \ldots & b_k(x_1) \\ \ldots & \ldots & \ldots \\ b_1(x_n) & \ldots & b_k(x_n) \end{bmatrix}$ est la matrice de design engendrée par la spline f.

Régression polynomiale vs splines de régression :

Considérons le modèle suivant :

$$y_i = f(x_i) + \epsilon_i \quad \text{avec} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Choisissons 500 valeurs de x dans $[0; 1]$ et simulons les 500 valeurs de y correspondantes à partir de :

$$f(x) = sin\left(18x - 8x^3\right) \quad \text{et} \quad \sigma = 0,3$$

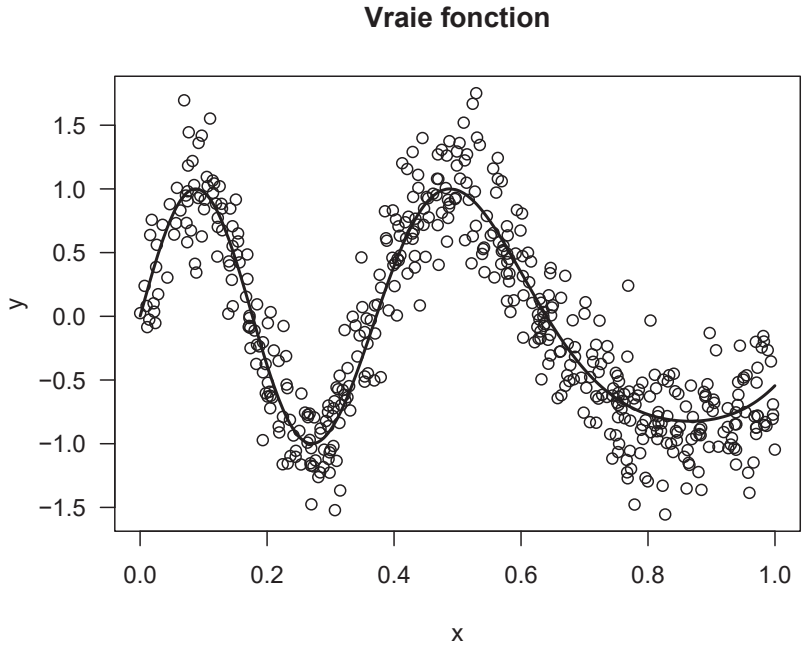

Figure 1.1 – Données simulées et vraie fonction

La figure 1.1 montre les données simulées ainsi que la vraie fonction utilisée. La figure 1.2 compare les ajustements de trois régressions polynomiales de degrés 5, 7 et 9 respectivement aux ajustements de trois splines (linéaire, quadratique et cubique) comportant chacune 6 nœuds intérieurs. Les barres verticales indiquent la position des nœuds (y compris les nœuds extérieurs).

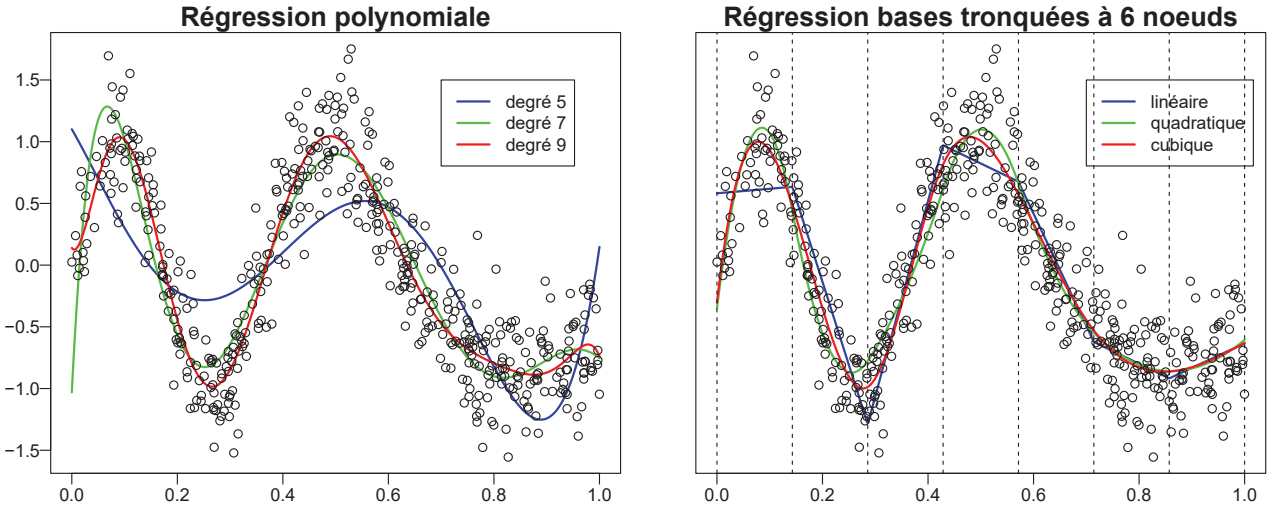

Figure 1.2 – Comparaison entre régression polynomiale et spline de régression

Dans l'exemple de la figure 1.2, une bonne approximation de la vraie fonction requiert un polynôme de degré 9 alors qu'une spline cubique voire quadratique suffit (sous réserve de correctement choisir le nombre et la position des nœuds).

1.2.1 Base des puissances tronquées

Supposons que l'on veuille construire un polynôme par morceaux f de degré d sur un intervalle $[a; b]$. On commence par définir M points $(x_1, \ldots, x_M)$ ou nœuds intérieurs tels que $a < x_1 < \cdots < x_M < b$.

L'une des manières les plus simples d'écrire la spline f est la suivante :

$$f(x) = \sum_{k=1}^{d} \beta_k x^k + \sum_{j=1}^{M} \beta_{d+j}(x - x_j)_+^d$$

avec $(x - x_j)_+ = max(x - x_j, 0)$.

Une des appellations de cette base est la base des puissances tronquées (voir Cornillon and Matzner-Løber, 2011, Section 10.2.2).

Cette base est certainement la plus pédagogique lorsque l'on souhaite présenter les splines. Toutefois, elle présente des problèmes de conditionnement au sens de l'analyse numérique (De Boor, 1986). En effet, pour $x > x_j$, les $d + j$ premiers termes de la spline sont non nuls en utilisant la base tronquée. Ainsi, l'évaluation de $f(x)$ réclame de sommer $d+j$ termes de la base, ce qui constitue un effort d'ordre computationnel maximal et inutile (Ueberhuber, 1997).

1.2.2 B-splines

Les B-splines permettent de pallier les problèmes de conditionnement de la régression polynomiale tout en assurant une évaluation efficiente de la spline.

Il existe plusieurs moyens d'évaluer une B-spline (De Boor, 1972). De manière traditionnelle, les B-splines sont définies à partir d'une formule de récurrence (De Boor, 1986).

À partir d'un certain nombre de nœuds x_i, la B-spline d'ordre 1 est composée par les fonctions :

$$B_{i1}(x) = \begin{cases} 1 & \text{si } x_i \leq x < x_{i+1} \\ 0 & \text{sinon.} \end{cases}$$

Les B-splines d'ordre supérieur sont alors obtenues par récurrence de la manière suivante :

$$B_{ik}(x) = \omega_{ik} B_{i,k-1} + (1 - \omega_{i+1,k}) B_{i+1,k-1}$$

avec

$$\omega_{ik}(x) = \begin{cases} \frac{x - x_i}{x_{i+k-1} - x_i} & \text{si } x_i \neq x_{i+k-1} \\ 0 & \text{sinon.} \end{cases}$$

Attention ici à la confusion entre **ordre** et **degré** de la spline. Une B-spline d'**ordre** k est composée de polynômes par morceaux de **degré** $k - 1$.

Définie de cette manière, chaque base d'une B-spline d'ordre k est non nulle sur k intervalles uniquement. C'est cette propriété dite de support local qui rend les B-splines plus attractives numériquement que la base des puissances tronquées.

1.2.3 Splines cubiques restreintes

On dira d'une spline cubique f qu'elle est restreinte (ou naturelle) si sa dérivée seconde s'annule en ses nœuds extérieurs (avec l'écriture utilisée plus haut cela revient à imposer $f''(x_1) = f''(x_M) = 0$).

Les splines cubiques restreintes ont l'avantage d'être moins sensibles aux « effets de bord » que les splines cubiques classiques. La figure 1.3 illustre cette propriété des splines cubiques restreintes, notamment pour la projection au-delà des nœuds extérieurs (lignes pointillées verticales).

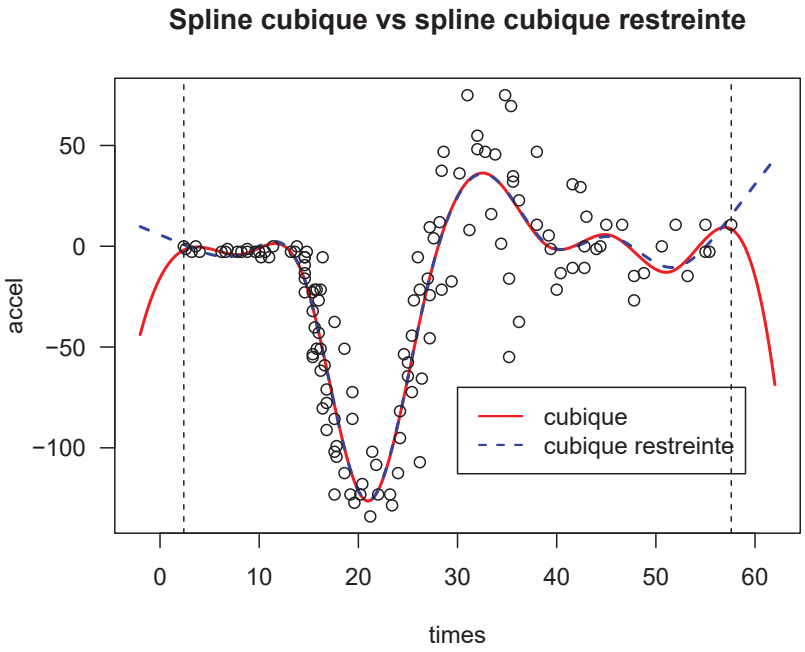

Figure 1.3 – Spline cubique et spline cubique restreinte à 10 nœuds ajustées sur le jeu de données *mcycle* du package R *MASS*

Base des *cubic regression splines*

Plusieurs bases existent pour représenter des splines cubiques, notamment la base des puissances tronquées ou les B-splines vues précédemment. Une autre base possible consiste à paramétrer la spline en fonction de ses valeurs aux nœuds (voir *cubic regression splines*, Wood 2017). Cette base nous servira dans la suite à engendrer le même espace que des B-splines restreintes (linéaires aux bords) tout en facilitant la définition de la pénalisation (voir chapitre 4). Notez également que le nom de la base est volontairement non traduit afin de renvoyer à Wood (2017) plus facilement.

Soit f une spline cubique avec k nœuds, $x_1, \ldots, x_k$. En posant $\beta_j = f(x_j)$ et $\delta_j = f''(x_j)$, f est définie sur l'intervalle $[x_j; x_{j+1}]$ de la manière suivante :

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1}$$

Avec

$$a_j^-(x) = \frac{x_{j+1} - x}{h_j}$$

$$a_j^+(x) = \frac{x - x_j}{h_j}$$

$$c_j^-(x) = \frac{1}{6}\left[\frac{(x_{j+1} - x)^3}{h_j} - h_j(x_{j+1} - x)\right]$$

$$c_j^+(x) = \frac{1}{6}\left[\frac{(x - x_j)^3}{h_j} - h_j(x - x_j)\right]$$

et $h_j = x_{j+1} - x_j$.

La fonction f étant une spline cubique, f' doit être continue. En outre, pour définir f comme une spline cubique naturelle, on doit imposer le fait que sa dérivée seconde s'annule en ses nœuds extérieurs.

Pour ce qui concerne la continuité de f', considérons le nœud x_{j+1}. On veut que la dérivée de f en ce point soit la même que l'on se place sur l'intervalle $[x_j; x_{j+1}]$ ou sur l'intervalle $[x_{j+1}; x_{j+2}]$.

Sur $[x_j; x_{j+1}]$, $f'(x)$ vaut

$$f'(x) = -\frac{1}{h_j}\beta_j + \frac{1}{h_j}\beta_{j+1} + \frac{1}{6}\left[\frac{-3(x_{j+1} - x)^2}{h_j} + h_j\right]\delta_j + \frac{1}{6}\left[\frac{3(x - x_j)^2}{h_j} - h_j\right]\delta_{j+1}$$

Tandis que sur $[x_{j+1}; x_{j+2}]$, $f'(x)$ vaut

$$f'(x) = -\frac{1}{h_{j+1}}\beta_{j+1} + \frac{1}{h_{j+1}}\beta_{j+2} + \frac{1}{6}\left[\frac{-3(x_{j+2} - x)^2}{h_{j+1}} + h_{j+1}\right]\delta_{j+1} + \frac{1}{6}\left[\frac{3(x - x_{j+1})^2}{h_{j+1}} - h_{j+1}\right]\delta_{j+2}$$

La condition de continuité de f' en x_{j+1} s'écrit donc

$$-\frac{1}{h_j}\beta_j + \frac{1}{h_j}\beta_{j+1} + \frac{h_j}{6}\delta_j + \frac{h_j}{3}\delta_{j+1} = -\frac{1}{h_{j+1}}\beta_{j+1} + \frac{1}{h_{j+1}}\beta_{j+2} - \frac{h_{j+1}}{3}\delta_{j+1} - \frac{h_{j+1}}{6}\delta_{j+2}$$

En réarrangeant les termes on obtient

$$-\frac{1}{h_j}\beta_j + \left(\frac{1}{h_j} + \frac{1}{h_{j+1}}\right)\beta_{j+1} - \frac{1}{h_{j+1}}\beta_{j+2} = -\frac{h_j}{6}\delta_j - \left(\frac{h_{j+1}}{3} + \frac{h_j}{3}\right)\delta_{j+1} - \frac{h_{j+1}}{6}\delta_{j+2}$$

Finalement, en multipliant par -1 il vient :

$$\frac{1}{h_j}\beta_j - \left(\frac{1}{h_j} + \frac{1}{h_{j+1}}\right)\beta_{j+1} + \frac{1}{h_{j+1}}\beta_{j+2} = \frac{h_j}{6}\delta_j + \left(\frac{h_{j+1}}{3} + \frac{h_j}{3}\right)\delta_{j+1} + \frac{h_{j+1}}{6}\delta_{j+2}$$

Cette contrainte s'applique sur tous les $k - 2$ nœuds intérieurs, c'est à dire pour x_{j+1} avec $j = 1, \ldots, k - 2$.

Nota Bene : Sur le même principe, on peut vérifier que la continuité de f'' est déjà assurée par la définition même de la base. En effet, sur $[x_j; x_{j+1}]$ comme sur $[x_{j+1}; x_{j+2}]$, on a $f''(x_{j+1}) = \delta_{j+1}$.

La contrainte de continuité de f' peut s'exprimer de manière matricielle. Soient $\boldsymbol{D}$ et $\boldsymbol{B}$ les matrices telles que, pour $i = 1, \ldots, k - 2$:

$$D_{i,i} = \frac{1}{h_i}, \quad D_{i,i+1} = -\frac{1}{h_i} - \frac{1}{h_{i+1}}, \quad D_{i,i+2} = \frac{1}{h_{i+1}} \quad \text{et} \quad B_{i,i} = \frac{h_i + h_{i+1}}{3}$$

Et pour $i = 1, \ldots, k-3$,

$$B_{i,i+1} = \frac{h_{i+1}}{6} \quad \text{et} \quad B_{i+1,i} = \frac{h_{i+1}}{6}$$

Les contraintes de continuité de f' et de dérivées secondes nulles en x_1 et x_k impliquent alors l'égalité suivante :

$$\boldsymbol{B}\boldsymbol{\delta}^- = \boldsymbol{D}\boldsymbol{\beta} \tag{1.1}$$

avec $\boldsymbol{\delta}^- = (\delta_2, \ldots, \delta_{k-1})^T$ et $\delta_1 = \delta_k = 0$.

Si l'on définit $\boldsymbol{F}^- = \boldsymbol{B}^{-1}\boldsymbol{D}$ et

$$\boldsymbol{F} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{F}^- \\ \mathbf{0} \end{bmatrix}$$

où $\mathbf{0}$ est une ligne remplie de zéros, il vient $\boldsymbol{\delta} = \boldsymbol{F}\boldsymbol{\beta}$.

Nous pouvons donc réécrire la spline f sur l'intervalle $[x_j; x_{j+1}]$ uniquement à partir des éléments de $\boldsymbol{\beta}$:

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\boldsymbol{F}_j\boldsymbol{\beta} + c_j^+(x)\boldsymbol{F}_{j+1}\boldsymbol{\beta}$$

Au final, on définit f sur l'intervalle $[x_1; x_k]$ de la manière suivante :

$$f(x) = \sum_{i=1}^{k} b_i(x)\beta_i$$

avec

$$b_i(x) = \sum_{j=1}^{k-1} \left(F_{j,i}c_j^-(x) + F_{j+1,i}c_j^+(x) + a_j^-(x)\mathbb{1}_{(i=j)} + a_j^+(x)\mathbb{1}_{(i=j+1)} \right) \mathbb{1}_{[x_j;x_{j+1}]}(x)$$

Ainsi, les bases sont entièrement déterminées par les nœuds spécifiés. La figure 1.4 donne les 6 bases b_j associées à une spline définie sur $[0; 1]$ par les nœuds 0, 0,2, 0,4, 0,6, 0,8 et 1. Les lignes verticales représentent ces mêmes nœuds.

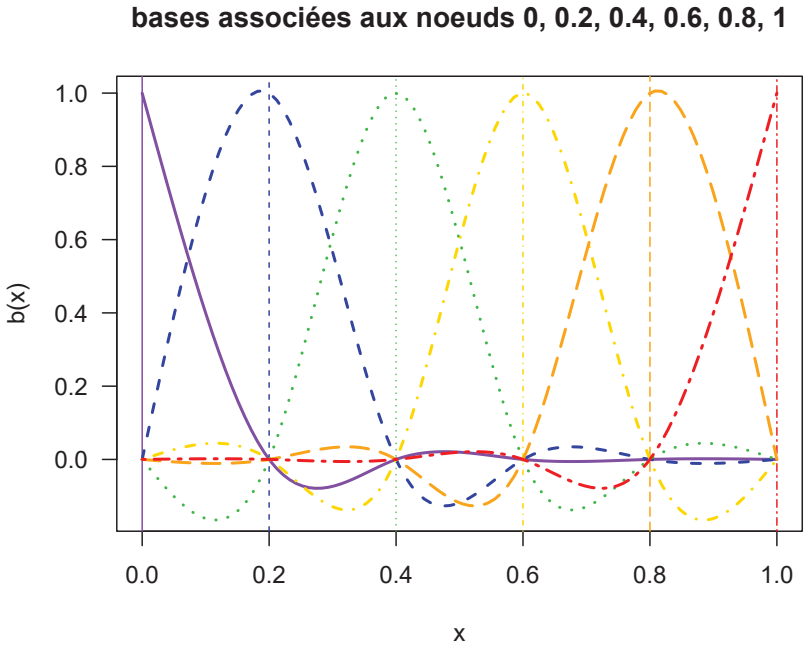

Figure 1.4 – Exemple de bases associées à une *cubic regression spline*

On remarque que chaque base vaut 1 sur son nœud associé et 0 sur tous les autres. Les bases de splines respectant cette propriété sont parfois appelées splines cardinales (Wood, 2017, Figure 5.4).

La spline en question s'écrit :

$$f(x) = \sum_{j=1}^{6} \beta_j b_j(x)$$

où les β_j sont des paramètres à estimer. Supposons que $\boldsymbol{\beta} = \begin{bmatrix} 0,5 \\ 0,9 \\ 0,7 \\ 0,5 \\ 0,55 \\ 0,6 \end{bmatrix}$

La figure 1.5 montre la spline f obtenue ainsi que ses bases multipliées par leur paramètre β_j respectif. Comme annoncé plus haut, on observe que la valeur de f au j^e nœud correspond bien à β_j.

Projection au-delà des nœuds extérieurs

Les *cubic regression splines* sont des splines dites naturelles, c'est à dire qu'elles sont linéaires en leurs nœuds extérieurs. On peut donc facilement extrapoler les valeurs de la spline en deçà du premier nœud x_1 et au-delà du dernier nœud x_k. En effet, il suffit de réaliser les extrapolations linéaires suivantes :

$$\begin{aligned} \text{si} \quad x < x_1, \quad & f(x) = f(x_1) + f'(x_1)(x - x_1) \\ \text{si} \quad x > x_k, \quad & f(x) = f(x_k) + f'(x_k)(x - x_k) \end{aligned}$$

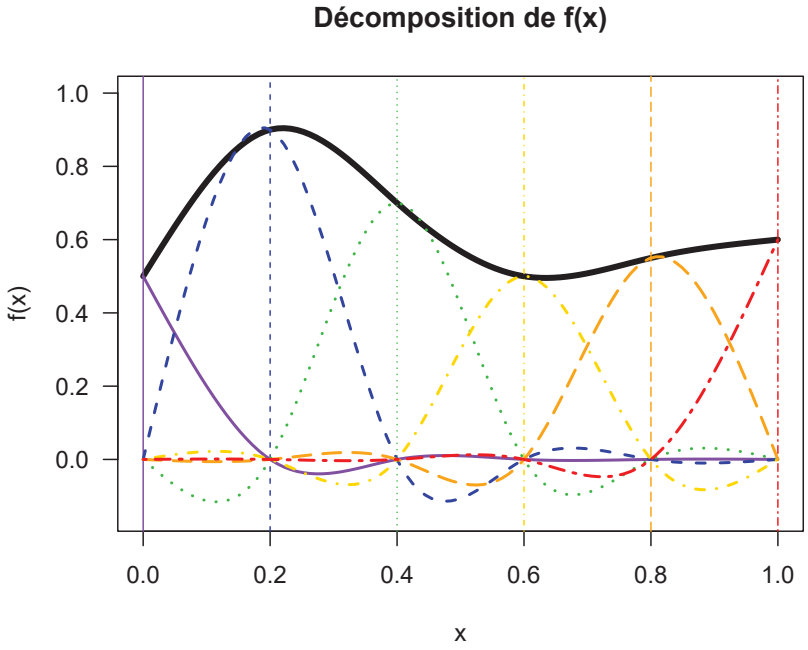

Figure 1.5 – Construction d'une *cubic regression spline*

1.3 Nombre et position des nœuds

Les splines de régression offrent une flexibilité très appropriée pour modéliser les effets des variables explicatives dans un modèle statistique. Leur forme entièrement paramétrique en font également un outil idéal pour l'inférence. Toutefois, l'utilisation des splines de régression s'accompagne des deux interrogations suivantes :

— Combien de nœuds choisir ?
— Où les positionner ?

Du point de vue de l'inférence, il est conseillé de choisir les nœuds a priori (voir le chapitre 3 sur la sélection de modèles). Pour ce qui est du nombre de nœuds, il convient de définir le nombre de degrés de liberté souhaité. Pour la position, en règle générale on dispose de très peu d'informations a priori, il est alors souhaitable d'utiliser les quantiles de la covariable concernée (Herndon and Harrell, 1990, 1995).

Chapitre 2

Modèles de survie

2.1 Généralités

L'ANALYSE DE SURVIE s'intéresse à l'étude du temps qui s'écoule jusqu'à ce qu'un événement particulier se produise. Il peut s'agir du temps jusqu'au décès d'une personne, jusqu'à la défaillance d'un système électronique, ou jusqu'à la prochaine éclaircie.
La principale difficulté de l'analyse de survie réside dans le phénomène de censure. Supposons que l'on s'intéresse à la survie d'un groupe de personnes choisies au hasard parmi tous les Français. Nous connaissons le nombre de personnes au début de notre étude, la date de début de notre étude mais nous devons également prévoir une date de fin à notre étude. Supposons que nous ayons les moyens de suivre tout le groupe pendant 30 ans. Il est alors fort probable qu'au bout de 30 ans, nous n'ayons pas observé le décès chez tous les individus. Si des personnes ont 20 ans au début de l'étude, elles n'auront que 50 ans à la fin et auront de grandes chances d'être encore en vie. Pour ces personnes, on dit que l'information est censurée à droite, c'est à dire que, tout ce que l'on sait, c'est que le temps de décès est supérieur à 30 ans. Ainsi, nous n'observons pas la réalisation de notre variable d'intérêt pour tous les individus.

2.2 Formalisme et indicateurs

En analyse de survie, pour un individu statistique i, on s'intéresse à la variable aléatoire positive T_i représentant le temps de survenue d'un événement (généralement le décès).

Toutefois, en présence d'une censure C_i, T_i n'est pas observable directement. Dans ce cas, on observe $T_i^* = min(T_i, C_i)$.

Ainsi, pour un individu i, on dispose :
- d'un temps d'observation t_i, réalisation de la variable T_i^*
- d'un indicateur de décès δ_i (1 si t_i correspond à un temps de décès et 0 sinon)

La fonction de survie en t associée à la variable T_i est la probabilité que T_i soit supérieur à t :

$$S(t) = P(T_i > t)$$

La fonction de densité de T_i est :

$$f(t) = \lim_{dt \to 0} \frac{P(t \leq T_i < t + dt)}{dt} = -S'(t)$$

Le taux de mortalité instantané (*hazard* en anglais) est défini par :

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T_i < t + dt | T_i > t)}{dt} = \frac{f(t)}{S(t)} \tag{2.1}$$

Le taux cumulé est défini comme l'intégrale du taux instantané :

$$H(t) = \int_0^t h(u)du$$

Finalement, la survie peut s'exprimer en fonction du taux de la manière suivante :

$$S(t) = exp\left[-H(t)\right] = exp\left(-\int_0^t h(u)du\right) \tag{2.2}$$

2.3 Estimateurs non-paramétriques

2.3.1 Estimateur de Kaplan-Meier

Un estimateur non-paramétrique de la survie asymptotiquement sans biais en présence de censure à droite fut proposé en 1958 (Kaplan and Meier, 1958).

À partir des temps de décès observés $t_1 \leq \cdots \leq t_N$, l'estimateur de Kaplan-Meier estime la survie conditionnelle au temps t_i, notée $\hat{s}_i$, sur l'intervalle $[t_i; t_{i+1}[$ de la manière suivante :

$$\hat{s}_i = \frac{n_i - d_i}{n_i}$$

avec n_i le nombre d'individus à risque juste avant le temps t_i et d_i le nombre de décès au temps t_i. La caractéristique fondamentale de cet estimateur est la manière dont il prend en compte des individus censurés au sein de la population à risque. Ainsi, pour passer de n_i à n_{i+1}, il faut non seulement retirer les décès d_i mais également tous les individus censurés entre t_i et t_{i+1}.

Si l'on cumule les survies conditionnelles depuis le début du suivi jusqu'au temps t alors l'estimation de la survie au temps t est :

$$\hat{S}(t) = \prod_{i:t_i \leq t} \hat{s}_i = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i}$$

La formule de Greenwood est souvent utilisée comme estimateur de la variance :

$$Var\left(\hat{S}(t)\right) \approx \hat{S}(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

Exemple de calcul d'une courbe de survie

Supposons que l'on dispose des temps de suivi de 12 individus : 1, 2, 2, 2*, 4, 4*, 5*, 6, 6, 6*, 7*, 7*.

L'astérisque indique un temps de censure tandis que l'absence d'astérisque indique un temps de décès.

Pour illustrer l'estimateur de Kaplan-Meier, nous pouvons construire le tableau 2.1 à partir des temps de décès :

temps de décès	nombre de décès	nombre d'individus à risque	survie conditionnelle	survie
1	1	12	1 - 1/12	91,7%
2	2	11	1 - 2/11	75,0%
4	1	8	1 - 1/8	65,6%
6	2	5	1 - 2/5	39,4%

Tableau 2.1 – Estimateur de Kaplan-Meier

Interprétons le tableau 2.1 ligne à ligne :
Au temps 1 il y a un seul décès donc il n'y a pas de problème particulier ici. La probabilité de décès vaut 1/12 et celle de survie vaut 1 - 1/12.
Au temps 2, il reste donc 12 - 1 = 11 individus à risque. Il y a deux décès et une censure. La probabilité de décès conditionnelle vaut donc 2/11.
Au temps 4, il reste maintenant 11 - (2+1) = 8 individus à risque car au temps 2 nous avons eu deux décès et une censure (donc 3 sorties en tout).
Au temps 6, il reste maintenant 8 - (1+2) = 5 individus à risque. Attention ici car il y a eu un décès et une censure au temps 4 ainsi qu'une **censure au temps 5** donc on a bien 3 sorties au total. Par construction, l'estimateur de Kaplan-Meier se sert des temps de décès pour construire son estimation, c'est la raison pour laquelle nous ne recalculons pas la survie au temps 5 car il ne s'agit que d'un temps de censure. Ici on a donc $\hat{S}(5) = \hat{S}(4)$

La figure 2.1 donne la courbe de survie associée au tableau 2.1.

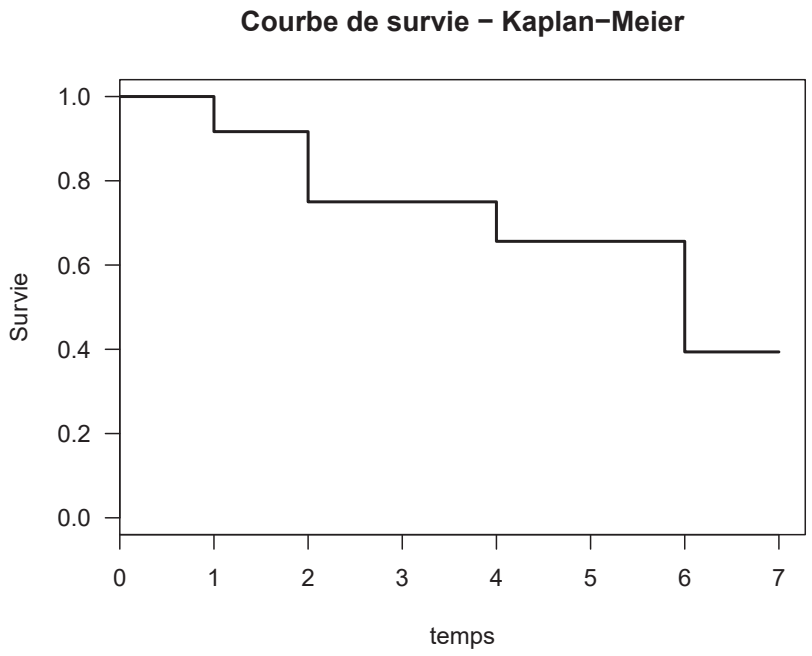

Figure 2.1 – Fonction de survie obtenue par l'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est certainement le plus utilisé en analyse de survie. Toutefois, dès que l'on souhaite prendre en compte les effets de covariables, son utilisation est limitée. En revanche, étant

donné qu'il est asymptotiquement sans biais, il peut être utilisé comme référence afin de contrôler les prédictions d'un modèle paramétrique.

2.3.2 Estimateur de Nelson-Aalen

Une alternative à l'estimateur de Kaplan-Meier consiste à estimer le taux de mortalité cumulé à chaque temps d'événement de la manière suivante :

$$\hat{H}(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i}$$

Cet estimateur, dit de Nelson-Aalen, permet alors d'estimer la survie au temps t :

$$\hat{S}(t) = exp\left(-\hat{H}(t)\right)$$

À partir des données de la section précédente, on peut construire le tableau 2.2.

temps de décès	nombre de décès	nombre d'individus à risque	taux	taux cumulé $\hat{H}(t)$	survie
1	1	12	1/12	1/12	92,0%
2	2	11	2/11	$\hat{H}(1) + 2/11$	76,7%
4	1	8	1/8	$\hat{H}(2) + 1/8$	67,7%
6	2	5	2/5	$\hat{H}(3) + 2/5$	45,4%

Tableau 2.2 – Estimateur de Nelson-Aalen

Nota Bene : il existe des variantes de cet estimateur qui diffèrent dans leur manière de gérer les ex æquo. Par exemple, au temps de décès 2, au lieu d'ajouter 2/11 au taux cumulé on pourrait ajouter $1/11 + 1/10$ (Fleming and Harrington, 1984).

2.4 Interprétation et intérêt du taux

Le modèle le plus utilisé en analyse de survie, à savoir le modèle de Cox (Cox, 1972), considère le taux de mortalité de base (c'est à dire la partie du taux dépendant uniquement du temps) comme un paramètre de nuisance. Il s'agit d'une technique astucieuse pour contourner les contraintes numériques de l'époque qui a permis l'analyse de nombreuses bases de données de survenue d'événements. Néanmoins, ces contraintes numériques n'existent plus et le fait de ne pas avoir accès au taux de base constitue aujourd'hui un défaut majeur dans bon nombre d'études. En effet, la notion de taux de mortalité instantané (ou plus simplement, la notion de taux) est essentielle en survie. En épidémiologie notamment, le taux intéresse fortement les cliniciens. Tandis que la survie offre une vision cumulée de la mortalité, le taux instantané restitue la force de mortalité qui s'applique à chaque instant. Il est donc très important de pouvoir restituer à la fois la survie et le taux.

2.4.1 Intérêt du taux par rapport à la survie

La figure 2.2 montre trois courbes de survie et les trois courbes de taux correspondantes. Comme annoncé en introduction, la courbe du taux sera appelée « dynamique du taux » dans la suite de

cette thèse car cette appellation permet de parler de façon concise « d'évolution au cours du temps du taux » et d'insister sur la notion de mouvement. Cette figure 2.2 illustre le fait que des courbes de survie très semblables peuvent correspondre à des dynamiques du taux très différentes.

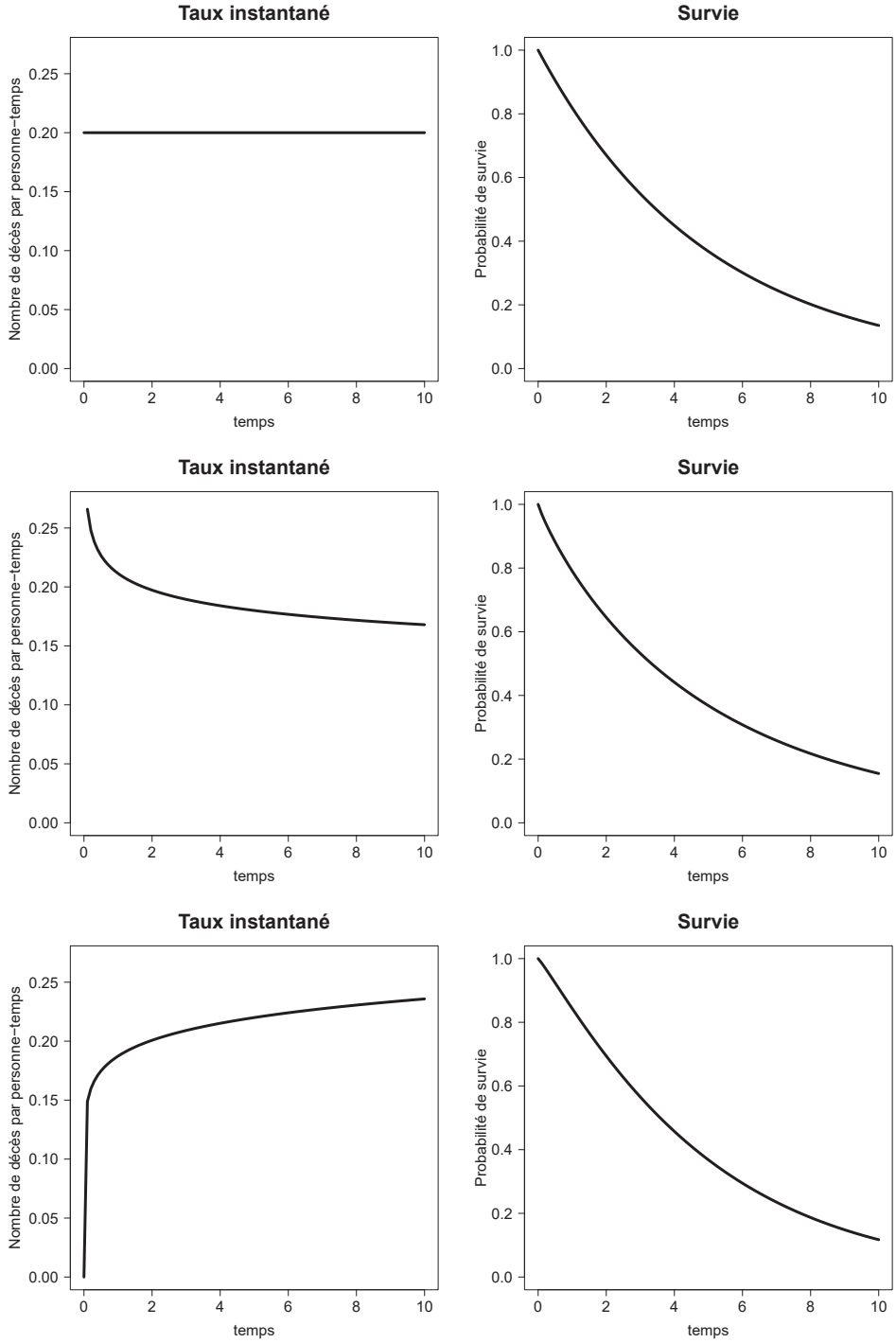

Figure 2.2 – Comparaison de courbes de taux et de courbes de survie. La première ligne correspond à une loi exponentielle de paramètre $\lambda = 0,2$. La deuxième ligne à une loi Weibull de paramètres $\lambda = 0,2$ et $\gamma = 0,9$. La troisième ligne à une loi Weibull de paramètres $\lambda = 0,2$ et $\gamma = 1,1$.

On se rend alors compte que la restitution du taux est essentielle. La courbe de survie est souvent plus difficile à lire que la courbe du taux en raison des contraintes auxquelles elle est soumise. En effet, la courbe de survie est nécessairement décroissante et comprise entre 0 et 1, et il est souvent compliqué d'apprécier la force de mortalité que subissent les individus à chaque instant à partir de la courbe de survie. De même, le taux cumulé H ne décrit pas la force subie par les individus car il est nécessairement croissant et compris entre 0 et $+\infty$. Quant au taux de mortalité instantané, bien qu'il doive être positif, il n'est pas contraint à la monotonie. Ainsi, ses valeurs extrêmes et sa dynamique sont autant d'informations précieuses sur la mortalité qui agit sur la population étudiée.

2.4.2 Le taux comme densité conditionnelle

Pour rappel, le taux de mortalité instantané associé à une variable aléatoire positive T est défini de la manière suivante :

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T > t)}{dt} = \frac{f(t)}{S(t)}$$

Il s'agit donc de la densité de la variable aléatoire T conditionnellement au fait que $T > t$. Cette manière de se représenter le taux permet de mieux comprendre son intérêt. En effet, lorsque qu'un médecin s'intéresse aux chances de survie d'un patient par exemple, c'est toujours conditionnellement au fait que son patient a survécu jusque-là. Les questions que le médecin se posent sont souvent conditionnelles : sachant que mon patient a 40 ans aujourd'hui, quelles sont ces chances de survie dans l'année qui suit ? S'il atteint l'âge de 70 ans, quelles seront alors ses chances ?

2.4.3 L'unité du taux

La définition du taux (2.1) indique que celui-ci dépend de l'unité de temps considéré (jour, mois, année, etc). On dit que le taux a une dimension (contrairement à une probabilité par exemple, qui est sans dimension).
Lorsque l'on parle de taux, il est donc essentiel de préciser son unité. Une conséquence importante est que le taux peut être supérieur à 1 puisqu'il ne s'agit pas d'une probabilité.

Exemple : Que signifie un taux constant de 2 décès par personne-année ?

Puisqu'il s'agit d'un taux constant, les survies à 1 et 2 ans s'écrivent :

$$S(1) = exp(-2) = 13,53\% \qquad S(2) = exp(-4) = 1,83\%$$

La probabilité de mourir la première année vaut ainsi : $1 - S(1) = 86,47\%$. Tandis que la probabilité de mourir la deuxième année sachant qu'on a survécu la première vaut :

$$\frac{S(1) - S(2)}{S(1)} = \frac{13,53\% - 1,83\%}{13,53\%} = 86,47\%$$

De manière générale, si l'on a un taux de mortalité constant λ, alors la probabilité de mourir dans l'unité de temps suivante sachant qu'on survécu jusque-là est :

$$\frac{S(t) - S(t+1)}{S(t)} = \frac{exp(-\lambda t) - exp(-\lambda(t+1))}{exp(-\lambda t)} = 1 - exp(-\lambda)$$

$$\frac{S(t) - S(t+1)}{S(t)} \underset{\lambda \approx 0}{\approx} \lambda$$

Lorsque le taux est proche de zéro, sa valeur se rapproche de celle de la probabilité conditionnelle de décès dans l'intervalle considéré.
Ainsi, pour retrouver une approche intuitive du taux de mortalité (dans le sens où sa valeur est proche de la probabilité conditionnelle de décès) il faut changer le pas de temps afin de réduire la valeur du taux et pouvoir faire l'approximation ci-dessus.

Par exemple, 2 décès par personne-année
$= 2/12 = 16{,}67\%$ décès par personne-mois $\approx 1 - exp(-\frac{2}{12}) = 15,35\%$ soit une probabilité de décès dans le mois de $15,35\%$
$= 2/365 = 0{,}548\%$ décès par personne-jour $\approx 1 - exp(-\frac{2}{365}) = 0,546\%$ soit une probabilité journalière de décès de $0,546\%$

2.4.4 La dynamique du taux

Taux constant

La dynamique la plus simple possible que l'on puisse imaginer pour le taux de mortalité est qu'il soit constant quel que soit le temps de suivi. Cette dynamique particulière est associée à la loi exponentielle. En effet, si le temps T jusqu'à un certain événement suit une loi exponentielle de paramètre λ, c'est à dire $T \sim \mathcal{E}(\lambda)$, alors :

$$h(t) = \lambda \qquad S(t) = exp\,(-\lambda t)$$

Un taux constant signifie simplement que le temps qui passe n'augmente ni ne diminue mes chances de voir l'événement se produire **sachant qu'il ne s'est pas produit jusque-là** : on parle de **phénomène sans mémoire**. Un exemple de temps jusqu'à événement associé à un taux constant pourrait donc être le temps que va mettre un individu à gagner au Loto. En effet, le fait d'avoir joué et perdu chaque semaine pendant 1, 5, 10 ou 30 ans ne change absolument rien à la probabilité de gagner la semaine suivante.

Taux croissant

Un taux croissant est synonyme d'une augmentation au cours du temps du risque de voir se présenter l'événement sachant qu'il ne s'est toujours pas produit. Ce phénomène peut être illustré en s'intéressant au temps jusqu'à la défaillance d'une pièce d'usinage ou d'une machine (en supposant que l'on a atteint une phase d'usure). La pièce ou la machine en question va présenter une usure naturelle et quelle que soit sa robustesse on sait qu'elle finira par défaillir. Ainsi, plus la pièce ou la machine sera utilisée, plus celle-ci va s'user avec le temps et plus la probabilité d'observer une défaillance le jour suivant va augmenter.

Taux décroissant

Un taux décroissant implique une diminution au cours du temps du risque de voir se présenter l'événement sachant qu'il ne s'est toujours pas produit. Supposons que l'on s'intéresse à la mortalité infantile. On sait que les nourrissons et les enfants en bas-âge sont plus vulnérables (accidents

domestiques, maladies) que les enfants sur le point d'entrer dans l'adolescence. Ainsi, de la naissance jusqu'à l'adolescence, le taux de mortalité diminue. De même, la probabilité de retrouver une personne disparue diminue avec le temps.

Dynamique plus complexe

Évidemment, la dynamique du taux ne se résume pas aux trois cas : taux constants, taux croissants et taux décroissants. En effet, le taux n'a pas de raison d'être monotone en toutes circonstances. Toutefois, ces trois cas de dynamiques simples permettent d'expliquer comment il est possible d'obtenir des formes plus complexes et notamment non monotones (voir la section 2.5 sur le mélange de taux).

2.5 Mélange de taux

2.5.1 Mélange de deux taux constants

Supposons que l'on étudie une population composée de deux groupes. Le groupe 1 présente un taux instantané constant de $\lambda_1 = 0,6$ événements par personne-temps tandis que celui du groupe 2 est $\lambda_2 = 0,1$. On suppose également, qu'initialement, il y a autant de personnes du groupe 1 que du groupe 2 dans la population d'étude. La figure 2.3 donne la dynamique du taux dans les deux groupes (en pointillé) et pour la population totale (trait plein).

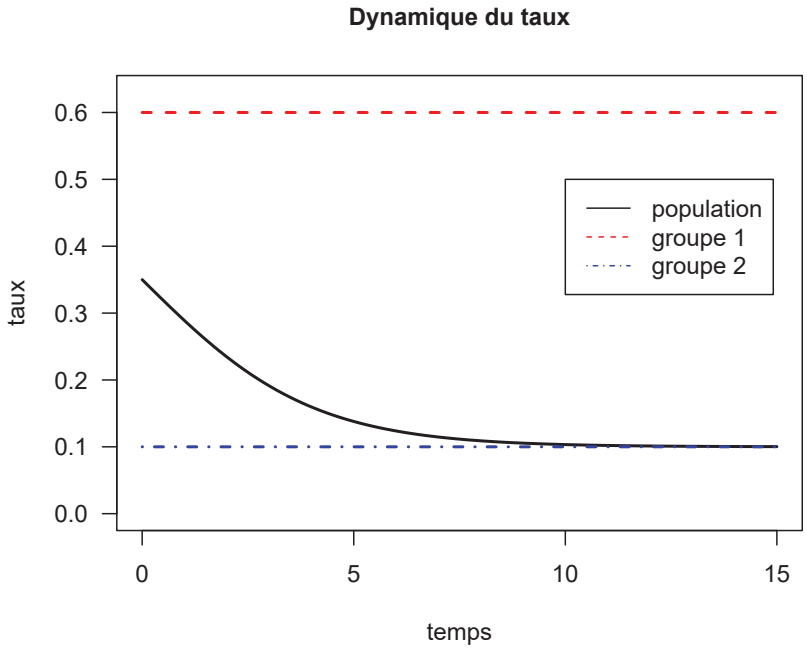

Figure 2.3 – Mélange de deux taux constants

On s'aperçoit qu'en 0, le taux de la population ou « taux marginal » correspond logiquement à la moyenne des taux des deux groupes. Cependant, on observe une diminution progressive jusqu'à ce que le taux marginal tende asymptotiquement vers celui du groupe 2.

Ce phénomène s'explique assez bien si l'on décrit la proportion que représentent les individus du groupe 1 par rapport à toute la population au cours du temps.

Notons $p(t)$ la proportion d'individus du groupe 1 par rapport à la population totale au temps t. Au départ (en $t = 0$), cette proportion vaut $p(0) = 0,5$ car on a supposé qu'il y avait autant de personnes du groupe 1 que du groupe 2. Si l'on note $S_1(t)$ et $S_2(t)$ les survies au temps t dans les groupes 1 et 2 respectivement, alors il vient :

$$p(t) = \frac{p(0)S_1(t)}{p(0)S_1(t) + [1 - p(0)]S_2(t)}$$

Et le taux $h_{pop}(t)$ de la population au temps t s'écrit :

$$h_{pop}(t) = \lambda_1 p(t) + \lambda_2(1 - p(t))$$

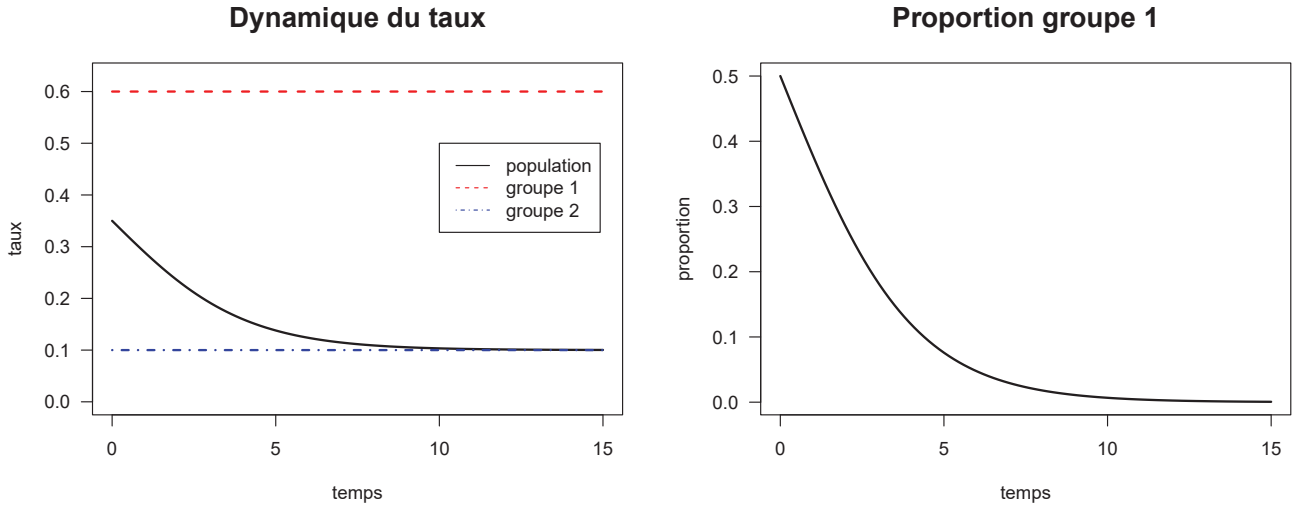

Figure 2.4 – Mélange de deux taux constants, proportion du groupe1

La figure 2.4 présente le même mélange que la figure 2.3 en y ajoutant l'évolution de la proportion d'individus du groupe 1 au cours du temps. Ainsi, à partir de 10 ans environ, les individus du groupe 1 sont presque tous décédés et le taux populationnel correspond au taux des individus restants (i.e ceux du groupe 2).

2.5.2 Mélange de deux taux croissants

De la même manière, la figure 2.5 présente le mélange de deux taux croissants.

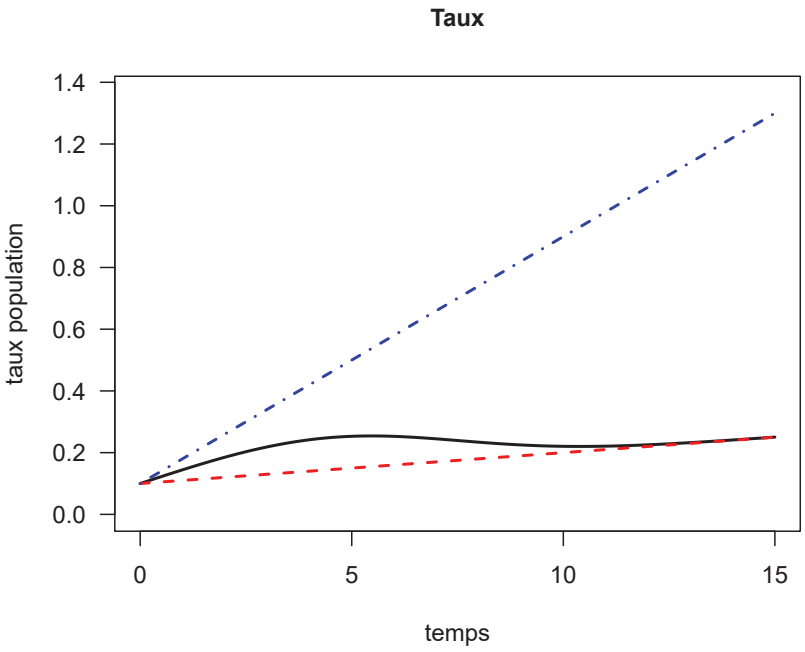

Figure 2.5 – Mélange de deux taux croissants

2.5.3 Mélange d'un taux constant et d'un taux croissant

La figure 2.6 montre le mélange d'un taux constant et d'un taux croissant.

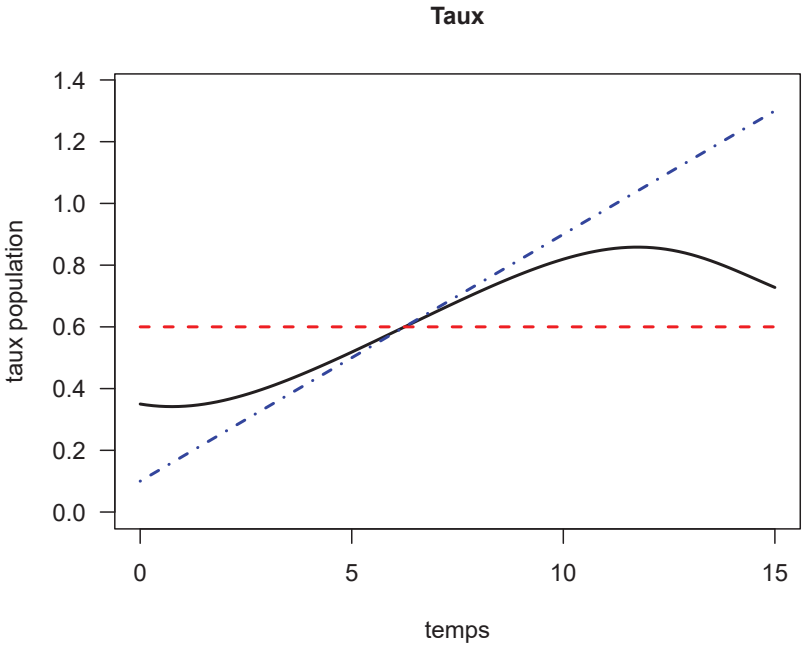

Figure 2.6 – Mélange d'un taux constant et d'un taux croissant

En résumé, l'**hétérogénéité** de la population conduit à un phénomène de **sélection** : la population survivante diffère de la population initiale (Vaupel and Yashin, 1985).

Étant donné que les populations étudiées sont très souvent constituées de mélanges hétérogènes d'autres populations, la non-monotonie du taux est un phénomène presque omniprésent en survie.

2.6 Vraisemblance

Pour un individu i, supposons que la loi de T_i est paramétrée par le vecteur $\boldsymbol{\beta}$. On observe une réalisation (t_i, δ_i) de $T_i^* = min(T_i, C_i)$. On note G la fonction de survie de C_i et g sa densité.

Quelle est la contribution de l'individu i à la vraisemblance ?

Si $\delta_i = 1$ alors :

$$\mathcal{L}_i(\boldsymbol{\beta}) = \lim_{dt \to 0} P(T_i \in [t_i; t_i + dt],\, C_i > T_i | \boldsymbol{\beta}) = f(t_i|\boldsymbol{\beta})G(t_i)$$

Si $\delta_i = 0$ alors :

$$\mathcal{L}_i(\boldsymbol{\beta}) = \lim_{dt \to 0} P(C_i \in [t_i; t_i + dt],\, C_i < T_i | \boldsymbol{\beta}) = g(t_i)S(t_i|\boldsymbol{\beta})$$

Il vient donc, dans le cas général :

$$\mathcal{L}_i(\boldsymbol{\beta}) = \left[f(t_i|\boldsymbol{\beta})G(t_i)\right]^{\delta_i} \left[g(t_i)S(t_i|\boldsymbol{\beta})\right]^{1-\delta_i}$$

Attention ici, T_i et C_i doivent être supposés indépendants. Si cette hypothèse est faite et étant donné que G et g ne dépendent pas de $\boldsymbol{\beta}$, la vraisemblance devient :

$$\mathcal{L}_i(\boldsymbol{\beta}) = f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = h(t_i)^{\delta_i} S(t_i)$$

$$\mathcal{L}_i(\boldsymbol{\beta}) = h(t_i)^{\delta_i} exp\left(-\int_0^{t_i} h(u)\mathrm{d}u\right)$$

En règle générale, on travaille sur la log-vraisemblance :

$$l_i(\boldsymbol{\beta}) = \delta_i log\left[h(t_i)\right] - \int_0^{t_i} h(u)\mathrm{d}u \tag{2.3}$$

La prise en compte de données tronquées à gauche (ou entrées tardives) est possible en considérant le temps d'entrée $t_{0,i}$. La formule (2.3) devient alors :

$$l_i(\boldsymbol{\beta}) = \delta_i log\left[h(t_i)\right] - \int_{t_{0,i}}^{t_i} h(u)\mathrm{d}u$$

Les méthodes exposées dans cette thèse s'appliquent également aux données tronquées à gauche. Toutefois, par simplicité, la suite de cette thèse s'appuiera sur la formule (2.3).

Dans un modèle de survie, la première problématique réside dans le calcul de l'intégrale présente dans (2.3).

2.7 Modèles de taux

Le taux étant notre indicateur d'intérêt, il est naturel de proposer des modèles sur cette échelle. De plus, on a vu dans la section 2.4 que l'échelle du taux était la moins contrainte pour restituer l'information sur des données de survie. En effet, seule la contrainte de positivité s'applique. Ainsi, en modélisant non pas le taux mais le logarithme du taux, on peut construire un modèle dont l'estimation des paramètres se fera sans contrainte. Dans la littérature, on peut trouver des modélisations sur le logarithme du taux cumulé (Liu et al., 2018) mais il est alors très contraignant de garantir à la fois la croissance du taux cumulé et la positivité du taux instantané (des méthodes de pénalisation sont utilisées ce qui alourdit clairement le processus d'optimisation). En outre, en modélisant le taux cumulé, il faut toujours pouvoir estimer le taux instantané afin de calculer la vraisemblance (comme vu au-dessus). Ainsi, le problème d'intégration du taux instantané disparaît mais se transforme en problème de dérivation du taux cumulé. Or les techniques de dérivation numérique sont beaucoup plus instables que celles d'intégration numérique, notamment en termes de sensibilité aux erreurs d'arrondi.

De nombreux arguments nous indiquent donc que la bonne échelle de modélisation est le logarithme du taux instantané.

2.7.1 Spécification du modèle

Pour des raisons de lisibilité, les lettres minuscules en gras (e.g., $\boldsymbol{\lambda}$ et $\boldsymbol{\beta}$) indiqueront des vecteurs tandis que les lettres majuscules en gras (e.g., $\boldsymbol{S}$ et $\boldsymbol{X}$) désigneront des matrices. La i^e ligne d'une matrice $\boldsymbol{M}$ est notée $\boldsymbol{M}_i$ et l'élément i, j est noté M_{ij}. Le i^e élément d'un vecteur $\boldsymbol{v}$ est noté v_i. Pour construire un modèle de taux, on a besoin de spécifier les effets propres du temps t et de chaque covariable x_j (parmi un vecteur de covariables $\boldsymbol{x}$) mais également les effets conjoints (interactions). La forme générale du modèle est la suivante :

$$log\{h(t, \boldsymbol{x})\} = f(t, \boldsymbol{x})$$

La fonction multidimensionnelle f peut prendre de multiples formes. Par exemple, dans un modèle de taux contenant uniquement le temps t et l'âge a, nous pourrions proposer le modèle suivant :

$$mod1 : log\{h(t, a)\} = \beta_0 + \beta_t \times t + \beta_a \times a$$

Les effets propres du temps et de l'âge sont ici linéaires et il n'y a pas d'interactions entre les deux. Une interaction linéaire est facilement intégrée de la manière suivante :

$$mod2 : log\{h(t, a)\} = \beta_0 + \beta_t \times t + \beta_a \times a + \beta_{t,a} \times t \times a$$

Une fois les covariables connues, spécifier le modèle revient à choisir deux choses :

1. la forme fonctionnelle du temps et des covariables continues
2. la structure d'interactions

Soit n le nombre d'individus à analyser, $\boldsymbol{t}$ le vecteur de taille n contenant les temps de suivi de tous les individus et $\boldsymbol{\delta}$ le vecteur de taille n contenant les indicatrices de décès de tous les individus ($\delta_i = 1$ quand l'individu i est décédé, $\delta_i = 0$ lorsqu'il est censuré). Soit $\boldsymbol{X}(\boldsymbol{t})$ la matrice de design, alors le modèle peut s'écrire :

$$log\{h(t_i; \boldsymbol{x}_i)\}_{1 \leq i \leq n} = \boldsymbol{X}(\boldsymbol{t})\boldsymbol{\beta}$$

2.7.2 Vraisemblance du modèle

Comme vu en 2.6, en l'absence de censure informative, la contribution à la log-vraisemblance de l'individu i s'écrit :

$$l_i(\boldsymbol{\beta}) = \delta_i log\left[h(t_i; \boldsymbol{x}_i)\right] - \int_0^{t_i} h(u; \boldsymbol{x}_i)du \tag{2.4}$$

La log-vraisemblance de tous les individus est donc :

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta})$$

2.7.3 Approximation du taux cumulé

Calculer l'intégrale du taux présente dans la log-vraisemblance (c'est à dire le taux cumulé au temps t_i) est un problème spécifique aux modèles de survie. Cette intégrale peut être approchée par la quadrature de Gauss-Legendre, qui nécessite : i) q, le nombre de nœuds associés à la quadrature ; ii) $(d_k^*)_{(1\leq k\leq q)}$, le vecteur de nœuds de longueur q pour une intégration sur $[-1; +1]$; et, iii) $(w_k^*)_{(1\leq k\leq q)}$ le vecteur de poids de longueur q pour une intégration sur $[-1; +1]$. Pour un individu i, les nœuds $(d_k^i)_{(1\leq k\leq q)}$ et poids $(w_k^i)_{(1\leq k\leq q)}$ correspondant à une intégration sur $[0; t_i]$ sont donnés par : $d_k^i = \frac{t_i}{2}d_k^* + \frac{t_i}{2}$ et $w_k^i = \frac{t_i}{2}w_k^*$. La contribution à la log-vraisemblance devient :

$$l_i(\boldsymbol{\beta}) \approx \delta_i log\left[h(t_i; \boldsymbol{x}_i)\right] - \sum_{k=1}^{q} h(d_k^i; \boldsymbol{x}_i)$$

D'un point de vue calculatoire, en plus de la matrice de design du modèle, nous pouvons construire les q matrices de design de dimensions (n, p) suivantes : $\boldsymbol{GL}^k = \boldsymbol{X}(\frac{\boldsymbol{t}}{2}d_k^* + \frac{\boldsymbol{t}}{2})$. Le vecteur des temps de suivi $\boldsymbol{t}$ est ainsi remplacé par le vecteur $\frac{\boldsymbol{t}}{2}d_k^* + \frac{\boldsymbol{t}}{2}$ de longueur n. De manière similaire, les q vecteurs de poids $\boldsymbol{w}^k$ de longueur n sont construits tels que $\boldsymbol{w}^k = \frac{\boldsymbol{t}}{2}\boldsymbol{w}_k^*$. Au final,

$$l_i(\boldsymbol{\beta}) \approx \delta_i \boldsymbol{X}_i\boldsymbol{\beta} - \sum_{k=1}^{q} w_i^k exp\left(\boldsymbol{GL}_i^k\boldsymbol{\beta}\right) \tag{2.5}$$

où $\boldsymbol{X}_i$ et $\boldsymbol{GL}_i^k$ sont des vecteurs de longueur p qui correspondent respectivement à la i^e ligne de $\boldsymbol{X}(\boldsymbol{t})$ et de $\boldsymbol{GL}^k$. Le choix de q est crucial car il représente le compromis entre la précision et le temps d'exécution.

D'après Charvat et al. (2016) et à des simulations (voir annexe A), $q = 20$ semble être une valeur raisonnable.

2.8 Approche de Poisson pour les modèles de survie

Dans cette section nous allons faire un rappel sur les modèles de Poisson et présenter une approche permettant d'ajuster des modèles de taux en utilisant une équivalence entre vraisemblance d'un modèle de taux et celle d'un modèle de Poisson.

2.8.1 Modèle de Poisson classique

On dispose d'un vecteur aléatoire à expliquer $\boldsymbol{Y}$ dont les composantes sont indépendantes et suivent une loi de Poisson, $Y_i \sim \mathcal{P}(\mu_i)$. Le modèle classique (non pénalisé) s'écrit :

$$\log(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta} = \eta_i$$

Pour estimer les coefficients $\boldsymbol{\beta}$, nous écrivons la log-vraisemblance du modèle :

$$l(\boldsymbol{\beta}) = \sum_i \log\left(\frac{\mu_i^{y_i}}{y_i!}e^{-\mu_i}\right)$$

$$l(\boldsymbol{\beta}) = \sum_i y_i \log(\mu_i) - \log(y_i!) - \mu_i$$

Pour toutes les composantes j du vecteur de paramètres $\boldsymbol{\beta}$, on doit résoudre :

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = 0 \Leftrightarrow \sum_i \frac{y_i - \mu_i}{\mu_i}\frac{\partial \mu_i}{\partial \beta_j} = 0$$

Ce problème est équivalent à résoudre des moindres carrés pondérés. En effet, cela revient à minimiser la fonction suivante :

$$MCP = \sum_i \frac{(y_i - \mu_i)^2}{\mu_i}$$

On peut résoudre ce problème en procédant de manière itérative et en considérant le dénominateur μ_i comme des poids connus à chaque itération. En définissant, à chaque itération k, la matrice diagonale $\boldsymbol{V}_{[k]}$ telle que $V_{[k]ii} = {\mu_i}^{[k]}$, on doit minimiser :

$$MCP = \left\|\sqrt{\boldsymbol{V}_{[k]}^{-1}}[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})]\right\|^2$$

On effectue alors un développement de Taylor de $\boldsymbol{\mu}$ autour de $\hat{\boldsymbol{\beta}}^{[k]}$ (le vecteur de paramètres estimés à l'itération k), on obtient :

$$MCP = \left\|\sqrt{\boldsymbol{V}_{[k]}^{-1}}[\boldsymbol{Y} - \boldsymbol{\mu}^{[k]} - \boldsymbol{J}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{[k]})]\right\|^2$$

Avec $\boldsymbol{J}$ la matrice Jacobienne de $\boldsymbol{\mu}$.

$$J_{ij} = \left.\frac{\partial \mu_i}{\partial \beta_j}\right|_{\hat{\boldsymbol{\beta}}^{[k]}} = X_{ij}{\mu_i}^{[k]}$$

Si l'on pose $G_{ii} = \frac{1}{{\mu_i}^{[k]}}$, alors : $\boldsymbol{J} = \boldsymbol{G}^{-1}\boldsymbol{X}$

$$MCP = \left\|\sqrt{\boldsymbol{V}_{[k]}^{-1}}\boldsymbol{G}^{-1}[\boldsymbol{G}(\boldsymbol{Y} - \boldsymbol{\mu}^{[k]}) + \boldsymbol{eta}^{[k]} - \boldsymbol{X}\boldsymbol{\beta}]\right\|^2$$

$$MCP = \left\|\sqrt{\boldsymbol{W}_{[k]}}[\boldsymbol{z}^{[k]} - \boldsymbol{X}\boldsymbol{\beta}]\right\|^2$$

En posant $\boldsymbol{z}^{[k]} = \frac{\boldsymbol{Y} - \boldsymbol{\mu}^{[k]}}{\boldsymbol{\mu}^{[k]}} + \boldsymbol{\eta}^{[k]}$. La matrice diagonale des poids $\boldsymbol{W}_{[k]}$ est ici égale à $\boldsymbol{V}_{[k]}$ (cela n'est vrai que dans le cas du modèle de Poisson).

Algorithme IRLS (*Iteratively Reweighted Least Squares*)
À la première itération, on se donne un $\boldsymbol{\beta}^0$ puis on répète jusqu'à la convergence :

1. En utilisant $\boldsymbol{\beta}^{[k]}$, on peut calculer $\boldsymbol{\mu}^{[k]}, \boldsymbol{\eta}^{[k]} = \boldsymbol{X}\boldsymbol{\beta}^{[k]}, \boldsymbol{z}^{[k]}$ et $\boldsymbol{W}^{[k]}$
2. On minimise $MCP = \left\| \sqrt{\boldsymbol{W}_{[k]}}[\boldsymbol{z}^{[k]} - \boldsymbol{X}\boldsymbol{\beta}] \right\|^2$ par rapport à $\boldsymbol{\beta}$ pour trouver $\boldsymbol{\beta}^{[k+1]}$
3. On augmente k de 1.

Pour minimiser MCP on utilise tout simplement un modèle linéaire sur les pseudo-données pondérées.
Dans R, on utilise la commande $lm\left(\sqrt{\boldsymbol{W}_{[k]}}\boldsymbol{z}^{[k]} \sim \sqrt{\boldsymbol{W}_{[k]}}\boldsymbol{X}\right)$

2.8.2 Équivalence des vraisemblances entre modèle de Poisson et modèle de survie

Cette équivalence a été décrite pour la première fois par Friedman (1982).

L'idée principale derrière cette équivalence est le **découpage du temps de suivi** en petits intervalles de temps. Ce découpage implique une augmentation du jeu de données initial.

D'après (2.4), la contribution d'un individu i à la log-vraisemblance d'un modèle de survie s'écrit :

$$l_i(\boldsymbol{\beta}) = \delta_i log\left[h(t_i; \boldsymbol{x}_i)\right] - \int_0^{t_i} h(u; \boldsymbol{x}_i)du$$

Divisons les observations (t_i, δ_i) selon $n+1$ bandes de suivi arbitraires $[b_j; b_{j+1}[$ pour $j = 0, \ldots, n$ et avec $b_0 = 0$ et $b_{n+1} = t_i$.

La contribution à la log-vraisemblance de l'individu i sur la j^e bande de suivi est :

$$l_{i,j}(\boldsymbol{\beta}) = \delta_{i,j} log\left[h(t_i; \boldsymbol{x}_i)\right] - \int_{b_j}^{b_{j+1}} h(u; \boldsymbol{x}_i)du \tag{2.6}$$

Avec $\delta_{i,j} = \delta_i$ si $j = n$ et 0 sinon.

Si les bandes de suivi sont suffisamment nombreuses, on voit que l'intégrale peut être approchée par

$$\int_{b_j}^{b_{j+1}} h(u; \boldsymbol{x}_i)du \approx (b_{j+1} - b_j)h\left(c_{ij}; \boldsymbol{x}_i\right)$$

avec $c_{ij} = t_i$ si $j = n$ et $\frac{b_{j+1}+b_j}{2}$ sinon. On approche donc l'intégrale du taux par sa valeur au milieu de la bande de suivi multipliée par la longueur de la bande de suivi. Il s'agit de l'approche du point milieu, notamment utilisée dans les modèles de survie par Remontet et al. (2007).

Finalement, (2.6) devient :

$$l_{i,j}(\boldsymbol{\beta}) = \delta_{i,j} log\left[h(t_i; \boldsymbol{x}_i)\right] - (b_{j+1} - b_j)h\left(c_{ij}; \boldsymbol{x}_i\right)$$

Pour l'instant, nous avons utilisé l'augmentation du jeu de données afin d'approximer le taux cumulé et cette approximation est beaucoup plus coûteuse que la quadrature de Gauss-Legendre présentée plus haut. Pour vraiment comprendre l'intérêt de (2.6), supposons que $\delta_{i,j}$ soit la variable à expliquer dans un modèle de Poisson : $\delta_{i,j} \sim \mathcal{P}(\mu_{i,j})$. La contribution à la log-vraisemblance de l'individu (i, j) est :

$$l_{i,j}^{Poisson}(\boldsymbol{\beta}) = \delta_{i,j} \log(\mu_{i,j}) - \log(\delta_{i,j}!) - \mu_{i,j}$$

En posant $\mu_{i,j} = (b_{j+1} - b_j)h\left(c_{ij}; \boldsymbol{x}_i\right)$, il vient :

$$l_{i,j}^{Poisson}(\boldsymbol{\beta}) = \delta_{i,j} \log\left[h\left(c_{ij}; \boldsymbol{x}_i\right)\right] + \delta_{i,j} \log(b_{j+1} - b_j) - \log(\delta_{i,j}!) - (b_{j+1} - b_j)h\left(c_{ij}; \boldsymbol{x}_i\right)$$

En ne gardant que les termes qui dépendent de $\boldsymbol{\beta}$ on a :

$$l_{i,j}^{Poisson}(\boldsymbol{\beta}) = \delta_{i,j} \log\left[h\left(c_{ij}; \boldsymbol{x}_i\right)\right] - (b_{j+1} - b_j)h\left(c_{ij}; \boldsymbol{x}_i\right)$$

$\delta_{i,j}$ étant forcément nul excepté pour $j = n$ on peut même écrire :

$$l_{i,j}^{Poisson}(\boldsymbol{\beta}) = \delta_{i,j} \log\left[h(t_i; \boldsymbol{x}_i)\right] - (b_{j+1} - b_j)h\left(c_{ij}; \boldsymbol{x}_i\right)$$

Et finalement, $l_{i,j} = l_{i,j}^{Poisson}$ (à une constante près, mais vis-à-vis de l'estimation de $\boldsymbol{\beta}$, on peut les considérer égales). Ce résultat est très intéressant car il implique qu'un modèle de survie sur le taux peut être ajusté à partir de la machinerie dédiée aux modèles linéaires généralisés (GLM).

Nota Bene : C'est une équivalence de vraisemblance qui permet d'utiliser les techniques d'optimisation (algorithme IRLS) propres aux GLM mais la variable $\delta_{i,j}$ **ne suit pas une loi de Poisson**.

2.9 Survie nette et modèles de taux en excès

2.9.1 Décomposition du taux observé

En analyse de survie, l'excès de mortalité des patients par rapport à la mortalité d'une population de référence peut être plus informatif que la mortalité toutes causes. Par exemple, les oncologues souhaiteraient connaître l'impact des tumeurs sur la mortalité. Cependant, les causes de décès ne sont pas toujours disponibles (en particulier dans les données de registres de cancer) ou fiables et la responsabilité du cancer dans le décès est difficile à établir pour de longs temps de suivi (le cancer peut par exemple mener au suicide et certains traitements peuvent avoir une toxicité à long terme entraînant la mort des patients).
Dans la suite de cette section nous ferons référence au cancer comme la pathologie d'intérêt mais la logique fonctionne avec tout autre type de maladie chronique.
Le concept d'« excès de mortalité » représente ainsi une alternative pertinente au concept de « cause de décès », comme tentative d'isoler les décès supplémentaires attribuables (directement ou indirectement) au cancer. Cet excès de mortalité peut être estimé en supposant que, chez les patients atteints de cancer, la mortalité attribuable à toutes les causes sauf le cancer étudié peut être approximée par la mortalité toutes causes de la population générale (la « mortalité attendue »). D'après Estève et al. (1990), si l'on considère que le temps jusqu'au décès dû au cancer et celui dû aux autres causes sont indépendants, on peut écrire :

$$h_O(t, \boldsymbol{x}, \boldsymbol{z}) = h_E(t, \boldsymbol{x}) + h_P(a + t, \boldsymbol{z}) \tag{2.7}$$

Dans cette équation, h_O est la mortalité observée chez les patients atteints de cancer, h_E est l'excès de mortalité dû au cancer, t est le temps écoulé depuis le diagnostic, a est l'âge au diagnostic du cancer, $\boldsymbol{x}$ est un vecteur de covariables ayant un effet sur h_E, et h_P est la mortalité attendue à l'âge $a + t$ dans la population générale avec les caractéristiques démographiques $\boldsymbol{z}$. La survie nette, notée SN, est la survie qui serait observée si le cancer étudié était la seule cause de décès ; elle peut être obtenue directement à partir de h_E en utilisant la relation classique entre survie et taux de mortalité (2.2) :

$$SN(t, \boldsymbol{x}) = exp\left(-\int_0^t h_E(u, \boldsymbol{x})\mathrm{d}u\right)$$

Pour analyser les données de registre de cancer, les taux de mortalité attendus h_P sont obtenus via les données de mortalité de l'INSEE, déclinées par sexe, âge, année et département. Avant d'être intégrés dans le modèle, les taux attendus sont lissés via un modèle GAM incluant une spline bi-dimensionnelle pénalisée de l'âge et de l'année de décès, à partir de l'âge de 15 ans, séparément pour chaque département et pour chaque sexe.

2.9.2 Estimateur de la survie nette de Pohar-Perme

Afin d'estimer la survie nette, un équivalent de l'estimateur de Nelson-Aalen a été introduit par Perme et al. (2012). La mise en œuvre de l'estimateur de Pohar-Perme est toutefois autrement plus complexe que celle de Nelson-Aalen ou de Kaplan-Meier puisqu'elle fait appel à des notions de processus stochastiques.

Pour estimer la survie nette, on se place dans un monde hypothétique où seul le cancer étudié (ou une autre maladie chronique) peut mener au décès. Partant de ce postulat, l'idée originale de Perme et al. (2012) est la suivante : les patients qui décèdent des autres causes que le cancer dans le monde réel seraient en fait encore en vie et donc encore à risque de décéder du cancer étudié dans le monde hypothétique. Les décès dus aux autres causes induisent ainsi une censure informative lorsque l'on souhaite estimer la survie nette. Perme et al. (2012) proposent de réajuster le nombre de décès dus au cancer et le nombre d'individus à risque au temps t en pondérant par la survie attendue.

Si l'on ne tenait pas compte de cette censure informative, l'estimateur du taux de mortalité en excès cumulé $H_c(t)$ serait, à chaque temps d'événement ou de censure t :

$$\hat{H}_c(t) = \int_0^t \frac{\mathrm{d}N(u)}{Y(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i(u)\mathrm{d}H_{a,i}(u)}{Y(u)}$$

où n est le nombre d'individus, $H_{a,i}$ le taux cumulé attendu d'après les caractéristiques de l'individu i, $Y_i(u) = 1$ si l'individu est à risque de décéder au temps u et 0 sinon, $Y(u) = \sum_{i=1}^n Y_i(u)$, $N_i(u) = 1$ si l'individu est décédé au temps u et 0 sinon, $N(u) = \sum_{i=1}^n N_i(u)$.

Cependant, comme vu plus haut, les quantités $Y_i(u)$ et $N_i(u)$ ne sont pas représentatives du monde hypothétique dans lequel on se place pour définir la survie nette. On doit en effet diviser ces quantités par la survie attendue au temps u pour l'individu i notée $S_{a,i}(u)$.

L'estimateur de Pohar-Perme devient alors :

$$\hat{H}_c(t) = \int_0^t \frac{\mathrm{d}N^w(u)}{Y^w(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i^w(u)\mathrm{d}H_{a,i}(u)}{Y^w(u)}$$

avec $N_i^w(u) = \frac{N_i(u)}{S_{a,i}(u)}$, $N^w(u) = \sum_{i=1}^n N_i^w(u)$, $Y_i^w(u) = \frac{Y_i(u)}{S_{a,i}(u)}$ et $Y^w(u) = \sum_{i=1}^n Y_i^w(u)$.

Afin de mieux voir ce qu'il se passe, reprenons les données de la section 2.3 : 12 individus dont les temps de décès et de censure sont 1, 2, 2, 2*, 4, 4*, 5*, 6, 6, 6*, 7*, 7*.

Pour estimer la survie nette, nous avons besoin d'informations supplémentaires sur nos individus : leur âge au diagnostic, leur année de diagnostic et leur sexe. Afin de simplifier l'exemple, nous considérons qu'il s'agit de 12 hommes tous diagnostiqués en 2000 à l'âge de 50 ans. Étant donné que tous les individus ont les mêmes caractéristiques, ils auront la même survie attendue à chaque temps, c'est à dire $S_{a,i}(t) = S_a(t)$ et $H_{a,i}(t) = H_a(t)$ pour tout t. L'estimateur de Pohar-Perme devient :

$$\hat{H}_c(t) = \int_0^t \frac{\mathrm{d}N(u)}{Y(u)} - \int_0^t \mathrm{d}H_a(u)$$

Le tableau 2.3 détaille les calculs de l'estimateur de Pohar-Perme.

temps	taux	taux attendu	taux en excès cumulé	survie nette
1	1/12	$\tau_{50,2000}$	$\hat{H}_c(1) = 1/12 - \tau_{50,2000}$	92,5%
2	2/11	$\tau_{51,2001}$	$\hat{H}_c(2) = \hat{H}_c(1) + 2/11 - \tau_{51,2001}$	77,9%
4	1/8	$\tau_{53,2003}$	$\hat{H}_c(4) = \hat{H}_c(2) + 1/8 - \tau_{52,2002} - \tau_{53,2003}$	69,9%
5	0	$\tau_{54,2004}$	$\hat{H}_c(5) = \hat{H}_c(4) - \tau_{54,2004}$	70,5%
6	2/5	$\tau_{55,2005}$	$\hat{H}_c(6) = \hat{H}_c(5) + 2/5 - \tau_{55,2005}$	47,7%
7	0	$\tau_{56,2006}$	$\hat{H}_c(7) = \hat{H}_c(6) - \tau_{56,2006}$	48,3%

Tableau 2.3 – Estimateur de Pohar-Perme

Vous remarquerez qu'ici la survie nette est également calculée aux temps de censure contrairement aux estimateurs de la survie brute.

La figure 2.7 compare la survie brute estimée par Nelson-Aalen à la survie nette estimée par Pohar-Perme :

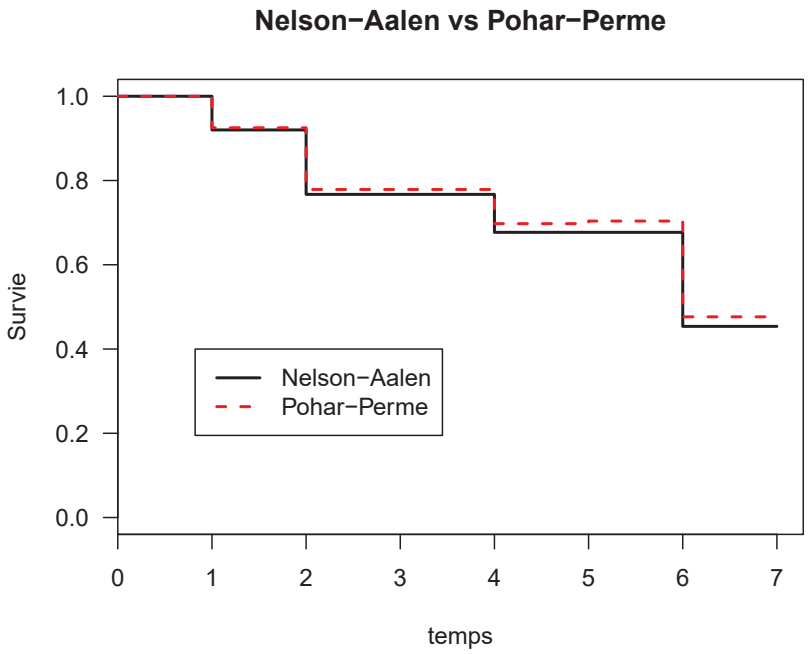

Figure 2.7 – Comparaison entre survie brute et survie nette (Nelson-Aalen vs Pohar-Perme)

2.9.3 Vraisemblance du modèle de taux en excès

Dans un modèle de taux en excès, h_P est considéré connu (obtenu à partir d'instituts statistiques nationaux comme l'INSEE pour la France) et nous souhaitons modéliser h_E de la manière suivante :

$$log\{h_E(t_i; \boldsymbol{x}_i)\}_{1 \leq i \leq n} = \boldsymbol{X(t)\beta}$$

Soit $\boldsymbol{a}$ le vecteur de taille n contenant les âges au diagnostic, la contribution à la log-vraisemblance d'un individu i s'écrit :

$$l_i(\boldsymbol{\beta}) = \delta_i log\left[h_E(t_i; \boldsymbol{x}_i) + h_P(t_i; \boldsymbol{z}_i)\right] - \int_0^{t_i} h_E(u; \boldsymbol{x}_i) + h_P(a_i + u; \boldsymbol{z}_i)du$$

Comme h_P ne dépend pas de $\boldsymbol{\beta}$, maximiser $l(\boldsymbol{\beta}) = \sum_i l_i(\boldsymbol{\beta})$ est équivalent à maximiser l'expression suivante :

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \delta_i log\left[h_E(t_i; \boldsymbol{x}_i) + h_P(t_i; \boldsymbol{z}_i)\right] - \int_0^{t_i} h_E(u; \boldsymbol{x}_i)du \right\}$$

2.9.4 Exemples de modèles de taux en excès

Afin d'éviter tout problème d'intégration du taux, le premier modèle de taux en excès fut un modèle de taux constant par intervalles (Estève et al., 1990). Supposons que l'intervalle de temps $[t_0; t_r]$ soit découpé en r intervalles définis par les $r + 1$ valeurs $t_0 \leq t_1 \leq \cdots \leq t_k$. Le modèle d'Estève s'écrit alors :

$$\text{modèle d'Estève} : h_E(t, \boldsymbol{x}) = exp\left(\sum_{i=1}^{p} \beta_i x_i\right) \sum_{k=1}^{r} \tau_k \mathbb{1}_{[t_{k-1} \leq t \leq t_k]}(t)$$

où p est le nombre de covariables, et τ_k est le taux en excès de base sur le k^e intervalle de temps.

Giorgi et al. (2003) étendirent le modèle d'Estève en proposant de modéliser à la fois le taux de base et la non-proportionnalité des effets des covariables par des B-splines quadratiques. Afin de tenir compte de la non-linéarité des effets des covariables et du temps, de la non-proportionnalité des effets des covariables ainsi que de la structure d'interactions, Remontet et al. (2007) développèrent le modèle suivant (en prenant l'âge comme exemple de covariable) :

$$\text{modèle de Remontet} : log\{h_E(t, a)\} = f(t) + g(a) + a \times h(t)$$

où f et h sont des splines cubiques avec un nœud à un an (le temps maximal était de cinq ans) et g est une spline cubique avec un nœud à l'âge médian. Ce modèle offre une grande flexibilité pour la prise en compte de la non-linéarité, de la non-proportionnalité ainsi que de l'interaction entre les effets de l'âge et du temps.

2.9.5 Approche de Poisson pour le taux en excès

De la même manière qu'en survie brute, l'approche de Poisson peut être utilisée pour maximiser la vraisemblance d'un modèle de taux en excès (Remontet et al., 2007, 2019). Toutefois, la mise en place est beaucoup plus fastidieuse : la fonction de lien *log* doit notamment être modifiée afin de prendre en compte les taux de mortalité attendus.

2.10 Validation des modèles de taux

Lorsque l'on ajuste un modèle sur des données observées, il est essentiel de vérifier si les hypothèses induites par le modèle sont vérifiées et si les prédictions effectuées par le modèle sont en accord avec les observations. Le problème des modèles de taux et des modèles de survie en général, c'est qu'il n'existe pas de valeurs directement observables de la variable aléatoire d'intérêt (à cause de la censure) et il est donc impossible de comparer graphiquement les prédits et les observés comme on pourrait le faire dans le cadre d'un GLM par exemple. Malgré cette contrainte, nous proposons ci-dessous deux façons élémentaires de juger de l'adéquation d'un modèle de taux.

2.10.1 Survie populationnelle et comparaison avec un estimateur non-paramétrique

Supposons que l'on ait ajusté un modèle de taux paramétrique en fonction du temps t et d'un vecteur de caractéristiques $\boldsymbol{x}$,

$$log\left\{h(t,\boldsymbol{x})\right\} = f(t,\boldsymbol{x})$$

et l'on souhaite savoir si le modèle est une représentation acceptable des données.

Pour chaque individu i du jeu de données ayant servi à l'ajustement, pour un temps t quelconque, on dispose de l'estimation de son taux instantané $\hat{h}_i(t)$ grâce à son vecteur $\boldsymbol{x}$. Nous pouvons donc calculer la survie de l'individu au temps t :

$$\hat{S}_i(t) = exp\left(-\int_0^t \hat{h}_i(u)\mathrm{d}u\right)$$

On peut également prédire la survie au temps t d'un groupe d'individus contenant n individus, c'est à dire la survie populationnelle de ce groupe :

$$\hat{S}_{pop}(t) = \frac{1}{n}\sum_{i=1}^{n}\hat{S}_i(t)$$

Dans le cadre d'un modèle de taux en excès, cette définition reste inchangée :

$$\widehat{SN}_{pop}(t) = \frac{1}{n}\sum_{i=1}^{n}\widehat{SN}_i(t)$$

Pour différentes valeurs de t, on peut donc obtenir $\widehat{SN}_{pop}$ et construire une courbe de survie populationnelle prédite par le modèle. L'intérêt est ensuite de comparer graphiquement cette courbe prédite à la courbe restituée par l'estimateur de Kaplan-Meier (survie brute) ou Pohar-Perme (survie nette). Évidemment, si notre modèle contient plusieurs variables explicatives, il faudra vérifier que l'adéquation du modèle est correcte quelles que soient les valeurs de ces covariables. Par exemple, pour un modèle ajusté sur l'âge, il faudra construire plusieurs classes d'âges et effectuer autant de comparaisons qu'il y a de classes d'âges.

2.10.2 Taux populationnel et comparaison à un modèle de taux constant par intervalles

La comparaison avec un estimateur non-paramétrique nous oblige à valider l'ajustement sur l'échelle de la survie (ou du taux cumulé). Or l'idéal serait de pouvoir valider l'ajustement sur l'échelle du taux instantané étant donné que c'est ce dernier que l'on modélise.

De la même manière qu'il est possible de prédire la survie d'un groupe d'individus à partir des survies individuelles (on parle de survie populationnelle, voir précédemment), il est possible de prédire le taux d'un groupe d'individus à partir des taux individuels, on parle alors de taux populationnel (voir la section 2.5) :

$$\hat{h}_{pop}(t) = \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{h}_i(t)\hat{S}_i(t)}{\frac{1}{n}\sum_{i=1}^{n} \hat{S}_i(t)}$$

Dans le cadre d'un modèle de taux en excès, on a :

$$\hat{h}_{E,pop}(t) = \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{h}_{E,i}(t)\widehat{SN}_i(t)}{\frac{1}{n}\sum_{i=1}^{n} \widehat{SN}_i(t)}$$

Le taux populationnel peut être comparé à des estimations provenant d'un modèle de taux constant par intervalles (Estève et al., 1990). Ces taux constants représentent alors une sorte de nuage de points à ajuster. La figure 2.8 illustre la comparaison entre un taux populationnel et un taux constant par intervalles.

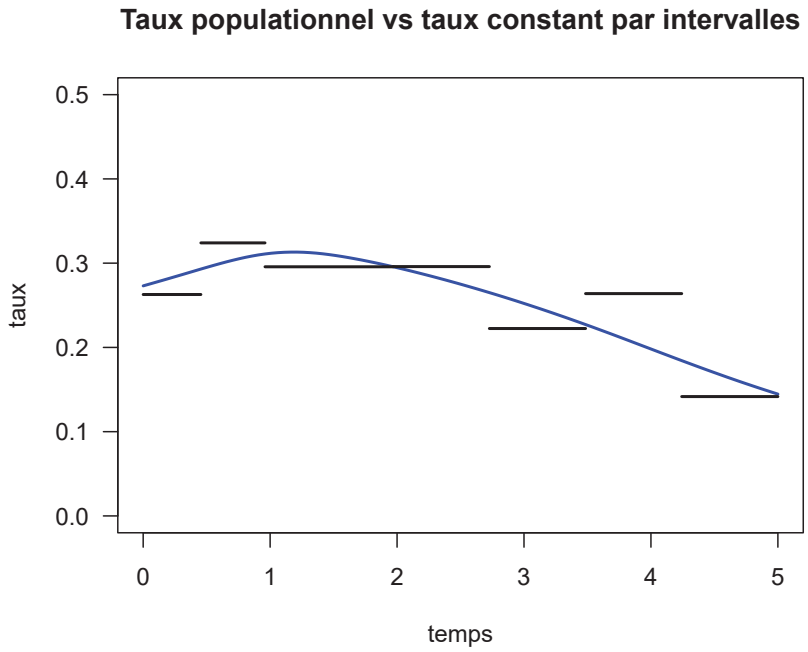

Figure 2.8 – Comparaison entre taux populationnel et taux constant par intervalles

De même que pour la survie populationnelle, il est nécessaire de vérifier l'adéquation pour différentes valeurs de covariables.

2.10.3 Autres outils diagnostiques

Notons que différents tests statistiques ont été proposés pour les modèles de régression en survie. Nous pouvons notamment citer les résidus de Schoenfeld (Schoenfeld, 1982) qui permettent de tester l'hypothèse des taux proportionnels. De tels résidus ont été étendus aux modèles de taux en excès par Stare et al. (2005).
En s'appuyant sur le cadre théorique des transformées de martingale et notamment les travaux de Lin and Spiekerman (1996), Danieli et al. (2017) ont proposé deux tests formels afin de tester l'hypothèse de proportionnalité et la forme fonctionnelle des covariables dans les modèles de taux en excès.

Chapitre 3

Sélection de modèle

Lorsque l'on construit un modèle afin de quantifier l'influence de covariables, on est confronté à deux problèmes majeurs :

1. **Sélection de variable** : Comment déterminer les variables pertinentes parmi toutes celles disponibles ?
2. **Choix de modèle** : Comment spécifier l'effet de ces variables ? Quelles formes fonctionnelles utiliser ? Quelles interactions envisager ?

Afin de faire face efficacement à ces deux problèmes, il est nécessaire de les considérer de manière simultanée. En effet, lorsque l'on teste si une certaine covariable doit être incluse dans un modèle, le résultat du test dépend de la forme fonctionnelle considérée.
Ce chapitre vise à présenter brièvement quelques principes et exemples de procédures de sélection de modèles utiles notamment en analyse de survie. Ce sujet est traité de manière détaillée par Anderson and Burnham (2004), Burnham and Anderson (2004) ou encore Kneib et al. (2009).

3.1 Particularité de la sélection de modèle en survie

La sélection de modèle constitue une part très importante de la modélisation statistique. Dans les modèles de survie, le choix d'un modèle adéquat repose sur trois aspects majeurs :

— la non-proportionnalité (ou dépendance au temps) de l'effet d'une covariable
— la non-linéarité des effets des variables continues
— les interactions entre les effets des covariables

Notons également que ces trois aspects doivent être gérés simultanément. En particulier, concernant la nécessité de modéliser conjointement la non-linéarité et la non-proportionnalité, le lecteur pourra se tourner vers Abrahamowicz and MacKenzie (2007).

3.2 Exemples de procédures de sélection de modèle

Pour sélectionner des variables ou choisir un modèle, il faut tout d'abord définir un critère ou une mesure reflétant l'adéquation d'un modèle à des données. La théorie de l'information (Shannon, 1948) apparaît alors comme une alliée indispensable.

3.2.1 Divergence de Kullback-Liebler

Soient p et q deux densités absolument continues par rapport à la mesure de Lebesgue, alors on définit la quantité suivante, appelée divergence de Kullback-Liebler (K-L, Kullback and Leibler 1951) :

$$KL(p||q) = \begin{cases} \int_{-\infty}^{+\infty} p(x) log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x & \text{si p est absolument continue par rapport à q} \\ \infty & \text{sinon} \end{cases}$$

La divergence K-L est intimement liée à l'estimateur du maximum de vraisemblance (MLE). En effet, soit $Y_1, ..., Y_n$ un échantillon de variables aléatoires de densité q_i. On essaie d'expliquer les Y_i à partir d'un modèle paramétrique avec la densité $p(Y_i, \boldsymbol{\beta})$ où $\boldsymbol{\beta}$ est le vecteur de paramètres à estimer. On a donc :

$$\hat{\boldsymbol{\beta}}_{MLE} = \underset{\boldsymbol{\beta}}{argmax}[l(\boldsymbol{\beta})] = \underset{\boldsymbol{\beta}}{argmax} \sum_i log\left[p(Y_i, \boldsymbol{\beta})\right] = \underset{\boldsymbol{\beta}}{argmin} \sum_i log\left[\frac{1}{p(Y_i, \boldsymbol{\beta})}\right]$$

Dans l'expression de droite nous pouvons introduire les densités q_i qui sont indépendantes de $\boldsymbol{\beta}$ et il vient :

$$\hat{\boldsymbol{\beta}}_{MLE} = \underset{\boldsymbol{\beta}}{argmin} \sum_i log\left[\frac{q_i}{p(Y_i, \boldsymbol{\beta})}\right]$$

L'espérance du terme de droite correspond à $KL(\mathbf{q}||\mathbf{p}(\boldsymbol{\beta}))$ avec $\mathbf{q} = \prod_i q_i$ et $\mathbf{p}(\boldsymbol{\beta}) = \prod_i p(Y_i, \boldsymbol{\beta})$.

Ainsi, le MLE cherche le vecteur de paramètres $\boldsymbol{\beta}$ pour lequel la densité estimée sera la plus proche de la véritable densité au sens de la divergence K-L.

L'intérêt de la divergence K-L est donc d'apporter une définition rigoureuse de l'information contenue dans un modèle mais également de la distance entre deux modèles (voir des exemples d'application dans la section 2.1.1 d'Anderson and Burnham 2004).

3.2.2 Critère d'information d'Akaike (AIC)

Depuis sa proposition en 1973, l'AIC (pour *Akaike Information Criterion*, Akaike 1973) s'est imposé comme une mesure incontournable de la qualité statistique d'un modèle. Construit à partir de la divergence K-L, il permet de comparer des modèles très différents y compris non-emboîtés (c'est à dire qu'il ne sont pas des cas particuliers les uns des autres), ce qui n'est pas possible avec le test du rapport de vraisemblance notamment.
On sait que la vraisemblance d'un modèle constitue une mesure objective de sa proximité aux données. Malheureusement, plus un modèle est complexe, plus sa vraisemblance est importante. Ainsi, si l'on choisit le modèle ayant la plus grande vraisemblance, on retiendra toujours le modèle le plus complexe. L'idée de l'AIC est de trouver un compromis entre la proximité aux données (donc la vraisemblance) et la complexité (le nombre de paramètres). Si L est la vraisemblance d'un modèle et k son nombre de paramètres alors :

$$AIC = -2log(L) + 2k \tag{3.1}$$

L'AIC cherche ainsi à respecter le principe de parcimonie. Voir l'Appendix A.7 de Wood (2017) pour une présentation détaillée.

3.2.3 Choix parmi un ensemble de modèles candidats

Dans le cadre de la sélection de variables et du choix de modèle, une stratégie assez simple consiste à construire un ensemble de modèles candidats du plus simple au plus compliqué (en termes de nombre de covariables et de complexité d'effets) dans les limites d'un ensemble de covariables de départ et des effets que l'on cherche à capter. Le modèle retenu sera alors celui de plus faible AIC.

Cette stratégie de sélection de modèles par AIC dans le cadre des modèles de survie est notamment présente dans Uhry et al. (2017).

Exemple : supposons que l'on cherche à construire un modèle de taux incluant les effets du temps et de l'âge. On sait que l'effet du temps peut être modélisé de manière acceptable par une spline cubique. En revanche, on s'interroge sur l'effet de l'âge ainsi qu'à la présence d'une interaction avec le temps. Des modèles candidats possibles sont :

$$
\begin{aligned}
mod1 &: log[h(t, age)] = f_1(t) \\
mod2 &: log[h(t, age)] = f_1(t) + age \\
mod3 &: log[h(t, age)] = f_1(t) + age + age \times t \\
mod4 &: log[h(t, age)] = f_1(t) + age + age \times f_2(t) \\
mod5 &: log[h(t, age)] = f_1(t) + g_1(age) \\
mod6 &: log[h(t, age)] = f_1(t) + g_1(age) + age \times t \\
mod7 &: log[h(t, age)] = f_1(t) + g_1(age) + age \times f_2(t) \\
mod8 &: log[h(t, age)] = f_1(t) + g_1(age) + g_2(age) \times f_2(t)
\end{aligned}
$$

où les f_i et g_i sont des splines cubiques dont les nœuds sont connus.

À partir d'un exemple très simple (deux covariables), on arrive ici à huit modèles candidats (sans même évoquer le degré de la spline, le nombre de nœuds ou leur position). Cette méthode présente donc des inconvénients pratiques majeurs du fait du possible grand nombre de modèles candidats.

3.2.4 Procédures *backward* et *forward*

Une alternative à la construction d'un ensemble de modèles candidats consiste à construire un processus itératif selon trois options :

- L'option *forward* : on construit un modèle initial dans lequel se trouvent les effets minimum attendus. On va alors complexifier le modèle de proche en proche en lui ajoutant des briques élémentaires (variable ou forme fonctionnelle de variable) préalablement constituées. À chaque étape, on utilise un critère (AIC, rapport de vraisemblance, etc) pour savoir si la brique élémentaire qui vient d'être rajoutée doit rester dans le modèle ou être retirée.
- L'option *backward* : il s'agit du même principe que l'option précédente excepté que l'on démarre désormais du modèle le plus complexe possible et qu'on le simplifie à chaque étape.
- L'option *stepwise* : il s'agit d'une alternance entre étapes *forward* et étapes *backward*.

Dans le contexte des modèles de survie, ces approches sont notamment illustrées par Wynant and Abrahamowicz (2014) qui proposent une stratégie de type *backward* afin de sélectionner les effets non-linéaires et non-proportionnels des covariables.

3.2.5 Limites

Ces différentes procédures de sélection sont malheureusement heuristiques et contiennent plusieurs failles :

- on ne peut pas être sûr qu'un modèle adéquat se trouve parmi les modèles construits
- on répète les tests statistiques au cours d'un processus *backward* ou *forward* sans manière simple d'ajuster le risque alpha (en général, aucun ajustement n'est prévu)
- le nombre de modèles construits peut rapidement devenir très important
- une légère perturbation des données entraînerait le choix d'un nouveau « meilleur » modèle
- si l'on ne tient pas compte de l'incertitude sur le modèle choisi, les intervalles de confiance estimés à partir de ce même modèle auront des probabilités de couverture plus faibles qu'attendues (Burnham and Anderson, 2004)
- si aucun des modèles candidats ne sort vraiment du lot au sens du critère de comparaison utilisé et que les prédictions issues des différents modèles candidats peuvent s'avérer assez différentes, il paraît risqué de ne retenir qu'un seul modèle pour faire de l'inférence
- le modèle issu d'une sélection a toutes les chances de sur-ajuster les données et d'offrir une sous-estimation des variances (Anderson and Burnham, 2004)

Des solutions existent pour éviter les différents écueils cités ici (Ye, 1998). Par exemple, au lieu de choisir le « meilleur » modèle parmi un ensemble de modèles candidats, on peut se servir de tous les modèles vraisemblables que l'on a construit pour en extraire une sorte de modèle moyen (on parle de *model averaging*). Toutefois, l'inférence issue d'une telle approche peut s'avérer ardue (Banner and Higgs, 2017).

Dans la suite de cette thèse nous allons présenter une autre approche très répandue qui permet de simplifier les processus de sélection et de spécification de modèle tout en luttant contre le sur-ajustement : la pénalisation.

Chapitre 4

Pénalisation

L'une des meilleures manières de lutter contre l'incertitude de modèle est de se contenter d'ajuster un unique modèle. Quitte à en choisir un seul, autant qu'il soit flexible pour capturer le maximum d'information. Il ne doit pas non plus être trop flexible sans quoi le sur-ajustement est inévitable. La pénalisation permet de trouver un compromis entre ces deux aspects.

4.1 Interpolation et lissage

Considérons les points $\{x_i, y_i : i = 1, \ldots, n\}$ avec $x_i < x_{i+1}$.

Supposons que l'on veuille interpoler ces points en utilisant une fonction g, c'est à dire $g(x_i) = y_i$ pour tout i. Simplement, nous ne voulons pas n'importe quelles propriétés pour cette fonction g. Ainsi, nous souhaiterions que la fonction soit la plus lisse possible. Mathématiquement, nous souhaiterions par exemple qu'elle soit continue et que ses dérivées première et seconde soient également continues sur l'ensemble de définition $[x_1; x_n]$. En outre, nous imposons qu'elle soit linéaire en ces bornes extérieures x_1 et x_n, c'est à dire $g''(x_1) = g''(x_n) = 0$.

Green and Silverman (1993) ont montré que de toutes les fonctions g respectant ces contraintes, la plus lisse dans le sens de la minimisation de :

$$J(g) = \int_{x_1}^{x_n} g''(x)^2 \mathrm{d}x$$

était une spline cubique naturelle. La fonction g est donc une collection de polynômes cubiques, un pour chaque intervalle $[x_i; x_{i+1}]$, et assemblés de telle manière aux points x_i que g est continue jusqu'à sa dérivée seconde, que $g(x_i) = y_i$ et que $g''(x_1) = g''(x_n) = 0$.

Ainsi, on voit que si l'on souhaite représenter des données à l'aide d'une fonction lisse et souple, les splines cubiques naturelles constituent une solution intéressante.

Toutefois, en statistiques, les y_i sont mesurés avec un certain bruit aléatoire et il n'est donc pas nécessairement indiqué d'interpoler les données. Plutôt que d'imposer $g(x_i) = y_i$, il est préférable de considérer les $g(x_i)$ comme n paramètres à estimer en minimisant :

$$\sum_{i=1}^{n} [y_i - g(x_i)]^2 + \lambda \int_{x_1}^{x_n} g''(x)^2 \mathrm{d}x$$

λ est le paramètre de lissage qui contrôle le compromis entre la fidélité aux données et la régularité (ou lissage) de la fonction g. La solution g à ce problème de minimisation est une spline de lissage, ou *smoothing spline* (Reinsch, 1967). En pratique, la fonction g serait écrite de manière paramétrique dans une certaine base de splines (typiquement les B-splines vues en section 1.2.2, comme avec la fonction *smooth.spline* dans R) avec autant de paramètres à estimer que de x_i.

Nota Bene : le terme de régression **non-paramétrique** est très souvent utilisé pour décrire les splines de lissage car très peu d'hypothèses sont faites sur la véritable forme de l'effet à estimer et que les données sont utilisées pour quantifier le degré de lissage nécessaire. Toutefois, ce terme apporte beaucoup de confusion car, en pratique, dès que l'on travaille avec des splines, on finit toujours par écrire l'effet en question comme une forme **paramétrique** associée à un certain nombre de paramètres à estimer (aussi nombreux soient-ils).

Les splines cubiques de lissage semblent donc être une approche intéressante pour capturer des effets non-linéaires tout en proposant une fonction la plus lisse possible. Elles présentent pourtant un problème de taille : leur complexité algorithmique proportionnelle au cube du nombre d'individus à analyser. Des approximations ont été proposées afin de contourner ce problème (Du and Gu, 2006) tout en conservant la nature non-paramétrique des splines de lissage.

4.2 Splines de régression pénalisées

Les splines de régression pénalisées sont nées de la nécessité d'établir un compromis entre les bonnes propriétés des splines de lissage et le faible coût algorithmique des splines de régression (Wood, 2017). En réponse à Silverman (1985), Parker and Rice (1985) proposèrent d'approcher les splines de lissage par des B-splines de dimension inférieure en spécifiant les nœuds a priori et en utilisant une pénalité sur le carré de la dérivée seconde. À l'époque, ils nommèrent cette approche *least square splines*.

Les splines de régression pénalisées ont ensuite été popularisées par Wood (2000, 2006b, 2011); Wood et al. (2016b).

Pour rappel, une spline de régression f à k bases s'écrit :

$$f(x) = \sum_{j=1}^{k} \beta_j b_j(x)$$

où les b_j sont les bases de la spline (connues mais à spécifier) et les β_j sont les paramètres à estimer.

Dans un modèle de régression classique, l'objectif est d'estimer les paramètres β_j en maximisant la log-vraisemblance qui à $\boldsymbol{\beta}$ associe $l(\boldsymbol{\beta})$. L'idée de base de la pénalisation est de restreindre le domaine de définition de ces paramètres afin que les estimations respectent un certain nombre de contraintes. On parle donc d'optimisation sous contrainte. Le degré de pénalisation est géré à l'aide d'un ou plusieurs paramètres de lissage qui interviennent directement dans l'expression de la nouvelle fonction à optimiser. Pour un unique paramètre de lissage λ, on cherche à maximiser la quantité suivante :

$$\mathcal{L}(\boldsymbol{\beta}, \lambda) = l(\boldsymbol{\beta}) - \lambda \left[Pen(\boldsymbol{\beta}) \right]$$

où $\mathcal{L}$ est la log-vraisemblance pénalisée et $Pen(\boldsymbol{\beta})$ représente le terme de pénalité qui sert à contraindre l'espace de définition des paramètres de régression $\boldsymbol{\beta}$. La maximisation de $\mathcal{L}$ permet d'obtenir les estimations des paramètres de régression $\hat{\boldsymbol{\beta}}^{\lambda}$ à λ fixé.

La première utilisation d'un score faisant intervenir la vraisemblance et un terme de pénalité remonte à Good and Gaskins (1971). Depuis lors, le rationnel de la définition est resté inchangé : $\mathcal{L}$ représente un compromis entre la fidélité aux données (représentée par la log-vraisemblance) et la pénalité. Le paramètre de lissage λ contrôle le poids que représente la pénalité par rapport à la log-vraisemblance.

Des exemples de pénalisation très souvent utilisées sont le LASSO, qui correspond à $Pen(\beta) = \sum_j |\beta_j|$, et la pénalisation Ridge, qui correspond à $Pen(\beta) = \sum_j \beta_j^2$. Ces deux types de pénalisation permettent de diminuer l'erreur quadratique moyenne de l'estimateur des moindres carrés : l'idée est de biaiser les prédictions pour en diminuer la variance.
La particularité du LASSO est que la nature de sa contrainte permet d'éliminer du modèle des variables peu ou pas explicatives en annulant leur paramètre de régression correspondant. Le LASSO est ainsi particulièrement utile lorsque qu'il y a plus de paramètres que d'individus statistiques $p > n$.
Quant au Ridge, contrairement au LASSO, il fait partie d'une classe de pénalités qui va particulièrement nous intéresser dans cette thèse : les pénalités quadratiques (voir la section 4.2.2).

Dans le cas général, plusieurs paramètres de pénalisation peuvent être associés à un même modèle et la log-vraisemblance pénalisée s'écrit :

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = l(\boldsymbol{\beta}) - \sum_m \lambda_m Pen_m(\boldsymbol{\beta})$$

Les définitions données ici ne fonctionnent que si les paramètres de lissage sont connus. En général ce n'est évidemment pas le cas et il est nécessaire de les estimer. Pour l'instant, considérons qu'ils sont connus et intéressons-nous à la construction de la pénalité dans le cadre des splines cubiques.

4.2.1 Pénalité sur la dérivée seconde

On a vu que les splines cubiques et notamment les splines cubiques restreintes présentent de nombreux avantages pour la modélisation flexible des formes fonctionnelles de variables continues.

De la même manière que pour les splines de lissage, il est naturel de construire une pénalité sur la dérivée seconde d'une spline de régression. En effet, la dérivée seconde d'une fonction mesure la variation de sa dérivée première, c'est à dire sa courbure qui s'apparente à sa « régularité ».

À la spline de régression f telle que $f(x) = \sum_{j=1}^{k} \beta_j b_j(x)$, on associe donc la pénalité : $Pen(\boldsymbol{\beta}) = \int f''(x)^2 \mathrm{d}x$.

La figure 4.1 illustre le lien entre régularité d'une fonction et la valeur de l'intégrale définie ci-dessus. Les quatre fonctions prises en exemple sont définies sur $[0; 1]$ et valent 0 en 0 et 1 en 1. Elles partent donc toutes du même point et arrivent également toutes au même point. En revanche, on voit que leur parcours est plus ou moins sinueux en fonction de la valeur de l'intégrale.

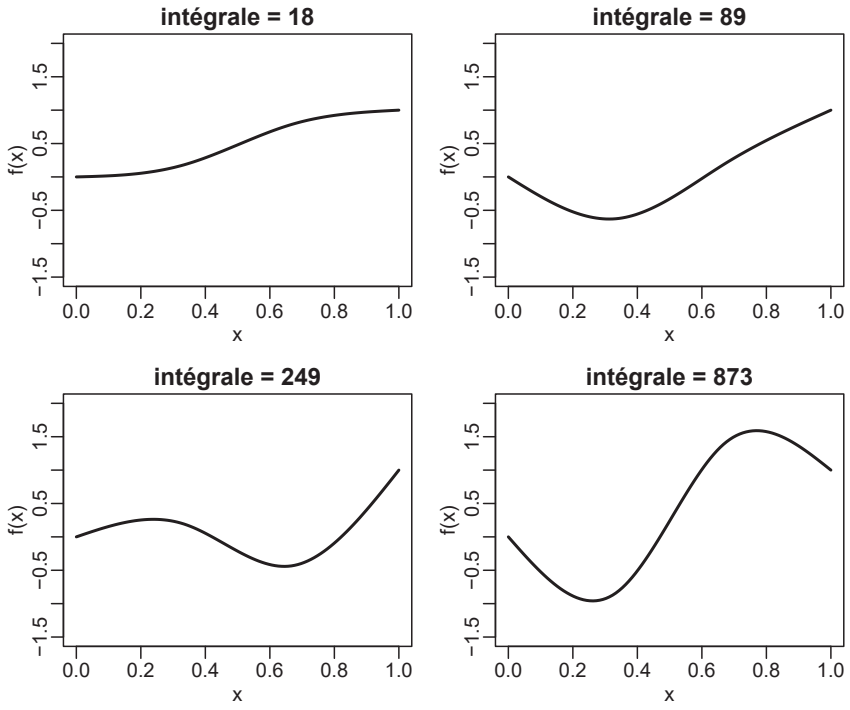

Figure 4.1 – Lien entre régularité et valeur de l'intégrale de la dérivée seconde au carré pour quatre fonctions définies sur $[0; 1]$

4.2.2 Matrice de pénalisation

Toute spline de régression s'écrit

$$f(x) = \sum_{j=1}^{k} \beta_j b_j(x)$$

Sa dérivée seconde est donc

$$f''(x) = \sum_{j=1}^{k} \beta_j b_j''(x)$$

Nous savons que les β_j sont les composantes du vecteur $\boldsymbol{\beta}$. De la même manière, définissons le vecteur $\boldsymbol{b''}(x)$ tel que ses composantes soient les $b_j''(x)$:

$$\boldsymbol{b''}(x) = \begin{bmatrix} b_1''(x) \\ \dots \\ b_k''(x) \end{bmatrix}$$

Il vient alors

$$f''(x) = \boldsymbol{\beta}^T \boldsymbol{b''}(x)$$

Et

$$\begin{aligned} Pen(\boldsymbol{\beta}) &= \int f''(x)^2 \mathrm{d}x \\ &= \int [\boldsymbol{\beta}^T \boldsymbol{b''}(x)]^2 \mathrm{d}x \\ &= \int [\boldsymbol{\beta}^T \boldsymbol{b''}(x)][\boldsymbol{\beta}^T \boldsymbol{b''}(x)]^T \mathrm{d}x \\ &= \int \boldsymbol{\beta}^T \boldsymbol{b''}(x)\boldsymbol{b''}(x)^T \boldsymbol{\beta} \mathrm{d}x \\ &= \boldsymbol{\beta}^T \left[\int \boldsymbol{b''}(x)\boldsymbol{b''}(x)^T \mathrm{d}x \right] \boldsymbol{\beta} \end{aligned}$$

On note $\boldsymbol{S} = \left[\int \boldsymbol{b''}(x)\boldsymbol{b''}(x)^T \mathrm{d}x \right]$ la matrice de pénalisation associée à la spline f. Elle est entièrement définie par la base utilisée pour écrire la spline.

Le terme de pénalité associé est ainsi une forme quadratique des paramètres de régression :

$$Pen(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}$$

Cette écriture est extrêmement pratique pour la maximisation de la log-vraisemblance pénalisée comme nous le verrons par la suite.

Exemple de matrice de pénalisation :

Soit f un polynôme cubique défini sur $[0, 1]$ par :

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3$$

f est donc de la forme $f(x) = \sum \beta_j b_j(x)$ avec $\boldsymbol{b}(x) = \begin{bmatrix} 1 & x & x^2 & x^3 \end{bmatrix}^T$ et $\boldsymbol{b''}(x) = \begin{bmatrix} 0 & 0 & 2 & 6x \end{bmatrix}^T$.
La base polynomiale choisie est ici volontairement simplifiée par rapport aux bases de splines généralement utilisées.
On a donc :

$$\boldsymbol{b''}(x)\boldsymbol{b''}(x)^T = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 6x \end{bmatrix} \begin{bmatrix} 0 & 0 & 2 & 6x \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 12x \\ 0 & 0 & 12x & 36x^2 \end{bmatrix}$$

et

$$S = \int_0^1 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 12x \\ 0 & 0 & 12x & 36x^2 \end{bmatrix} \mathrm{d}x = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 6 \\ 0 & 0 & 6 & 12 \end{bmatrix}$$

Finalement, la pénalité s'écrit $\boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta} = 4\beta_3^2 + 12\beta_3\beta_4 + 12\beta_4^2$.

Dans le cas d'une pénalisation Ridge, c'est à dire $Pen(\boldsymbol{\beta}) = \sum_j \beta_j^2$, on aurait $S = I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

Matrice de pénalisation associée aux *cubic regression splines*

Un des atouts des *cubic regression splines* présentées plus haut (section 1.2.3) est la relative facilité avec laquelle on peut définir une pénalisation sur la dérivée seconde.

En effet, on peut montrer que (en reprenant les notations de la section 1.2.3) :

$$\int_{x_1}^{x_k} f''(x)^2 \mathrm{d}x = \boldsymbol{\beta}^T \boldsymbol{D}^T \boldsymbol{B}^{-1} \boldsymbol{D} \boldsymbol{\beta}$$

Ainsi, $\boldsymbol{S} = \boldsymbol{D}^T \boldsymbol{B}^{-1} \boldsymbol{D}$ est la matrice de pénalisation associée à une *cubic regression spline*.

Démonstration : sur l'intervalle $[x_j; x_{j+1}]$, $f''(x)$ s'écrit :

$$f''(x) = \delta_j \frac{x_{j+1} - x}{h_j} + \delta_{j+1} \frac{x - x_j}{h_j}$$

Sur $[x_1; x_k]$, on peut écrire :

$$f''(x) = \sum_{i=2}^{k-1} \delta_i d_i(x)$$

avec

$$d_i(x) = \begin{cases} (x - x_i)/h_{i-1} & \text{si } x_{i-1} \leq x \leq x_i \\ (x_{i+1} - x)/h_i & \text{si } x_i \leq x \leq x_{i+1} \\ 0 & \text{sinon.} \end{cases}$$

Soit $\boldsymbol{d}(x)$ le vecteur contenant les éléments $d_i(x)$ pour $i = 2, \ldots, k-1$. On peut alors écrire f'' de manière matricielle :

$$f''(x) = \boldsymbol{\delta^{-T}} \boldsymbol{d}(x)$$

D'où

$$\int_{x_1}^{x_k} f''(x)^2 \mathrm{d}x = \int_{x_1}^{x_k} \left(\boldsymbol{\delta^{-T}} \boldsymbol{d}(x) \right)^2 \mathrm{d}x$$

$$\int_{x_1}^{x_k} f''(x)^2 \mathrm{d}x = \int_{x_1}^{x_k} \boldsymbol{\delta^{-T}} \boldsymbol{d}(x) \boldsymbol{d}(x)^T \boldsymbol{\delta^{-}} \mathrm{d}x$$

$$\int_{x_1}^{x_k} f''(x)^2 \mathrm{d}x = \boldsymbol{\delta^{-T}} \left(\int_{x_1}^{x_k} \boldsymbol{d}(x) \boldsymbol{d}(x)^T \mathrm{d}x \right) \boldsymbol{\delta^{-}}$$

Chaque $d_i(x)$ est non nul sur seulement deux intervalles, la matrice $\int_{x_1}^{x_k} \boldsymbol{d}(x) \boldsymbol{d}(x)^T \mathrm{d}x$ est donc symétrique et tri-diagonale par construction. Les éléments diagonaux (pour $i = 2, \ldots, k-1$) sont :

$$\int_{x_{i-1}}^{x_{i+1}} d_i(x)^2 \mathrm{d}x = \left[\frac{(x - x_{i-1})^3}{3h_{i-1}^2} \right]_{x_{i-1}}^{x_i} - \left[\frac{(x_{i+1} - x)^3}{3h_i^2} \right]_{x_i}^{x_{i-1}} = \frac{h_{i-1}}{3} + \frac{h_i}{3}$$

Les éléments extra-diagonaux $(i-1, i)$ et $(i, i-1)$ sont :

$$\int_{x_{i-1}}^{x_i} d_i(x)d_{i-1}(x)\mathrm{d}x = \int_{x_{i-1}}^{x_i} \frac{x - x_{i-1}}{h_{i-1}}\frac{x_i - x}{h_{i-1}}\mathrm{d}x = \frac{h_{i-1}}{6}$$

Et finalement,

$$\int_{x_1}^{x_k} \boldsymbol{d}(x)\boldsymbol{d}(x)^T\mathrm{d}x = \boldsymbol{B}$$

Étant donné que $\boldsymbol{\delta}^- = \boldsymbol{B}^{-1}\boldsymbol{D}\boldsymbol{\beta}$ (1.1), on a bien $\boldsymbol{S} = \boldsymbol{D}^T\boldsymbol{B}^{-1}\boldsymbol{D}$.

4.3 Pénaliser plusieurs splines unidimensionnelles

4.3.1 Construction d'un modèle additif

Considérons la modélisation des effets de deux variables explicatives x et y. On associe à chaque variable un effet marginal représenté par une spline :

$$f_x(x) = \sum_{i=1}^{I} \alpha_i a_i(x) \qquad f_y(y) = \sum_{j=1}^{J} \beta_j b_j(y)$$

Pour chacune de ces splines nous savons construire sa matrice de pénalisation :

$$\boldsymbol{S}_x = \int \boldsymbol{a''}(x)\boldsymbol{a''}(x)^T\mathrm{d}x \qquad \boldsymbol{S}_y = \int \boldsymbol{b''}(y)\boldsymbol{b''}(y)^T\mathrm{d}y$$

On s'intéresse à la modélisation additive suivante :

$$f_{xy}(x, y) = f_x(x) + f_y(y)$$

Tel quel, le modèle n'est pas identifiable. En effet, chaque spline contient une constante (un *intercept*) de telle sorte que si l'on soustrait un certain nombre à f_x tout en l'ajoutant à f_y, les prédictions du modèle seront exactement les mêmes. Il est ainsi impératif d'appliquer une contrainte sur les coefficients à estimer, on parle de contrainte de centrage (voir la section 4.3.2). Après application de la contrainte de centrage sur les deux bases unidimensionnelles, notre effet bidimensionnel s'écrit :

$$f_{xy}(x, y) = \beta_0 + \tilde{f}_x(x) + \tilde{f}_y(y)$$

où $\tilde{f}$ indique la version centrée de f.

La fonction f_{xy} s'accompagne donc de $1 + (I - 1) + (J - 1) = I + J - 1$ paramètres de régression à estimer.

La matrice de design associée est :

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \tilde{\boldsymbol{X}}_x & \tilde{\boldsymbol{X}}_y \end{bmatrix}$$

avec $\tilde{\boldsymbol{X}}_x$ la matrice de design engendrée par $\tilde{f}_x$ et $\tilde{\boldsymbol{X}}_y$ celle engendrée par $\tilde{f}_y$. $\mathbf{1}$ indique une colonne remplie de 1 et correspond à l'intercept.

Le modèle est associé aux deux matrices de pénalisation suivantes :

$$\tilde{\boldsymbol{S}}_{x|xy} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{S}}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \qquad \text{et} \qquad \tilde{\boldsymbol{S}}_{y|xy} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \tilde{\boldsymbol{S}}_y \end{bmatrix}$$

avec $\tilde{\boldsymbol{S}}_x$ la matrice de pénalisation engendrée par $\tilde{f}_x$ et $\tilde{\boldsymbol{S}}_y$ celle engendrée par $\tilde{f}_y$. $\tilde{\boldsymbol{S}}_x$ et $\tilde{\boldsymbol{S}}_y$ sont respectivement de dimensions $(I-1)\times(I-1)$ et $(J-1)\times(J-1)$.

Le principe exposé ici se généralise bien entendu à un nombre arbitraire de splines pénalisées.

4.3.2 Contrainte de centrage

Lorsque l'on cherche à estimer plusieurs splines pénalisées de manière additive, il convient d'appliquer une contrainte à chaque spline afin que le modèle soit identifiable. Supposons que notre modèle additif contienne J splines pénalisées f_j chacune associée à une variable x_j. On note x_{ji} la valeur de la variable x_j pour l'individu statistique i $(i = 1, \ldots, n)$.

Pour l'individu i, le prédicteur linéaire s'écrit

$$\sum_{j=1}^{J} f_j(x_{ji})$$

Pour tout j, Wood (2017) propose d'imposer la contrainte suivante

$$\sum_{i=1}^{n} f_j(x_{ji}) = 0 \tag{4.1}$$

ce qui équivaut à

$$\mathbf{1}^T \boldsymbol{X}^j \boldsymbol{\beta}^j = 0$$

avec $\boldsymbol{X}^j$ et $\boldsymbol{\beta}^j$ la matrice de design et le vecteur de paramètres de régression respectivement associés à la spline f_j.

Nota Bene : D'autres types de contraintes sont également possibles : on pourrait par exemple forcer la spline f_j à s'annuler pour une certaine valeur de covariable. Une alternative encore plus simple consisterait à tout simplement retirer l'intercept des matrices de design $\boldsymbol{X}^j$ (enlever la première colonne). La raison pour laquelle on utilise (4.1) est que les colonnes soumises à la contrainte deviennent orthogonales à l'intercept, ce qui assure une stabilité algorithmique maximum et une estimation plus précise (les intervalles de confiance sont légèrement plus étroits qu'avec d'autres contraintes, voir l'aide de *mgcv* à la page *identifiability*).

Point technique
Pour estimer les paramètres d'un modèle sous contrainte de type $\boldsymbol{C}\boldsymbol{\beta} = \mathbf{0}$, avec $\boldsymbol{C}$ la matrice de contrainte de dimension $m \times p$ et $\boldsymbol{\beta}$ de taille p, on peut réécrire le modèle en fonction de $p - m$ paramètres non contraints (Wood, 2017, section 5.4.1). Il faut trouver la matrice carrée orthogonale $\boldsymbol{Z}$ de dimension $p \times (p - m)$ et telle que $\boldsymbol{C}\boldsymbol{Z} = \mathbf{0}$. Ainsi, on obtient une nouvelle matrice de design $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{Z}$ et un nouveau vecteur de paramètres à estimer $\tilde{\boldsymbol{\beta}}$ de taille $p - m$ tel que $\boldsymbol{\beta} = \boldsymbol{Z}\tilde{\boldsymbol{\beta}}$ et qui satisfait automatiquement la contrainte. La nouvelle matrice de pénalisation est $\tilde{\boldsymbol{S}} = \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}$.

Pour trouver $\boldsymbol{Z}$, il faut tout d'abord effectuer une décomposition QR de $\boldsymbol{C}^T$:

$$\boldsymbol{C}^T = \boldsymbol{U} \begin{bmatrix} \mathbf{P} \\ \mathbf{0} \end{bmatrix}$$

Avec U une matrice orthogonale $p \times p$ et $\boldsymbol{P}$ une matrice triangulaire supérieure $m \times m$. On peut décomposer $\boldsymbol{U} = (\boldsymbol{D} : \boldsymbol{Z})$ avec $\boldsymbol{Z}$ de dimension $p \times (p - m)$. Vérifions qu'on a bien $\boldsymbol{CZ} = \mathbf{0}$:

$$\boldsymbol{CZ} = \begin{bmatrix} \mathbf{P}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{D}^T \\ \mathbf{Z}^T \end{bmatrix} \boldsymbol{Z} = \begin{bmatrix} \mathbf{P}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{p-m} \end{bmatrix} = \mathbf{0}$$

En posant $\boldsymbol{C}_j = \mathbf{1}^T \boldsymbol{X}^j$ pour la j^e spline, on applique la procédure ci-dessus pour les J splines du modèle. On pourra ensuite rajouter un intercept à la matrice de design globale (il suffira d'ajouter une colonne de 1).

4.4 Pénaliser les interactions

Intégrer plusieurs variables dans un modèle purement additif est assez restrictif. Dans l'idéal nous aimerions pouvoir étendre la définition de la pénalisation aux interactions entre plusieurs variables.

4.4.1 Interactions entre une spline pénalisée et une variable continue

Considérons l'effet bidimensionnel suivant :

$$f(x, y) = \beta_0 + \beta_1 \times y + f_x(x) + g_x(x) \times y$$

où f_x et g_x sont des splines pénalisées. f_x et g_x sont ici soumises à une contrainte de centrage car l'intercept et l'effet propre de la variable y sont déjà présents.

Le terme d'interaction $g_x(x) \times y$ ne pose pas de problème particulier dans la construction du modèle. Notons $\tilde{\boldsymbol{X}}_x$ et $\tilde{\boldsymbol{S}}_x$ les matrices de design et de pénalisation respectivement associées à g_x (après application de la contrainte de centrage). Le terme d'interaction $g_x(x) \times y$ sera quant à lui associé à la matrice de design $\boldsymbol{X}$ telle que

$$\boldsymbol{X}_i = \tilde{\boldsymbol{X}}_{x,i} \times y_i$$

et à la matrice de pénalisation $\boldsymbol{S}$ telle que

$$\boldsymbol{S} = \tilde{\boldsymbol{S}}_x$$

Chaque élément de la ligne i de la matrice de design initiale est donc tout simplement multiplié par la valeur de la variable y pour l'individu statistique i.

L'intérêt d'un tel modèle est que f_x et g_x sont associées à des paramètres de lissage différents. Ainsi, l'effet propre de x peut être d'une complexité différente de celle de son effet en interaction avec y.

4.4.2 Interactions entre une spline pénalisée et une variable catégorielle

Pénaliser chaque modalité

Une interaction complète entre une spline pénalisée f et une variable catégorielle cat à K modalités cat_k s'écrit

$$f(x, cat) = \sum_{k=1}^{K} f_k(x) \mathbb{1}_{cat=cat_k} \tag{4.2}$$

où f_k représente la spline pénalisée associée à la modalité cat_k. Elle est définie avec la même base que f.

Le modèle est associé à K matrices de pénalisation identiques. En revanche, les K paramètres de lissage correspondant sont a priori tous différents et la complexité de l'effet de x peut varier d'une modalité à l'autre. Cette approche est donc équivalente à ajuster un modèle par modalité en modélisant l'effet de x par la spline f.

Nous pouvons imposer que les paramètres de lissage soient identiques, auquel cas nous obtiendrons la même complexité pour l'effet de x au sein de chaque modalité.

Pénaliser les différences entre modalité

Reprenons l'exemple d'une spline pénalisée f et d'une variable catégorielle cat.

Réécrivons (4.2) en considérant la première modalité comme référence :

$$f(x, cat) = f(x) + \sum_{k=2}^{K} f_k(x) \mathbb{1}_{cat=cat_k} \tag{4.3}$$

Ici, chaque spline f_k représente la différence entre l'effet de x dans la modalité cat_k et l'effet de x dans la modalité de référence (ici la première modalité). En effet,

$$f(x, cat_k) - f(x, cat_1) = f(x) + f_k(x) - f(x) = f_k(x)$$

Sans pénalisation, cette décomposition n'aurait qu'un intérêt restreint. En revanche, ici, l'intérêt est de changer la cible de pénalisation : (4.2) permet de lisser l'effet de x indépendamment au sein de chaque modalité tandis que (4.3) permet de lisser les différences entre modalités.

Le choix entre 4.2 et 4.3 dépend donc de l'échelle sur laquelle on souhaite restituer un effet lissé. La pénalisation sur la différence est particulièrement utile dans les modèles de taux car elle permet de lisser le rapport de taux (*hazard ratio*) plutôt que le taux. Comme nous le verrons par la suite (section 9.4), cela peut être particulièrement pertinent pour modéliser l'effet d'une covariable comme le sexe ou le stade : l'idée est de tirer de l'information d'un stade ou d'un sexe de référence afin de « caler » un stade ou un sexe qui présenterait plus de variabilité (du fait d'une mortalité ou d'un effectif plus réduits).

4.4.3 Splines multidimensionnelles pénalisées

Produit tensoriel

Lorsque l'on ne cherche plus à lisser une courbe mais une surface (ou une hyper-surface de dimension quelconque), et que l'on souhaite lisser différemment selon les différentes directions, il est nécessaire de construire un produit tensoriel de splines marginales (De Boor et al., 1978; Wood, 2006b).

Reprenons l'exemple des deux splines marginales suivantes :

$$f_x(x) = \sum_{i=1}^{I} \alpha_i a_i(x) \qquad f_y(y) = \sum_{j=1}^{J} \beta_j b_j(y)$$

Pour convertir f_x en fonction souple de x et y à la fois, une solution serait que les coefficients α_i varient eux-mêmes de manière souple en fonction de y. Pour ce faire, nous pouvons écrire les α_i comme des splines

$$\alpha_i(y) = \sum_{j=1}^{J} \beta_{ij} b_j(y)$$

Alors la spline bidimensionnelle construite par produit tensoriel (ou plus simplement « le tensor ») serait :

$$f_{x,y}(x,y) = \sum_{i=1}^{I}\sum_{j=1}^{J} \beta_{ij} a_i(x) b_j(y) \tag{4.4}$$

Du point de vue des matrices de design, la ligne i de la matrice de design associée au produit tensoriel est : $\boldsymbol{X}_i = \boldsymbol{X}_{i,x} \otimes \boldsymbol{X}_{i,y}$ où $\otimes$ est le produit de kronecker et avec $\boldsymbol{X}_{i,x}$ et $\boldsymbol{X}_{i,y}$ les lignes i des matrices de design respectivement associées à f_x et f_y.

Cette construction s'étend à un nombre arbitraire de covariables.

Exemple : on considère les deux covariables temps et âge. On dispose des bases suivantes : $f_{temps}(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$ et $f_{age}(a) = \beta_0 + \beta_1 a$.
La spline bidimensionnelle associée sera le produit deux à deux des bases unidimensionnelles. Ainsi :

$$f_{temps,age}(t,a) = (\alpha_0 + \alpha_1 t + \alpha_2 t^2)(\beta_0 + \beta_1 a)$$

$$f_{temps,age}(t,a) = \alpha_0\beta_0 + \alpha_0\beta_1 a + \alpha_1\beta_0 t + \alpha_1\beta_1 ta + \alpha_2\beta_0 t^2 + \alpha_2\beta_1 t^2 a$$

$$f_{temps,age}(t,a) = \gamma_1 + \gamma_2 a + \gamma_3 t + \gamma_4 ta + \gamma_5 t^2 + \gamma_6 t^2 a$$

On remarque que le coefficient β_1 associé à la pente de l'effet de l'âge va varier de manière quadratique en fonction du temps. Dans cet exemple, il y a $2 \times 3 = 6$ paramètres à estimer. Une fois les bases unidimensionnelles spécifiées, toutes les interactions entre chaque élément des bases marginales sont donc gérées par le modèle.

Dans le cadre unidimensionnel, les pénalisations sont de la forme $J_x(f_x) = \lambda_x \int f_x''(x)^2 \mathrm{d}x$ et $J_y(f_y) = \lambda_y \int f_y''(y)^2 \mathrm{d}y$. Pour le produit tensoriel, la construction est plus complexe. Définissons tout d'abord $f_{x|y}$ comme la fonction qui à x associe $f_{x,y}(x,y)$ et $f_{y|x}$ comme la fonction qui à y associe $f_{x,y}(x,y)$.

Une pénalisation naturelle pour $f_{x,y}$ serait alors :

$$J(f_{x,y}) = \lambda_x \int_y J_x(f_{x|y})\mathrm{d}y + \lambda_y \int_x J_y(f_{y|x})\mathrm{d}x$$

Dans le cas d'une pénalisation sur la dérivée seconde, cela donne :

$$J(f_{x,y}) = \int_{x,y} \lambda_x \left(\frac{\partial^2 f_{x,y}}{\partial x^2}\right)^2 + \lambda_y \left(\frac{\partial^2 f_{x,y}}{\partial y^2}\right)^2 \mathrm{d}x\mathrm{d}y$$

Les constructions présentées ici dans le cadre bidimensionnel sont généralisables à tout nombre de covariables. Reiss et al. (2014) donnent une manière exacte de calculer $J(f_{x,y})$. Toutefois, Wood (2017) propose une approximation très satisfaisante en pratique. Après une reparamétrisation adéquate (voir la section 5.6.2 de Wood 2017), le produit tensoriel est associé à autant de matrices de pénalisation qu'il y a de splines marginales. Dans notre exemple bidimensionnel, nous avons les deux matrices

$$\bar{\boldsymbol{S}}_x = \boldsymbol{S}_x \otimes \boldsymbol{I}_J \qquad \bar{\boldsymbol{S}}_y = \boldsymbol{I}_I \otimes \boldsymbol{S}_y$$

$\boldsymbol{S}_x$ et $\boldsymbol{S}_y$ étant les matrices de pénalisations marginales et $\boldsymbol{I}_J$ et $\boldsymbol{I}_I$ les matrices identité de dimension J et I respectivement (dans l'exemple bidimensionnel ci-dessus nous avions $I = 3$ pour le temps et $J = 2$ pour l'âge).

Décomposition du produit tensoriel en effets propres + interactions

Le terme principal du tensor présenté dans la formule (4.4) s'écrit :

$$f_{x,y}(x, y) = \sum_{i=1}^{I} \sum_{j=1}^{J} \beta_{ij} a_i(x) b_j(y)$$

Les effets propres des deux covariables ainsi que leurs interactions sont intégrés dans une même composante du modèle. Toutefois, il peut être intéressant de séparer cet unique terme en les trois suivants :

$$f_x(x) + f_y(y) + \tilde{f}_{x,y}(x, y) \tag{4.5}$$

où $\tilde{f}_{x,y}$ correspond au tensor sans les effets propres de x et y. Wood parle d'*ANOVA decompositions of smooths* (voir la section 5.6.3 de Wood 2017). Cette décomposition trouve son sens dans la pénalisation. En effet, là où dans le tensor classique, $f_{x,y}$ est associé à deux paramètres de lissage, le tensor décomposé en effets propres et interactions est associé à quatre paramètres de lissage : un pour f_x, un pour f_y et deux pour $\tilde{f}_{x,y}$.

L'intérêt de cette décomposition est de pouvoir pénaliser les effets propres et les interactions de manière indépendante. Par exemple, si l'on s'attend à un effet propre complexe pour x mais à une forme fonctionnelle plus simple dans son interaction avec y, alors la décomposition (4.5) paraît intéressante. En revanche, cette décomposition a un coût en termes d'estimation de paramètres de lissage. Par exemple, pour trois covariables, le tensor classique est associé à trois paramètres de lissage tandis que la décomposition engendre douze paramètres de lissage à estimer. En effet, la décomposition avec trois covariables x, y et z s'écrit avec les termes

$$f_x(x) + f_y(y) + f_z(z) + \tilde{f}_{x,y}(x, y) + \tilde{f}_{x,z}(x, z) + \tilde{f}_{y,z}(y, z) + \tilde{f}_{x,y,z}(x, y, z)$$

respectivement associés à $1 + 1 + 1 + 2 + 2 + 2 + 3 = 12$ paramètres de lissage.

Chapitre 5

Inférence

Dans ce chapitre, nous rappellerons les propriétés de l'estimateur du maximum de vraisemblance. Nous présenterons ensuite l'estimateur du maximum de vraisemblance pénalisée accompagné de ses propriétés dans un cadre fréquentiste puis bayésien.

5.1 Estimateur du maximum de vraisemblance

À partir d'un vecteur d'observations $\boldsymbol{y}$, d'un modèle m et d'un vecteur de paramètres de régression $\boldsymbol{\beta}$ de taille p, la log-vraisemblance associée s'écrit $l(\boldsymbol{\beta}|\boldsymbol{y}, m)$, que l'on simplifie souvent en $l(\boldsymbol{\beta})$.

On définit l'estimateur du maximum de vraisemblance (MLE) comme suit :

$$\hat{\boldsymbol{\beta}}_{MLE} = \underset{\boldsymbol{\beta}}{argmax}\,[l(\boldsymbol{\beta})]$$

Notons que le maximum global peut très bien ne pas exister ou bien ne pas être unique. En outre, il peut exister plusieurs maxima locaux de la log-vraisemblance.

Sous certaines conditions de régularité de la vraisemblance (Casella and Berger, 2008, section 10.6.2), le MLE a les propriétés remarquables suivantes :

1. Il est consistant :

$$\hat{\boldsymbol{\beta}}_{MLE} \underset{n\to+\infty}{\xrightarrow{P}} \boldsymbol{\beta}_0$$

   avec $\boldsymbol{\beta}_0$ le vecteur des vraies valeurs des paramètres de régression et $\xrightarrow{P}$ la convergence en probabilité.

2. Il est asymptotiquement normalement distribué :

$$\hat{\boldsymbol{\beta}}_{MLE} \underset{n\to+\infty}{\sim} \mathcal{N}(\boldsymbol{\beta}_0, \mathcal{I}(\boldsymbol{\beta}_0)^{-1})$$

   Avec $\mathcal{I}$ l'information de Fisher définie par :

$$\mathcal{I}(\boldsymbol{\beta}) = -\mathbb{E}_{\mathbf{y}}\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] \tag{5.1}$$

3. Il est asymptotiquement efficace, c'est à dire qu'il est de variance minimale quand $n \to +\infty$.

4. Il est invariant, c'est à dire que

$$\widehat{g(\boldsymbol{\beta})}_{MLE} = g(\hat{\boldsymbol{\beta}}_{MLE})$$

pour toute fonction g.

En pratique, comme on ne connaît pas $\boldsymbol{\beta}_0$, si l'on veut construire des intervalles de confiance pour nos estimations, on approche la variance asymptotique $\mathcal{I}(\boldsymbol{\beta}_0)^{-1}$ par $\mathcal{I}(\hat{\boldsymbol{\beta}}_{MLE})^{-1}$.

En outre, la loi des grands nombres nous assure une deuxième approximation en remplaçant $-\mathbb{E}_{\mathbf{y}}\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right]$ par $-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$.

On note :

$$\mathcal{J}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$$

l'information de Fisher observée.

En pratique, comme ce qui a été fait pour $\mathcal{I}$, on remplacera $\mathcal{J}(\boldsymbol{\beta}_0)$ par $\mathcal{J}(\hat{\boldsymbol{\beta}}_{MLE})$.

5.2 Estimateur du maximum de vraisemblance pénalisée

Soit $\boldsymbol{\lambda}$ un vecteur de M paramètres de lissage connus. En reprenant les mêmes écritures que précédemment, la log-vraisemblance pénalisée s'écrit :

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = l(\boldsymbol{\beta}) - \sum_{m=1}^{M} \frac{\lambda_m}{2} \boldsymbol{\beta}^T \boldsymbol{S}^m \boldsymbol{\beta} \tag{5.2}$$

Le facteur $\frac{1}{2}$ multipliant les paramètres de lissage n'a aucun impact dans l'estimation mais permet de simplifier l'écriture des dérivées de la log-vraisemblance pénalisée.

Maximiser $\mathcal{L}$ en $\boldsymbol{\beta}$ conduit à l'estimateur du maximum de vraisemblance pénalisée (MPLE) :

$$\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE} = \underset{\boldsymbol{\beta}}{argmax}[\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda})]$$

Depuis la proposition de Good and Gaskins (1971) d'utiliser la vraisemblance pénalisée, plusieurs auteurs se sont intéressés aux propriétés asymptotiques du MPLE. Ainsi, Silverman (1982) détaille ses propriétés asymptotiques et notamment l'existence, la consistance et la normalité asymptotique sous certaines conditions.

Les propriétés asymptotiques d'un estimateur pénalisé du taux de mortalité en survie sont donnés dans Cox and O'Sullivan (1990) et Gu and Qiu (1994). Dans le cadre des GLM et des modèles *Accelerated Failure Time* (AFT) le lecteur pourra également se tourner vers Gu and Kim (2002).

Commenges et al. (2014) rapportent que la propriété de consistance du MLE est conservée (quand $n \to +\infty$) à condition que la suite des paramètres de lissage λ_n tende vers zéro.

5.2.1 Approche fréquentiste via les M-estimateurs

Le MPLE fait partie de la famille des M-estimateurs (van der Vaart, 1998; Huber, 2011). Cette propriété est notamment utilisée par Commenges et al. (2014) afin de montrer la normalité asymptotique du MPLE (voir la section 3.3 de Commenges et al. 2014 et la section 6.3 de Huber 2011 pour plus de détails) :

$$\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE} \underset{n\to+\infty}{\sim} \mathcal{N}\left(E(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}), \boldsymbol{V}_{\hat{\beta}}\right)$$

avec

$$\boldsymbol{V}_{\hat{\beta}} = \left(\mathcal{I}(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}) + \boldsymbol{S}^{\lambda}\right)^{-1} \mathcal{I}(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE})^{-1} \left(\mathcal{I}(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}) + \boldsymbol{S}^{\lambda}\right)^{-1} \tag{5.3}$$

La matrice $\boldsymbol{V}_{\hat{\beta}}$, notamment utilisée par Gray (1992), conduit à des probabilités de couverture bien en-dessous de la valeur cible de 95% car elle ne tient pas compte du biais induit par la pénalisation. En effet, en général, on a $E(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}) \neq \boldsymbol{\beta}_0$. Toutefois, cette variance reste utile dans la construction de tests statistiques (voir la section 4.8 de Wood 2006a).

5.2.2 Approche Bayésienne

Le MPLE peut être considéré comme un estimateur du maximum a posteriori (MAP) dans un cadre bayésien (Leonard, 1978). Cette idée est fondamentale si l'on cherche à faire de l'inférence à partir de modèles pénalisés.

Fahrmeir et al. (2010) proposent la description suivante. Considérons un vecteur $\boldsymbol{y}$ à expliquer. Le modèle est défini par la distribution conditionnelle :

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\lambda})$$

Avec $\boldsymbol{\beta}$ le vecteur de paramètres de régression d'intérêt et $\boldsymbol{\lambda}$ le vecteur de paramètres secondaires (ou paramètres de nuisance) représentant par exemple la variance dans un modèle linéaire ou les paramètres de lissage dans un cadre pénalisé.

On spécifie une distribution a priori sur les $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}|\boldsymbol{\lambda})$$

On peut alors également spécifier la distribution a priori $p(\boldsymbol{\lambda})$ pour $\boldsymbol{\lambda}$.

La distribution a posteriori résumant le modèle est donc :

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\lambda}) p(\boldsymbol{\beta}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda})$$

On se rend mieux compte du lien entre cette approche bayésienne et la vraisemblance pénalisée si l'on considère $p(\boldsymbol{\lambda})$ comme fixe. En effet, l'estimateur MAP de $\boldsymbol{\beta}$ est :

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{MAP} &= \underset{\boldsymbol{\beta}}{argmax} \left[p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\lambda}) p(\boldsymbol{\beta}|\boldsymbol{\lambda})\right] \\
&= \underset{\boldsymbol{\beta}}{argmax} \left[log\{p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\lambda})\} + log\{p(\boldsymbol{\beta}|\boldsymbol{\lambda})\}\right]
\end{aligned}$$

Wood et al. (2016b) proposent alors le formalisme suivant :

$$p(\boldsymbol{\beta}|\boldsymbol{\lambda}) = \frac{|\boldsymbol{S}^{\lambda}|_{+}^{1/2}}{\sqrt{2\pi}^{p-M_0}} exp\left(-\frac{\boldsymbol{\beta}^T \boldsymbol{S}^{\lambda} \boldsymbol{\beta}}{2}\right)$$

Avec

$$\boldsymbol{S}^{\lambda} = \sum_{m=1}^{M} \lambda_m \boldsymbol{S}^m$$

et M_0 est le nombre de valeurs propres nulles de $\boldsymbol{S}^{\lambda}$. L'opérateur matriciel $|.|_+$ correspond au produit des valeurs propres non nulles. On rappelle également que p correspond à la longueur du vecteur $\boldsymbol{\beta}$.

Finalement, il vient :

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{MAP} &= \underset{\boldsymbol{\beta}}{argmax}\left[l(\boldsymbol{\beta}) - \frac{\boldsymbol{\beta}^T \boldsymbol{S}^{\lambda} \boldsymbol{\beta}}{2}\right] \\ &= \hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}\end{aligned}$$

L'approche Bayésienne permet d'écrire $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{S}^{\lambda-})$ a priori, avec $\boldsymbol{S}^{\lambda-}$ une matrice pseudo-inverse de Moore-Penrose de $\boldsymbol{S}^{\lambda}$.

On obtient ainsi une loi a posteriori asymptotique pour $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\lambda} \underset{n\to+\infty}{\sim} \mathcal{N}\left(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}, \left(\mathcal{I}(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}) + \boldsymbol{S}^{\lambda}\right)^{-1}\right) \tag{5.4}$$

Les détails sont donnés dans le *Supplementary Material* B.4 de Wood et al. (2016b).

5.3 Variance des estimateurs

L'approximation Bayésienne asymptotique décrite ci-dessus conduit à la matrice de covariance a posteriori suivante :

$$\boldsymbol{V}_{\beta} = \left(\mathcal{I}(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}) + \boldsymbol{S}^{\lambda}\right)^{-1}$$

En pratique, $\mathcal{I}(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE})$ est remplacée par la matrice observée :

$$\boldsymbol{H} = -\frac{\partial^2 l(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$$

En outre, on note

$$\boldsymbol{\mathcal{H}} = \boldsymbol{H} + \boldsymbol{S}^{\lambda}$$

De telle sorte, qu'en pratique,

$$\boldsymbol{V}_{\beta} = \boldsymbol{\mathcal{H}}^{-1}$$

Cette matrice de covariance Bayésienne a été décrite pour la première fois par Wahba (1983). En pratique, cette variance Bayésienne est très utilisée au vu de ses très bonnes propriétés en termes de probabilités de couverture. En effet, Nychka (1988) a montré que cette matrice tient compte du biais engendré par la pénalisation et assure donc une couverture satisfaisante **sur l'ensemble des valeurs des covariables** (cette bonne couverture moyenne peut toutefois être compensée par des sur- et sous-couvertures locales).

Avec les écritures précédentes, la variance fréquentiste (5.3) devient :

$$\boldsymbol{V}_{\hat{\beta}} = \boldsymbol{\mathcal{H}}^{-1}\boldsymbol{H}\boldsymbol{\mathcal{H}}^{-1}$$

Chapitre 6

Estimation des paramètres de lissage

Lorsque tous les paramètres de lissage $\boldsymbol{\lambda}$ associés à notre modèle d'étude sont connus, nous pouvons estimer les paramètres de régression $\hat{\boldsymbol{\beta}}^{\lambda}$ en maximisant la log-vraisemblance pénalisée $\mathcal{L}$.

Toutefois, en pratique, les paramètres de lissage sont inconnus et il convient de les estimer.

Deux types de critères sont utilisés pour estimer ces paramètres :
- la validation croisée : critère LCV
- la maximisation de la vraisemblance marginale des paramètres de lissage : critère LAML

6.1 Complexité et degrés de liberté effectifs

Avant d'estimer les paramètres de lissage, il est nécessaire de quantifier leur impact sur la complexité du modèle pénalisé. Dans le cadre non pénalisé, il est naturel d'associer la complexité d'un modèle à son nombre de paramètres. En revanche, l'idée de complexité dans les modèles pénalisés n'est pas triviale puisqu'elle recouvre des notions diverses (Höge et al., 2018).
En effet, la pénalisation a pour but de contraindre les paramètres estimés ; on s'attendrait donc à ce que les degrés de liberté associés à un modèle pénalisé aient une valeur plus faible que ceux associés à la version non pénalisée du même modèle.
Supposons une spline pénalisée f associée à p paramètres de régression et un paramètre de lissage λ. Intuitivement, lorsque $\lambda \to 0$, f se rapproche d'une spline de régression classique à p paramètres et le nombre de degrés de liberté doit tendre vers p. Tandis que lorsque $\lambda \to \infty$, alors f se rapproche d'une droite linéaire et le nombre de degrés de liberté tend vers 2.

En pratique, afin de déterminer la complexité d'un modèle pénalisé, on utilise le nombre de degrés de liberté effectifs. Pour comprendre sa construction, supposons un modèle gaussien tel que

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

avec $\boldsymbol{y}$ le vecteur à expliquer, $\boldsymbol{X}$ la matrice de design de dimensions (n, p), $\boldsymbol{\beta}$ le vecteur de paramètres à estimer et $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Pour estimer $\boldsymbol{\beta}$, il faut résoudre :

$$\underset{\boldsymbol{\beta}}{argmin} \quad ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2$$

La solution s'écrit :

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

On peut alors exprimer le vecteur des prédictions $\hat{\boldsymbol{y}}$ en fonction de $\boldsymbol{y}$:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Par commutativité de l'opérateur trace, on a :

$$tr\left[\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right] = tr\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\right] = tr(\boldsymbol{I}_p) = p$$

On vient ainsi de faire le lien entre la trace d'une matrice (la *hat matrix*) et le nombre de paramètres d'un modèle gaussien.

Supposons maintenant que les paramètres $\boldsymbol{\beta}$ soient pénalisés par une matrice de pénalisation $\boldsymbol{S}$ et un paramètre de lissage λ. Le problème devient :

$$\underset{\boldsymbol{\beta}}{argmin} \quad ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}$$

Et l'on a :

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{S}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

On définit alors le nombre de degrés de liberté effectifs d'un tel modèle de la manière suivante :

$$\tau = tr\left[\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{S}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\right]$$

En remarquant que $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{H}$ correspond à l'opposé de la hessienne de la log-vraisemblance et que $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{S} = \boldsymbol{\mathcal{H}}$ correspond à l'opposé de la hessienne de la log-vraisemblance pénalisée, on peut étendre la définition de τ à tout modèle possédant une vraisemblance suffisamment régulière :

$$\tau = tr\left[\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{H}\right] \tag{6.1}$$

Cette approximation du nombre de degrés de liberté effectifs (*effective degrees of freedom* ou edf) a notamment été proposée par Hastie and Tibshirani (1990).

6.2 Likelihood Cross Validation (LCV)

Les critères d'erreur de prédictions et les techniques de validation croisée (comme le critère de *generalized cross-validation* ou GCV) sont fréquemment utilisées dans les GAM afin d'estimer les paramètres de lissage (Wood, 2017). Dans les modèles de survie, des techniques de validation croisée fondées sur la vraisemblance ont déjà été proposées par O'Sullivan (1988), Verweij and Van Houwelingen (1993), Liu et al. (2018), et d'autres.

Classiquement, la validation croisée sur l'échelle de la vraisemblance s'écrit de la manière suivante :

$$LCV(\boldsymbol{\lambda}) = -l(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}) + \tau \tag{6.2}$$

avec

$$\tau = tr\left[\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{H}\right]$$

Un des atouts de la validation croisée sur la vraisemblance est sa propriété d'être un estimateur de la vraisemblance espérée (Liquet and Commenges, 2004). Commenges et al. (2007) rappellent les bonnes propriétés de cet estimateur démontrées par plusieurs études de simulation ainsi que son lien étroit avec le critère de Kullback-Leibler. Stone (1977) a en effet montré qu'un critère de *leave-one-out cross-validation* était asymptotiquement équivalent à l'AIC, lui-même dérivé de la divergence K-L. Toutes ces propriétés font du LCV un critère de choix pour l'estimation du ou des paramètres de lissage.

Voyons maintenant comment obtenir la formule (6.2).

Tout d'abord, on cherche à maximiser la log-vraisemblance pénalisée (5.2) :

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = l(\boldsymbol{\beta}) - \sum_{m=1}^{M} \frac{\lambda_m}{2} \boldsymbol{\beta}^T \boldsymbol{S}^m \boldsymbol{\beta}$$

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = l(\boldsymbol{\beta}) - \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{S}^\lambda \boldsymbol{\beta}$$

Le gradient et l'opposé de la hessienne en $\boldsymbol{\beta}$ s'écrivent :

$$\boldsymbol{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \boldsymbol{G}(\boldsymbol{\beta}) - \boldsymbol{S}^\lambda \boldsymbol{\beta}$$

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}, \lambda) = \boldsymbol{H}(\boldsymbol{\beta}) + \boldsymbol{S}^\lambda$$

Dans la suite, $\hat{\boldsymbol{\beta}}^\lambda_{MPLE}$ sera simplifié en $\hat{\boldsymbol{\beta}}$. La validation croisée sur la log-vraisemblance s'écrit :

$$CVL(\boldsymbol{\lambda}) = \sum_{i=1}^{n} l^i(\hat{\boldsymbol{\beta}}^{[-i]}) \tag{6.3}$$

avec

$$l^i(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - l^{[-i]}(\boldsymbol{\beta})$$

la contribution de l'individu i à la log-vraisemblance **non pénalisée**, $l^{[-i]}(\boldsymbol{\beta})$ la log-vraisemblance non pénalisée sans tenir compte de l'individu i, et $\hat{\boldsymbol{\beta}}^{[-i]}$ le vecteur de paramètres de régression qui maximise la vraisemblance pénalisée sans l'individu i (ce dernier terme est le seul qui dépend de $\boldsymbol{\lambda}$ dans l'expression de CVL).

Verweij and Van Houwelingen (1993) détaillent partiellement le passage de (6.3) à (6.2). On peut également citer l'article de Commenges et al. (2007) pour une démonstration plus récente et axée sur la vraisemblance pénalisée.

Notre premier problème est le calcul de $\hat{\boldsymbol{\beta}}^{[-i]}$ car ce dernier nécessite de réajuster le modèle autant de fois qu'il y a d'individus. Afin d'éviter cet écueil, le calcul qui suit permet d'exprimer $\hat{\boldsymbol{\beta}}^{[-i]}$ en fonction de $\hat{\boldsymbol{\beta}}$.

On réalise une approximation de Taylor d'ordre 1 de $\boldsymbol{\mathcal{G}}^{[-i]}$ (le gradient de la log-vraisemblance pénalisée sans l'individu i) autour de $\hat{\boldsymbol{\beta}}$. Pour simplifier l'écriture, la dépendance à $\boldsymbol{\lambda}$ sera implicite :

$$\boldsymbol{\mathcal{G}}^{[-i]}(\boldsymbol{\beta}) \approx \boldsymbol{\mathcal{G}}^{[-i]}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\mathcal{H}}^{[-i]}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

On recalcule l'expression pour $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{[-i]}$

$$0 \approx \boldsymbol{\mathcal{G}}^{[-i]}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\mathcal{H}}^{[-i]}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{[-i]} - \hat{\boldsymbol{\beta}})$$

Il vient donc :

$$\hat{\boldsymbol{\beta}}^{[-i]} \approx \hat{\boldsymbol{\beta}} + \left[\boldsymbol{\mathcal{H}}^{[-i]}(\hat{\boldsymbol{\beta}})\right]^{-1} \boldsymbol{\mathcal{G}}^{[-i]}(\hat{\boldsymbol{\beta}})$$

Or $\mathcal{L}^{[-i]}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}^{i}(\boldsymbol{\beta})$, donc $\boldsymbol{\mathcal{G}}^{[-i]}(\boldsymbol{\beta}) = \boldsymbol{\mathcal{G}}(\boldsymbol{\beta}) - \boldsymbol{\mathcal{G}}^{i}(\boldsymbol{\beta})$ et $\boldsymbol{\mathcal{H}}^{[-i]}(\boldsymbol{\beta}) = \boldsymbol{\mathcal{H}}(\boldsymbol{\beta}) - \boldsymbol{\mathcal{H}}^{i}(\boldsymbol{\beta})$

Et ainsi :

$$\hat{\boldsymbol{\beta}}^{[-i]} \approx \hat{\boldsymbol{\beta}} - \left[\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\mathcal{H}}^{i}(\hat{\boldsymbol{\beta}})\right]^{-1} \boldsymbol{\mathcal{G}}^{i}(\hat{\boldsymbol{\beta}})$$

On néglige alors le terme $\boldsymbol{\mathcal{H}}^{i}(\hat{\boldsymbol{\beta}})$ et il vient :

$$\hat{\boldsymbol{\beta}}^{[-i]} \approx \hat{\boldsymbol{\beta}} - \left[\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}})\right]^{-1} \boldsymbol{\mathcal{G}}^{i}(\hat{\boldsymbol{\beta}})$$

On sait que

$$\boldsymbol{\mathcal{G}}^{i}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\mathcal{G}}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\mathcal{G}}^{[-i]}(\hat{\boldsymbol{\beta}}) = \boldsymbol{G}(\hat{\boldsymbol{\beta}}) - \boldsymbol{G}^{[-i]}(\hat{\boldsymbol{\beta}}) + \left(\boldsymbol{S}^{\lambda^{[-i]}} - \boldsymbol{S}^{\lambda}\right)\hat{\boldsymbol{\beta}}$$

Or $\boldsymbol{S}^{\lambda^{[-i]}} = \boldsymbol{S}^{\lambda}$ et donc $\boldsymbol{\mathcal{G}}^{i}(\hat{\boldsymbol{\beta}}) = \boldsymbol{G}^{i}(\hat{\boldsymbol{\beta}})$.

Finalement, (6.3) devient :

$$CVL(\boldsymbol{\lambda}) \approx \sum_{i=1}^{n} l^{i}\left(\hat{\boldsymbol{\beta}} - \left[\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}})\right]^{-1} \boldsymbol{G}^{i}(\hat{\boldsymbol{\beta}})\right)$$

$$CVL(\boldsymbol{\lambda}) \approx \sum_{i=1}^{n} \left(l^{i}(\hat{\boldsymbol{\beta}}) - \boldsymbol{G}^{i}(\hat{\boldsymbol{\beta}})^{T}\left[\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}})\right]^{-1} \boldsymbol{G}^{i}(\hat{\boldsymbol{\beta}})\right)$$

$$CVL(\boldsymbol{\lambda}) \approx l(\hat{\boldsymbol{\beta}}) - \sum_{i=1}^{n} \left(\boldsymbol{G}^{i}(\hat{\boldsymbol{\beta}})^{T}\left[\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}})\right]^{-1} \boldsymbol{G}^{i}(\hat{\boldsymbol{\beta}})\right)$$

Or $\sum_{i=1}^{n} \left(\boldsymbol{G}^{i}(\hat{\boldsymbol{\beta}})^{T} \boldsymbol{G}^{i}(\hat{\boldsymbol{\beta}})\right)$ et $\boldsymbol{H}(\hat{\boldsymbol{\beta}})^{-1}$ tendent tous deux vers l'information de Fisher lorsque $n \to +\infty$. Nous obtenons donc l'approximation suivante :

$$CVL(\boldsymbol{\lambda}) \approx l(\hat{\boldsymbol{\beta}}) - tr\left[\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}})^{-1} \boldsymbol{H}(\hat{\boldsymbol{\beta}})\right]$$

Et par les simplifications introduites précédemment, il vient

$$CVL(\boldsymbol{\lambda}) \approx l(\hat{\boldsymbol{\beta}}) - tr\left[\boldsymbol{\mathcal{H}}^{-1} \boldsymbol{H}\right]$$

Enfin, en prenant l'opposé on obtient bien la formule (6.2).

6.3 *Laplace approximate marginal likelihood* (LAML)

Dans ce qui suit, par souci de lisibilité, nous poserons $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}$.

Wood et al. (2016b) proposent d'estimer les paramètres de lissage en maximisant le critère suivant :

$$LAML(\boldsymbol{\lambda}) = \mathcal{L}(\hat{\boldsymbol{\beta}}, \boldsymbol{\lambda}) + \frac{1}{2} log\left(|\boldsymbol{S}^{\lambda}|_{+}\right) - \frac{1}{2} log\left(|\boldsymbol{\mathcal{H}}|\right) + \frac{M_0}{2} log(2\pi) \tag{6.4}$$

où $|\boldsymbol{S}^{\lambda}|_{+}$ est le produit des valeurs propres positives de $\boldsymbol{S}^{\lambda}$ et M_0 est le nombre de valeurs propres nulles de $\boldsymbol{S}^{\lambda}$ quand tous les λ_j sont strictement positifs.

Dans le cadre des GAM, ce critère de sélection est connu sous le nom de REML (pour *restricted maximum likelihood*, Wood 2017). Bien qu'il s'agisse du même critère, Wood et al. (2016b) proposent d'utiliser le terme LAML pour *Laplace approximate marginal likelihood* dans le cas général.

Il faut noter que le *Supplementary Material* A de Wood et al. (2016b) démontre que le MPLE demeure consistent lorsque les paramètres de lissage sont estimés par le critère LAML.

Bien que les procédures d'estimation de cette thèse s'inscrivent dans un cadre fréquentiste, le critère LAML trouve son origine dans le bayésianisme. Commençons par reconnaître qu'il existe un parallèle entre pénalisation et a priori bayésien (voir la section 5.2.2). En effet, vouloir pénaliser un effet, cela correspond à avoir un a priori sur la régularité de cet effet. Or, en statistiques, avoir un a priori signifie avoir recours au bayésianisme.

Considérons un modèle de log-vraisemblance $l(\boldsymbol{\beta}) = log(f(\boldsymbol{y}|\boldsymbol{\beta}))$ avec un a priori impropre sur $\boldsymbol{\beta}$ de la forme :

$$f(\boldsymbol{\beta}) = \frac{|\boldsymbol{S}^{\lambda}|_{+}^{1/2}}{\sqrt{2\pi}^{p-M_0}} exp\left(-\frac{\boldsymbol{\beta}^T \boldsymbol{S}^{\lambda} \boldsymbol{\beta}}{2}\right)$$

Nota Bene : on parle d'a priori **impropre** lorsque

$$\int_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} = \infty$$

La loi a priori des $\boldsymbol{\beta}$ dépend donc des hyperparamètres $\boldsymbol{\lambda}$. Là où une approche complètement bayésienne supposerait de choisir une loi a priori sur $\boldsymbol{\lambda}$, Wood et al. (2016b) s'appuient sur un cadre bayésien empirique et s'intéressent à la vraisemblance marginale des paramètres de lissage, à savoir :

$$\int f(\boldsymbol{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta}$$

L'approximation de Taylor de $log\left[f(\boldsymbol{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta})\right]$ autour de $\hat{\boldsymbol{\beta}}$ conduit alors à (6.4) :

$$\begin{aligned}
\int f(\boldsymbol{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} &\approx \int exp\left(l(\hat{\boldsymbol{\beta}}) - \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{\mathcal{H}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{2} - \frac{\hat{\boldsymbol{\beta}}^T \boldsymbol{S}^{\lambda} \hat{\boldsymbol{\beta}}}{2} + log\left(|\boldsymbol{S}^{\lambda}|_{+}^{1/2}\right) - \frac{log(2\pi)(p - M_0)}{2}\right) \mathrm{d}\boldsymbol{\beta} \\
&= exp\left(\mathcal{L}(\hat{\boldsymbol{\beta}}, \boldsymbol{\lambda})\right) |\boldsymbol{S}^{\lambda}|_{+}^{1/2} \sqrt{2\pi}^{M_0 - p} \int exp\left(-\frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{\mathcal{H}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{2}\right) \mathrm{d}\boldsymbol{\beta} \\
&= exp\left(\mathcal{L}(\hat{\boldsymbol{\beta}}, \boldsymbol{\lambda})\right) \sqrt{2\pi}^{M_0} |\boldsymbol{S}^{\lambda}|_{+}^{1/2} / |\boldsymbol{\mathcal{H}}|^{1/2}
\end{aligned}$$

6.4 Lien entre paramètre de lissage et variance d'un effet aléatoire

Le critère LAML permet d'estimer les paramètres de lissage comme s'ils étaient les paramètres de variance d'un a priori Gaussien sur les paramètres de régression. Une conséquence intéressante de ce procédé est qu'il peut fonctionner à l'envers : il est possible de considérer des effets aléatoires comme des splines pénalisées (Wood, 2017, Section 5.8).

Soit une spline pénalisée g caractérisée par son vecteur de paramètres de régression $\boldsymbol{\beta}$, sa matrice de pénalisation $\boldsymbol{S}$ et son paramètre de lissage associé λ.

Alors la densité a priori du vecteur $\boldsymbol{\beta}$ s'écrit :

$$f(\boldsymbol{\beta}) \propto exp\left(-\frac{\lambda}{2}\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}\right)$$

or, pour un vecteur $\boldsymbol{\beta}$ suivant une loi normale multivariée, c'est à dire $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$, on a :

$$f(\boldsymbol{\beta}) \propto exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{I}\boldsymbol{\beta}\right)$$

Ainsi, dans un cadre pénalisé, si l'on note $\lambda = \frac{1}{\sigma^2}$ et $\boldsymbol{S} = \boldsymbol{I}$ (pénalisation Ridge), nous pouvons estimer la variance σ^2 d'un effet aléatoire avec le critère LAML.

6.5 Incertitude sur les paramètres de lissage

La section 5.3 présente deux variantes afin d'estimer la variance du MPLE. Toutefois, aucune de ces approches ne propose de prendre en compte l'incertitude sur les paramètres de lissage. En effet, les propriétés du MPLE sont valables lorsque les paramètres de lissage sont connus (ce qui n'est jamais le cas en pratique). Nous venons de voir deux critères permettant d'estimer ces paramètres. Mais qui dit estimation dit incertitude sur les valeurs estimées.
Wood et al. (2016b) proposent la prise en compte de cette incertitude.

6.5.1 Variance corrigée

Tout d'abord, nous avons besoin de quelques notations. En pratique, l'estimation des paramètres de lissage $\boldsymbol{\lambda}$ est contrainte car ceux-ci doivent être positifs. On définit alors le vecteur $\boldsymbol{\rho} = log(\boldsymbol{\lambda})$ dont les éléments ρ_m sont des réels non contraints.

La formule (5.4) devient alors

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho} \underset{n\to+\infty}{\sim} \mathcal{N}\left(\hat{\boldsymbol{\beta}}^{\rho}, \boldsymbol{V}_{\beta}\right) \tag{6.5}$$

Alors que, en pratique, la formule (6.5) est utilisée en remplaçant simplement $\boldsymbol{\rho}$ par $\hat{\boldsymbol{\rho}}$, Wood et al. (2016b) proposent d'utiliser la distribution asymptotique suivante :

$$\boldsymbol{\rho} \underset{n\to+\infty}{\sim} \mathcal{N}\left(\hat{\boldsymbol{\rho}}, \boldsymbol{V}_{\rho}\right) \tag{6.6}$$

où $\boldsymbol{V}_{\rho}$ est l'inverse de la hessienne de l'opposé de la log-vraisemblance marginale calculée en $\hat{\boldsymbol{\rho}}$.

À partir de (6.5) et (6.6), il vient :

$$\boldsymbol{\beta}|\boldsymbol{y} \overset{d}{=} \hat{\boldsymbol{\beta}}^{\rho^*} + \boldsymbol{R}_{\rho^*}^T \boldsymbol{z}$$

avec $\boldsymbol{\rho}^* \sim \mathcal{N}(\hat{\boldsymbol{\rho}}, \boldsymbol{V}_\rho)$, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{R}_{\rho^*}^T \boldsymbol{R}_{\rho^*} = \boldsymbol{V}_\beta$ et $\overset{d}{=}$ l'égalité en distribution.
Une approximation de Taylor entraîne alors

$$\boldsymbol{\beta}|\boldsymbol{y} \overset{d}{\approx} \hat{\boldsymbol{\beta}}^{\hat{\rho}} + \boldsymbol{J}(\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}) + \boldsymbol{R}_{\hat{\rho}}^T \boldsymbol{z} + \sum_m \left. \frac{\partial \boldsymbol{R}_\rho^T \boldsymbol{z}}{\partial \rho_m} \right|_{\hat{\rho}} (\rho_m - \hat{\rho}_m) \tag{6.7}$$

avec $\boldsymbol{J} = \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\rho}}|_{\hat{\boldsymbol{\rho}}}$
La formule (6.7) conduit à la matrice de covariance *corrigée* suivante :

$$\boldsymbol{V}_\beta' = \boldsymbol{V}_\beta + \boldsymbol{V'} + \boldsymbol{V''} \tag{6.8}$$

avec

$$\boldsymbol{V'} = \boldsymbol{J}\boldsymbol{V}_\rho \boldsymbol{J}^T$$

et

$$V_{jm}'' = \sum_i^p \sum_l^M \sum_k^M \frac{\partial R_{ij}}{\partial \rho_k} V_{\rho,kl} \frac{\partial R_{im}}{\partial \rho_l}$$

Pour rappel, p est le nombre de paramètres de régression et M le nombre de paramètres de lissage.

Cette nouvelle matrice de covariance est une matrice de covariance Bayésienne à laquelle on ajoute deux termes correcteurs.

Le véritable intérêt de cette matrice de covariance corrigée n'est pas la construction d'intervalle de confiance. En effet, Nychka (1988) a montré qu'en moyenne, la matrice Bayésienne $\boldsymbol{V}_\beta$ proposait déjà des propriétés de couverture satisfaisantes. Ainsi, la correction présentée dans (6.8) ne peut alors conduire qu'à de la sur-couverture. En réalité, l'intérêt de $\boldsymbol{V}_\beta'$ réside dans la sélection de modèle (voir ci-dessous).
Pour plus de détails concernant la construction de $\boldsymbol{V}_\beta'$, voir Wood et al. (2016b). Notons également que les techniques présentées ci-dessus sont récentes et font toujours l'objet de discussion (Wood et al., 2016a).

6.5.2 AIC corrigé

Afin de définir un critère de sélection de modèles pénalisés, il est naturel d'utiliser le nombre de degrés de liberté effectifs comme définis par (6.1) en lieu et place du nombre de paramètres k dans la formule classique de l'AIC (3.1).
Le problème d'une telle approche est que le critère qui en résulte est trop enclin à sélectionner des modèles complexes (Greven and Kneib, 2010). Cela est dû au fait qu'il existe une incertitude sur les paramètres de lissage estimés.

Afin de remédier à ce problème, Wood et al. (2016b) proposent la définition d'un AIC corrigé :

$$AIC2 = -2l(\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}) + 2\tau_2$$

avec

$$\tau_2 = tr\left[\boldsymbol{V}'_{\beta}\boldsymbol{H}\right]$$

τ_2 correspond aux degrés de liberté effectifs corrigés pour tenir compte de l'incertitude sur les paramètres de lissage. Pour rappel, la version non corrigée τ s'écrit :

$$\tau = tr\left[\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{H}\right]$$

c'est-à-dire

$$\tau = tr\left[\boldsymbol{V}_{\beta}\boldsymbol{H}\right]$$

On voit alors que la matrice de variance corrigée $\boldsymbol{V}'_{\beta}$ a pour but de remplacer la matrice de variance non corrigée $\boldsymbol{V}_{\beta}$ dans l'écriture de l'AIC.

Deuxième partie

Développement d'un modèle de taux pénalisé

Chapitre 7

Aperçu des modèles de survie pénalisés

Différentes approches pénalisées ont déjà été proposées pour modéliser des temps d'événement. Historiquement, les premières approches développées utilisaient les splines de lissage (Cox and O'Sullivan 1990).

Parmi les méthodes actuelles permettant d'ajuster des modèles de survie pénalisés, on peut distinguer quatre grandes catégories :

- Les méthodes s'appuyant sur les modèles de Poisson pénalisés dans le cadre des GAM grâce à l'équivalence entre vraisemblance de Poisson et vraisemblance d'un modèle de taux (voir la section 2.8).
- Les méthodes s'appuyant sur le lien qui existe entre estimation de paramètres de lissage et estimation de paramètres de variance dans le cadre des modèles mixtes.
- Les approches reposant sur l'inférence Bayésienne. Ces méthodes spécifient une loi a priori sur les paramètres de régression et de lissage et reposent essentiellement sur de l'estimation par *Markov chain Monte Carlo* (MCMC) ou plus récemment par une approximation efficiente nommée *integrated nested Laplace approximation* (INLA).
- Les approches reposant sur l'inférence fréquentiste et qui utilisent notamment la validation croisée pour estimer les paramètres de lissage.

Notez que ces catégories ne sont pas exclusives et que certaines approches peuvent donc appartenir à plus d'une catégorie. Dans ce qui suit se trouve un aperçu non exhaustif de plusieurs approches existantes classées selon les quatre familles décrites ci-dessus, ainsi qu'une présentation des méthodes disponibles pour la modélisation du taux en excès.

7.1 Approche de Poisson

Afin de bénéficier du cadre théorique des GAM, Kauermann (2005), Becher et al. (2009), et Rodríguez-Girondo et al. (2013) ont proposé des modèles de taux pénalisés dans lesquels la procédure d'estimation s'appuie sur une approche de Poisson et une augmentation du jeu de données.
Gasparrini et al. (2017) utilise le modèle de Poisson pénalisé afin d'étendre les modèles non-linéaires à retards échelonnés ou *distributed nonlinear lag models* (DLNM) aux données de survie. Ces modèles sont très utiles pour mesurer l'impact sur le taux de mortalité d'une exposition cumulée à un facteur de risque.
Récemment, Bender et al. (2018a) ont généralisé le recours au modèle de Poisson pénalisé afin de bénéficier de l'ensemble des méthodes des modèles mixtes additifs généralisés ou *generalized additive*

mixed models (GAMM). La résultante de leur approche à travers le package *pammtools* permet ainsi de prendre en compte la non-linéarité, la non-proportionnalité, les interactions mais également les effets aléatoires et les variables dépendantes du temps.
Bien que très pratique, le cadre des GAM oblige à augmenter le jeu de données initial et ainsi à accroître considérablement le temps de calcul et les ressources mémoires nécessaires.

7.2 Modèles mixtes

Dans une approche fondée sur les modèles mixtes, Kneib and Fahrmeir (2007) ont développé un modèle de taux pénalisé permettant de modéliser les effets non-linéaires et dépendants du temps. L'idée fondamentale de cette approche repose sur le lien qui existe entre pénaliser des paramètres de régression et spécifier des effets aléatoires associés à ces coefficients. Cette connexion est notamment explicitée dans la section 6.4 de cette thèse.
Cependant, cette approche présente certaines limites. En effet, elle ne permet pas de modéliser les interactions entre plus de deux variables continues. En outre, les interactions entre le temps et une seule covariable continue ne sont pas disponibles dans la version actuelle du logiciel *BayesX* (version 3.0.2, Brezger et al. 2005, qui inclut l'approche de Kneib and Fahrmeir) ou bien dans le package R *R2BayesX* (version 1.1-1, Umlauf et al. 2015).

7.3 Inférence Bayésienne

Des approches bayésiennes ont également été proposées, notamment par Hennerfeind et al. (2006) qui présentent ce qu'ils nomment des *geoadditive survival models* dans lesquels les paramètres de régression et de lissage sont estimés conjointement par une procédure MCMC. Martino et al. (2011) ont proposé un modèle pour lequel la procédure MCMC relativement chronophage est remplacée par du INLA. Cependant, dans cette dernière approche, les auteurs ne considèrent qu'un modèle de taux proportionnel à l'aide d'une loi de Weibull ou bien un modèle de taux constant par intervalles.
Récemment, Umlauf et al. (2018) ont proposé le package R *bamlss* afin de modéliser des structures additives complexes sur l'échelle de différents indicateurs dont le logarithme du taux instantané. Leur approche inclut les produits tensoriels du temps et des covariables. Bien qu'extrêmement flexible, l'estimation des paramètres s'appuie sur une procédure MCMC particulièrement demandeuse en temps de calcul.

7.4 Inférence Fréquentiste

Rondeau et al. (2003) ont proposé un modèle de taux pénalisé dans lequel le taux de base lui-même (et non son logarithme) est modélisé par une M-spline (Ramsay et al., 1988) cubique pénalisée. Des contraintes de positivité sont utilisées pour les coefficients de la spline. Les paramètres de lissage sont estimés par un critère de validation croisée équivalent au LCV.
Liu et al. (2018) ont récemment présenté un cadre pénalisé afin de modéliser sur différentes échelles fondées sur une transformation monotone de la survie (comme le logarithme du taux cumulé par exemple). Les paramètres de lissage sont estimés par LCV ou bien par le critère GBIC. Malheureusement, cette approche ne permet pas la modélisation sur l'échelle du taux instantané, échelle qui nous paraît la plus naturelle, le taux étant un indicateur de grand intérêt (voir la section 2.4 de cette thèse). De plus, si le choix de modéliser le log du taux cumulé évite le calcul de l'intégrale du taux nécessaire à l'écriture de la vraisemblance, le calcul à effectuer devient alors un problème de dérivation numérique

et des contraintes sont nécessaires pour garantir la positivité du taux. En outre, le processus d'optimisation de Liu et al. (2018) ne s'appuie pas sur le calcul explicite des dérivées de la vraisemblance ni sur celles du critère LCV.

7.5 Modèles de taux en excès pénalisés

Dans le cadre des modèles de taux en excès pénalisés, il faut noter que la littérature est bien moins fournie. Hennerfeind et al. (2008) ont proposé une approche bayésienne dans laquelle ils proposent notamment de modéliser les effets non-linéaires et dépendants du temps à l'aide de P-splines. Cependant, le recours à une procédure MCMC implique une nouvelle fois des temps de calcul particulièrement importants.
Plus récemment, Remontet et al. (2019) ont développé un modèle de taux en excès pénalisé qui s'appuie sur l'approche de Poisson. Bien que flexible, cette approche est fastidieuse à mettre en place car la traditionnelle famille de Poisson n'est plus adaptée. En effet, il est nécessaire d'écrire sa propre famille de GLM afin de prendre en compte les taux attendus. Notons également que l'approche de Poisson est coûteuse en temps de calcul car elle nécessite d'augmenter le jeu de données initial. Par exemple, le modèle proposé par Remontet et al. (2019) n'est pas adapté aux jeux de données volumineux, tels ceux du cancer de la prostate et du cancer du sein (respectivement 72 558 et 96 726 cas ; voir Cowppli-Bony et al. 2017), avec 16 Go de mémoire vive à partir du logiciel R 64-bit (R Core Team, 2018). Pour pallier ce problème, l'utilisation de nouveaux algorithmes dédiés aux données volumineuses (Wood et al., 2015, 2017; Li and Wood, 2019) paraît appropriée. Ces nouveaux algorithmes sont notamment implémentés dans la fonction *bam* du package *mgcv*. Malheureusement, la fonction *bam*, contrairement à la fonction *gam*, ne permet pas d'utiliser la famille de GLM évoquée plus haut et nécessaire à la gestion des taux attendus.

Enfin, l'implémentation de l'approche de Liu et al. (2018) dans le package R *rstpm2* propose la prise en compte de taux de mortalité attendus bien que le modèle sous-jacent ne soit pas publié à ma connaissance. De plus, cette implémentation souffre des problèmes propres à la modélisation sur le taux cumulé (ou toute échelle différente du taux instantané) ainsi qu'à l'absence de calcul explicite des dérivées de la fonction à optimiser. Nous reviendrons toutefois sur cette approche de Liu et al. (2018). car elle offre, théoriquement, des fonctionnalités proches de celles proposées par notre méthode (i.e elle permet d'ajuster des splines multidimensionnelles pénalisées en prenant en compte la mortalité attendue). Les propriétés statistiques de ces deux méthodes seront donc comparées dans une étude de simulation (voir la section 12.2).

Chapitre 8

Proposition d'un modèle de taux pénalisé

8.1 Le modèle

Pour construire notre modèle de taux pénalisé, on suppose que la base marginale du temps t et de chaque covariable continue x_j (parmi un vecteur de covariables $\boldsymbol{x}$) est une combinaison linéaire de k_j fonctions connues b_{ij}, ordinairement des splines

$$g_j(x) = \sum_{i=1}^{k_j} \beta_{ji} b_{ji}(x)$$

et les β_{ji} sont des paramètres à estimer (voir le chapitre 1).

Les interactions entre deux ou plus de deux covariables continues sont considérées comme des produits tensoriels des bases marginales. Par exemple, si g_j est un produit tensoriel du temps t et des covariables x_1 et x_2, il vient :

$$g_j(t, x_1, x_2) = \sum_{i_t=1}^{k_t} \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \beta_{i_t,i_1,i_2} b_{t,i_t}(t) b_{x_1,i_1}(x_1) b_{x_2,i_2}(x_2)$$

(voir le chapitre 4). En utilisant les notations et formulations de Wood et al. (2016b), la forme générale du modèle est la suivante :

$$log\{h(t, \boldsymbol{x})\} = \sum_{j=1}^{J} g_j(t, \boldsymbol{x})$$

Dans cette formule, h est le taux instantané et les g_j sont J splines. Chaque fonction g_j peut être soit un produit tensoriel de deux ou plus éléments de $(t, \boldsymbol{x})$, soit la base marginale d'une covariable ou bien la base marginale du temps. Une des caractéristiques importantes de ce modèle est qu'il inclut la modélisation flexible du taux de base, d'effets dépendants du temps et non-linéaires, ainsi que des interactions entre covariables continues via l'utilisation du produit tensoriel.
$\boldsymbol{\beta}$ étant le vecteur de taille p contenant les paramètres à estimer du modèle, chaque fonction g_j peut être pénalisée ou non et ainsi être associée à aucun, un ou plusieurs termes de pénalité (voir la section 4.2.2)

$$\lambda_m \boldsymbol{\beta}^T \boldsymbol{S}^m \boldsymbol{\beta}$$

où les $\boldsymbol{S}^m$ sont des matrices symétriques semi-définies positives connues et les λ_m sont des paramètres de lissage inconnus formant le vecteur $\boldsymbol{\lambda}$.
La construction des $\boldsymbol{S}^m$ à partir des dérivées secondes des g_j est couramment employée mais d'autres formes de pénalisation (comme la Ridge par exemple) sont évidemment possibles (voir le chapitre 4).
Une fois que les M matrices de pénalisation $\boldsymbol{S}^m$ ont été spécifiées, la matrice de pénalisation du modèle complet $\boldsymbol{S}^\lambda$ s'écrit

$$\boldsymbol{S}^\lambda = \sum_{m=1}^{M} \lambda_m \boldsymbol{S}^m$$

L'objectif est alors d'estimer $\boldsymbol{\beta}$ et $\boldsymbol{\lambda}$.
Notons enfin que tout effet paramétrique non pénalisé associé à une variable continue ou catégorielle peut être ajouté dans le modèle.

8.1.1 Exemple de spécification de modèles avec des variables continues

Supposons que l'on veuille modéliser les effets du temps t et de quatre covariables x_1, x_2, x_3, x_4. On souhaite spécifier une interaction flexible entre les effets de t, x_1 et x_2, ainsi qu'un effet non-linéaire et proportionnel de x_3. Enfin, l'effet de x_4 est linéaire et proportionnel. Le modèle correspondant est le suivant :

$$\text{modèle 1 :} \quad log[h(t, x_1, x_2, x_3, x_4)] = g_1(t, x_1, x_2) + g_2(x_3) + \beta_4 \times x_4$$

où g_1 est un produit tensoriel tridimensionnel associé à trois paramètres de lissage et g_2 est une spline cubique naturelle associée à un unique paramètre de lissage de sorte que le nombre total de paramètres de lissage est porté à quatre ($M = 4$).
Supposons que chaque base marginale a une dimension de 5 (intercept inclus), alors g_1 et g_2 sont respectivement associés à 125 et 5 paramètres de régression. Cependant, une contrainte d'identifiabilité doit être appliquée à g_1 et g_2 afin que leurs intercepts ne soient pas confondus (voir la section 4.3.2). Le nombre total de paramètres de régression est donc $p = 125 + 4 + 1 = 130$.

8.1.2 Exemple de spécification de modèle avec des variables catégorielles

Des variables catégorielles peuvent également être introduites dans le modèle. Supposons que, en plus du temps t et des variables continues x_1, x_2, x_3, x_4, l'on s'intéresse à l'effet du sexe (codé 1 pour les hommes et 0 pour les femmes). En gardant les mêmes spécifications que pour le modèle 1, le modèle le plus simple incluant l'effet d'une variable catégorielle est celui qui contient un effet proportionnel :

$$\text{modèle 2 :} \quad log[h(t, x_1, x_2, x_3, x_4, sexe)] = g_1(t, x_1, x_2) + g_2(x_3) + \beta_4 \times x_4 + \beta_5 \times sexe$$

Dans de nombreux cas néanmoins, le modèle 2 est trop restrictif. Des modèles plus complexes peuvent être considérés lorsque l'on souhaite incorporer une variable catégorielle (voir la section 4.4.2) ; par exemple, il est possible de répéter chaque spline pénalisée et chaque terme paramétrique autant de fois qu'il y a de modalités.
Chaque fonction serait ici répétée deux fois, en utilisant les indicatrices I_H pour les hommes et I_F pour les femmes :

$$\begin{aligned}\text{modèle 3 :} \quad log[h(t, x_1, x_2, x_3, x_4, sexe)] &= [g_{1,H}(t, x_1, x_2) + g_{2,H}(x_3) + \beta_{4,H} \times x_4] \times I_H \\ &+ [g_{1,F}(t, x_1, x_2) + g_{2,F}(x_3) + \beta_{4,F} \times x_4] \times I_F\end{aligned}$$

Dans le modèle 3, les paramètres de régression et de lissage sont deux fois plus nombreux que dans le modèle 1.

En pratique, le modèle 2 est trop simple pour être utile et le modèle 3 n'est pas forcément plus utile car il revient à effectuer une analyse séparée chez les hommes et les femmes. Trouver un compromis entre les deux est une tâche difficile et l'AIC corrigé peut être utile dans de telles situations (voir la section 6.5.2).

La pénalisation sur les différences entre modalités, présentée dans la section 4.4.2, peut s'avérer particulièrement utile dans le cadre des modèles de taux. En effet, considérons le modèle suivant :

$$\text{modèle 4 :} \quad log[h(t, sexe)] = g(t) + g_{diff}(t)$$

où g est la spline pénalisée de la modalité de référence (ici les hommes) et g_{diff} est la spline pénalisée représentant la différence entre le log du taux chez les femmes et le log du taux chez les hommes. En effet :

$$log[h(t, F)] - log[h(t, H)] = g(t) + g_{diff}(t) - g(t) = g_{diff}(t)$$

Ce qui équivaut à :

$$log\left[\frac{h(t, F)}{h(t, H)}\right] = g_{diff}(t)$$

Dans le cadre des modèles de taux, la pénalisation sur les différences entre modalités permet donc de pénaliser (et donc de lisser) le rapport de taux (*hazard ratio*) entre modalités. Cette caractéristique est particulièrement intéressante lorsque l'on cherche à garder une cohérence entre modalités (voir la section 9.4 pour une illustration dans le cadre des modèles de taux en excès), par exemple entre modalités d'une variable « stade au diagnostic ».

8.2 Calcul de la log-vraisemblance pénalisée

Soit n le nombre d'individus à analyser, $\boldsymbol{t}$ le vecteur de taille n contenant les temps de suivi de tous les individus et $\boldsymbol{\delta}$ le vecteur de taille n contenant les indicatrices de décès de tous les individus ($\delta_i = 1$ quand l'individu i est décédé, $\delta_i = 0$ lorsqu'il est censuré à droite). Soit $\boldsymbol{X}(\boldsymbol{t})$ la matrice de design telle que :

$$log\{h(t_i; \boldsymbol{x}_i)\}_{1 \leq i \leq n} = \boldsymbol{X}(\boldsymbol{t})\boldsymbol{\beta}$$

D'après (2.4), la contribution d'un individu i à la log-vraisemblance s'écrit :

$$l_i(\boldsymbol{\beta}) = \delta_i log\left[h(t_i; \boldsymbol{x}_i)\right] - \int_0^{t_i} h(u; \boldsymbol{x}_i)du$$

Comme expliqué dans la section 2.6, la prise en compte de données tronquées à gauche est également possible en remplaçant la borne inférieure de l'intégrale par une valeur $t_{0,i}$. Toutefois, afin de simplifier l'ensemble des formules à venir, nous présenterons le cas où $t_{0,i} = 0$.

L'équation (2.5) donne l'approximation par la quadrature de Gauss-Legendre :

$$l_i(\boldsymbol{\beta}) \approx \delta_i \boldsymbol{X}_i \boldsymbol{\beta} - \sum_{k=1}^{q} w_i^k exp\left(\boldsymbol{GL}_i^k \boldsymbol{\beta}\right)$$

La log-vraisemblance et la log-vraisemblance pénalisée du modèle sont respectivement :

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta})$$

et

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = l(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{S}^{\lambda} \boldsymbol{\beta}$$

Dans la suite, l'écriture $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ sera simplifiée en $\mathcal{L}(\boldsymbol{\beta})$.
La procédure d'estimation des paramètres de lissage $\boldsymbol{\lambda}$ est fondée sur des itérations de Newton-Raphson externes (*outer algorithm*) et celle des paramètres de régression $\boldsymbol{\beta}$ à $\boldsymbol{\lambda}$ fixés est fondée sur des itérations de Newton-Raphson internes (*inner algorithm*) (voir les sections 8.3 et 8.4).

8.3 Estimation des paramètres de régression à paramètres de lissage fixés (*inner algorithm*)

Lorsque les éléments de $\boldsymbol{\lambda}$ sont fixés, $\hat{\boldsymbol{\beta}}^{\lambda}_{MPLE}$ (simplifié en $\hat{\boldsymbol{\beta}}^{\lambda}$ dans la suite, et voir la section 5.2) est :

$$\hat{\boldsymbol{\beta}}^{\lambda} = \underset{\boldsymbol{\beta}}{argmax}\left[\mathcal{L}(\boldsymbol{\beta})\right]$$

La solution à ce problème s'obtient par un algorithme de Newton-Raphson qui nécessite les dérivées première et seconde de la log-vraisemblance pénalisée par rapport à $\boldsymbol{\beta}$. Les dérivées de la contribution à la log-vraisemblance non pénalisée sont :

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_l} = \delta_i X_{il} - \sum_{k=1}^{q} w_i^k GL_{il}^k exp\left(\boldsymbol{GL}_i^k \boldsymbol{\beta}\right)$$

et

$$\frac{\partial^2 l_i(\boldsymbol{\beta})}{\partial \beta_l \beta_m} = -\sum_{k=1}^{q} w_i^k GL_{il}^k GL_{im}^k exp\left(\boldsymbol{GL}_i^k \boldsymbol{\beta}\right)$$

Le gradient de la log-vraisemblance pénalisée est $\boldsymbol{\mathcal{G}}(\boldsymbol{\beta}) = \left(\sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T - \boldsymbol{S}^{\lambda}\boldsymbol{\beta}$. Par commodité, la hessienne de l'opposé de la log-vraisemblance, notée $\boldsymbol{H}$, s'écrit en notation matricielle :

$$\boldsymbol{H}(\boldsymbol{\beta}) = \sum_{k=1}^{q} \left(\boldsymbol{GL}^k\right)^T \left[\boldsymbol{w}^k \circ \boldsymbol{GL}^k \circ exp\left(\boldsymbol{GL}^k \boldsymbol{\beta}\right)\right]$$

Dans cette équation, $\circ$ est le produit de Hadamard. Le produit de Hadamard entre un vecteur $\boldsymbol{v}$ de taille n et une matrice $\boldsymbol{M}$ de dimension (n,p) est défini de la manière suivante : $(\boldsymbol{v} \circ \boldsymbol{M})_{(i,j)} = (\boldsymbol{M} \circ \boldsymbol{v})_{(i,j)} = v_i M_{ij}$.

La hessienne de l'opposé de la log-vraisemblance pénalisée est donc $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}) = \boldsymbol{H}(\boldsymbol{\beta}) + \boldsymbol{S}^{\boldsymbol{\lambda}}$. Le pas de l'algorithme de Newton-Raphson s'écrit alors $\boldsymbol{\Delta}(\boldsymbol{\beta}) = \boldsymbol{\mathcal{H}}(\boldsymbol{\beta})^{-1}\boldsymbol{\mathcal{G}}(\boldsymbol{\beta})$. Dès que nécessaire, une perturbation diagonale de $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta})$ est effectuée afin de la rendre définie positive et $\boldsymbol{\Delta}$ est divisé par deux jusqu'à ce que la log-vraisemblance soit maximisée (pour plus de détails, voir la section 8.8).

8.4 Les critères LCV et LAML pour l'estimation des paramètres de lissage (*outer algorithm*)

Comme évoqué précédemment, un algorithme de Newton-Raphson est implémenté afin d'estimer les paramètres de lissage $\boldsymbol{\lambda}$. Cet algorithme nécessite les dérivées première et seconde des critères LCV et LAML par rapport aux paramètres de lissage. Par souci de lisibilité, $\hat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}}$, $\boldsymbol{H}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}})$, et $\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}})$ seront respectivement notés $\hat{\boldsymbol{\beta}}$, $\boldsymbol{H}$, et $\boldsymbol{\mathcal{H}}$. Toutes les références concernant le calcul matriciel sont disponibles dans Petersen and Pedersen (2008).

8.4.1 Laplace approximate marginal likelihood criterion (LAML)

Wood et al. (2016b) décrivent une perspective bayésienne empirique dans laquelle les paramètres de lissage sont traités comme des hyperparamètres de variance. Ces paramètres peuvent être estimés en maximisant la vraisemblance marginale en $\boldsymbol{\lambda}$ (critère LAML), qui s'écrit :

$$LAML(\boldsymbol{\lambda}) = \mathcal{L}(\hat{\boldsymbol{\beta}}) + \frac{1}{2}log|\boldsymbol{S}^{\boldsymbol{\lambda}}|_+ - \frac{1}{2}log|\boldsymbol{\mathcal{H}}| + \frac{M_0}{2}log(2\pi)$$

Pour rappel, $|\boldsymbol{S}^{\boldsymbol{\lambda}}|_+$ est le produit des valeurs propres positives de $\boldsymbol{S}^{\boldsymbol{\lambda}}$ et M_0 est le nombre de valeurs propres nulles de $\boldsymbol{S}^{\boldsymbol{\lambda}}$ quand tous les λ_j sont strictement positifs. Maximiser le critère LAML en fonction de $\boldsymbol{\rho} = log(\boldsymbol{\lambda})$ afin que les paramètres de lissage estimés restent positifs durant toute la procédure implique de calculer les dérivées de LAML par rapport à $\boldsymbol{\rho}$.

D'après Wood et al. (2016a), les dérivées de LAML sont :

$$\frac{\partial LAML}{\partial \rho_l} = -\frac{\lambda_l}{2}\hat{\boldsymbol{\beta}}^T\boldsymbol{S}^l\hat{\boldsymbol{\beta}} + \frac{1}{2}\frac{\partial log|\boldsymbol{S}^{\boldsymbol{\lambda}}|_+}{\partial \rho_l} - \frac{1}{2}\frac{\partial log|\boldsymbol{\mathcal{H}}|}{\partial \rho_l}$$

et

$$\frac{\partial^2 LAML}{\partial \rho_l \partial \rho_m} = -\kappa_l^m\frac{\lambda_l}{2}\hat{\boldsymbol{\beta}}^T\boldsymbol{S}^l\hat{\boldsymbol{\beta}} - \frac{\partial \hat{\boldsymbol{\beta}}^T}{\partial \rho_m}\boldsymbol{\mathcal{H}}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} + \frac{1}{2}\frac{\partial^2 log|\boldsymbol{S}^{\boldsymbol{\lambda}}|_+}{\partial \rho_l \partial \rho_m} - \frac{1}{2}\frac{\partial^2 log|\boldsymbol{\mathcal{H}}|}{\partial \rho_l \partial \rho_m}$$

où $\kappa_l^m = 1$ si $l = m$, 0 sinon. Les dérivées du log-déterminant de $\boldsymbol{\mathcal{H}}$ (Wood et al. 2016a) sont les suivantes :

$$\frac{\partial log|\boldsymbol{\mathcal{H}}|}{\partial \rho_l} = tr\left(\boldsymbol{\mathcal{H}}^{-1}\frac{\partial \boldsymbol{\mathcal{H}}}{\partial \rho_l}\right)$$

et

$$\frac{\partial^2 log|\boldsymbol{\mathcal{H}}|}{\partial \rho_l \partial \rho_m} = -tr\left(\boldsymbol{\mathcal{H}}^{-1}\frac{\partial \boldsymbol{\mathcal{H}}}{\partial \rho_l}\boldsymbol{\mathcal{H}}^{-1}\frac{\partial \boldsymbol{\mathcal{H}}}{\partial \rho_m}\right) + tr\left(\boldsymbol{\mathcal{H}}^{-1}\frac{\partial^2 \boldsymbol{\mathcal{H}}}{\partial \rho_l \partial \rho_m}\right) \tag{8.1}$$

Les dérivées de $log|\boldsymbol{S}^{\boldsymbol{\lambda}}|_+$ sont données par (voir la section 8.8.6) :

$$\frac{\partial log|\boldsymbol{S}^{\lambda}|_{+}}{\partial \rho_l} = \frac{\partial log|\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j|}{\partial \rho_l} = tr\left[\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\frac{\partial\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j}{\partial\rho_l}\right] = tr\left[\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\lambda_l\tilde{\boldsymbol{S}}^l\right]$$

et

$$\begin{aligned}\frac{\partial^2 log|\boldsymbol{S}^{\lambda}|_{+}}{\partial\rho_l\partial\rho_m} &= -tr\left[\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\frac{\partial\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j}{\partial\rho_l}\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\frac{\partial\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j}{\partial\rho_m}\right]\\ &\quad + tr\left[\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\frac{\partial^2\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j}{\partial\rho_l\partial\rho_m}\right]\\ &= -tr\left[\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\lambda_l\tilde{\boldsymbol{S}}^l\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\lambda_m\tilde{\boldsymbol{S}}^m\right] + tr\left[\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\kappa_l^m\lambda_l\tilde{\boldsymbol{S}}^l\right]\end{aligned}$$

Dans un contexte de modèle de taux, les dérivées de $\boldsymbol{H}$ et $\boldsymbol{\mathcal{H}}$ sont :

$$\frac{\partial\boldsymbol{H}}{\partial\rho_l} = \sum_{k=1}^{q}\left(\boldsymbol{GL}^k\right)^T\left[\boldsymbol{w}^k\circ\boldsymbol{GL}^k\circ\left\{\left(\boldsymbol{GL}^k\circ exp\left(\boldsymbol{GL}^k\hat{\boldsymbol{\beta}}\right)\right)\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_l}\right\}\right] \tag{8.2}$$

$$\frac{\partial\boldsymbol{\mathcal{H}}}{\partial\rho_l} = \frac{\partial\boldsymbol{H}}{\partial\rho_l} + \lambda_l\boldsymbol{S}^l$$

et

$$\begin{aligned}\frac{\partial^2\boldsymbol{H}}{\partial\rho_l\partial\rho_m} = \sum_{k=1}^{q}\left(\boldsymbol{GL}^k\right)^T\Bigg[\boldsymbol{w}^k\circ\boldsymbol{GL}^k\circ\Bigg\{\Bigg(\boldsymbol{GL}^k\circ\left[\left(\boldsymbol{GL}^k\circ exp\left(\boldsymbol{GL}^k\hat{\boldsymbol{\beta}}\right)\right)\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_m}\right]\Bigg)\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_l}\\ + \left(\boldsymbol{GL}^k\circ exp\left(\boldsymbol{GL}^k\hat{\boldsymbol{\beta}}\right)\right)\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_l\partial\rho_m}\Bigg\}\Bigg]\end{aligned} \tag{8.3}$$

$$\frac{\partial^2\boldsymbol{\mathcal{H}}}{\partial\rho_l\partial\rho_m} = \frac{\partial^2\boldsymbol{H}}{\partial\rho_l\partial\rho_m} + \kappa_l^m\lambda_l\boldsymbol{S}^l \tag{8.4}$$

Les dérivées de $\hat{\boldsymbol{\beta}}$ sont obtenues par différentiation implicite (Wood et al., 2016b). L'idée de la dérivation implicite est la suivante : nous partons de la définition de l'estimateur du maximum de vraisemblance pénalisée

$$\forall i \qquad \frac{\partial\mathcal{L}(\hat{\boldsymbol{\beta}})}{\partial\beta_i} = 0$$

En décomposant la partie gauche, il vient :

$$\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial\beta_i} - \left(\sum_{m=1}^{M}\lambda_m\boldsymbol{S}_i^m\right)\hat{\boldsymbol{\beta}} = 0$$

où $\boldsymbol{S}_i^m$ est la i^e ligne de la matrice $\boldsymbol{S}^m$.

Étant donné que la quantité précédente est nulle, elle reste nulle si nous la dérivons par rapport aux log-paramètres de lissage :

$$\frac{\partial}{\partial \rho_l}\left[\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \beta_i} - \left(\sum_{m=1}^{M} \lambda_m \boldsymbol{S}_i^m\right)\hat{\boldsymbol{\beta}}\right] = 0$$

$$\sum_j \left[\frac{\partial^2 l(\hat{\boldsymbol{\beta}})}{\partial \beta_i \partial \beta_j}\frac{\partial \hat{\beta}_j}{\partial \rho_l}\right] - \left(\lambda_l \boldsymbol{S}_i^l\right)\hat{\boldsymbol{\beta}} - \left(\sum_{m=1}^{M} \lambda_m \boldsymbol{S}_i^m\right)\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} = 0$$

$$-\boldsymbol{H}_i \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} - \left(\lambda_l \boldsymbol{S}_i^l\right)\hat{\boldsymbol{\beta}} - \left(\sum_{m=1}^{M} \lambda_m \boldsymbol{S}_i^m\right)\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} = 0$$

$$-\boldsymbol{\mathcal{H}}_i \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} = \left(\lambda_l \boldsymbol{S}_i^l\right)\hat{\boldsymbol{\beta}}$$

Au final, il vient :

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} = -\boldsymbol{\mathcal{H}}^{-1}\left(\lambda_l \boldsymbol{S}^l\right)\hat{\boldsymbol{\beta}}$$

et

$$\frac{\partial^2 \hat{\boldsymbol{\beta}}}{\partial \rho_l \partial \rho_m} = -\left[\frac{\partial \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_m}\left(\lambda_l \boldsymbol{S}^l\right) + \boldsymbol{\mathcal{H}}^{-1}\left(\lambda_l \boldsymbol{S}^l\right)\kappa_l^m\right]\hat{\boldsymbol{\beta}} - \left[\boldsymbol{\mathcal{H}}^{-1}\left(\lambda_l \boldsymbol{S}^l\right)\right]\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_m}$$

où la dérivée première de $\boldsymbol{\mathcal{H}}^{-1}$ est donnée par :

$$\frac{\partial \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_m} = -\boldsymbol{\mathcal{H}}^{-1}\frac{\partial \boldsymbol{\mathcal{H}}}{\partial \rho_m}\boldsymbol{\mathcal{H}}^{-1}$$

8.4.2 Likelihood cross-validation criterion (LCV)

Le critère LCV est défini comme suit :

$$LCV(\boldsymbol{\lambda}) = -l(\hat{\boldsymbol{\beta}}) + tr\left(\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{H}\right)$$

Ce calcul nous permet d'optimiser le critère LCV via un algorithme de Newton-Raphson. En plus de la stabilité numérique, le principal avantage de cette implémentation est qu'elle n'est pas spécifique à la modélisation du taux. En effet, en spécifiant une autre log-vraisemblance avec ses dérivées première et seconde, les dérivées du critère LCV proposées ici peuvent être utilisées dans n'importe quel contexte couvert par Wood et al. (2016b).

La dérivée première de LCV est :

$$\frac{\partial LCV}{\partial \rho_l} = -\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \rho_l} + \frac{\partial tr\left(\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{H}\right)}{\partial \rho_l}$$

En utilisant la dérivation en chaîne sur le terme $\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \rho_l}$, il vient :

$$\frac{\partial LCV}{\partial \rho_l} = -\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} + \frac{\partial tr\left(\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{H}\right)}{\partial \rho_l}$$

Au final, par dérivation du produit à l'intérieur de la trace :

$$\frac{\partial LCV}{\partial \rho_l} = -\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} + tr\left(\frac{\partial \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_l}\boldsymbol{H} + \boldsymbol{\mathcal{H}}^{-1}\frac{\partial \boldsymbol{H}}{\partial \rho_l}\right)$$

La dérivée seconde du critère LCV est :

$$\frac{\partial^2 LCV}{\partial \rho_l \partial \rho_m} = -\frac{\partial^2 l(\hat{\boldsymbol{\beta}})}{\partial \rho_l \partial \rho_m} + \frac{\partial^2 tr\left(\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{H}\right)}{\partial \rho_l \partial \rho_m}$$

D'après la dérivée première ainsi qu'en utilisant la dérivation en chaîne et la dérivée d'un produit, il vient :

$$\begin{aligned}\frac{\partial^2 LCV}{\partial \rho_l \partial \rho_m} = &-\frac{\partial \hat{\boldsymbol{\beta}}^T}{\partial \rho_m}\boldsymbol{H}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} - \frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial^2 \boldsymbol{\beta}}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l \partial \rho_m} \\ &+ tr\left(\frac{\partial^2 \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_l \partial \rho_m}\boldsymbol{H} + \frac{\partial \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_l}\frac{\partial \boldsymbol{H}}{\partial \rho_m} + \frac{\partial \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_m}\frac{\partial \boldsymbol{H}}{\partial \rho_l} + \boldsymbol{\mathcal{H}}^{-1}\frac{\partial^2 \boldsymbol{H}}{\partial \rho_l \partial \rho_m}\right)\end{aligned} \tag{8.5}$$

avec

$$\frac{\partial^2 \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_l \partial \rho_m} = -\frac{\partial \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_m}\frac{\partial \boldsymbol{H}}{\partial \rho_l}\boldsymbol{\mathcal{H}}^{-1} - \boldsymbol{\mathcal{H}}^{-1}\frac{\partial^2 \boldsymbol{\mathcal{H}}}{\partial \rho_l \partial \rho_m}\boldsymbol{\mathcal{H}}^{-1} - \boldsymbol{\mathcal{H}}^{-1}\frac{\partial \boldsymbol{H}}{\partial \rho_l}\frac{\partial \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_m} \tag{8.6}$$

8.4.3 *Outer algorithm*

Pour chaque nouveau $\boldsymbol{\rho}$ dans la procédure d'optimisation de LAML ou LCV, un algorithme de Newton-Raphson interne est nécessaire afin de trouver $\hat{\boldsymbol{\beta}}^{\lambda}$ (noté $\hat{\boldsymbol{\beta}}^{\rho}$ à partir de maintenant). L'optimisation des critères LAML et LCV est ainsi appelée algorithme externe et conduit à l'estimation $\hat{\boldsymbol{\beta}}^{\hat{\rho}}$. Voir la section 8.8.1 pour plus de détails sur l'implémentation.

8.5 Comparaison de LCV et LAML

Dans cette section, notre objectif est de comparer les comportements des critères LCV et LAML (et notamment leur propension au sur-ajustement) à partir de quelques fichiers de données simulées.

Cette comparaison est librement inspirée de la figure 1 de Wood (2011) qui présentait une comparaison des critères GCV et REML dans le cadre des GAM. La simulation de Wood démontrait la plus forte susceptibilité du critère GCV à donner de multiples extrema et à sous-lisser par rapport au critère REML.

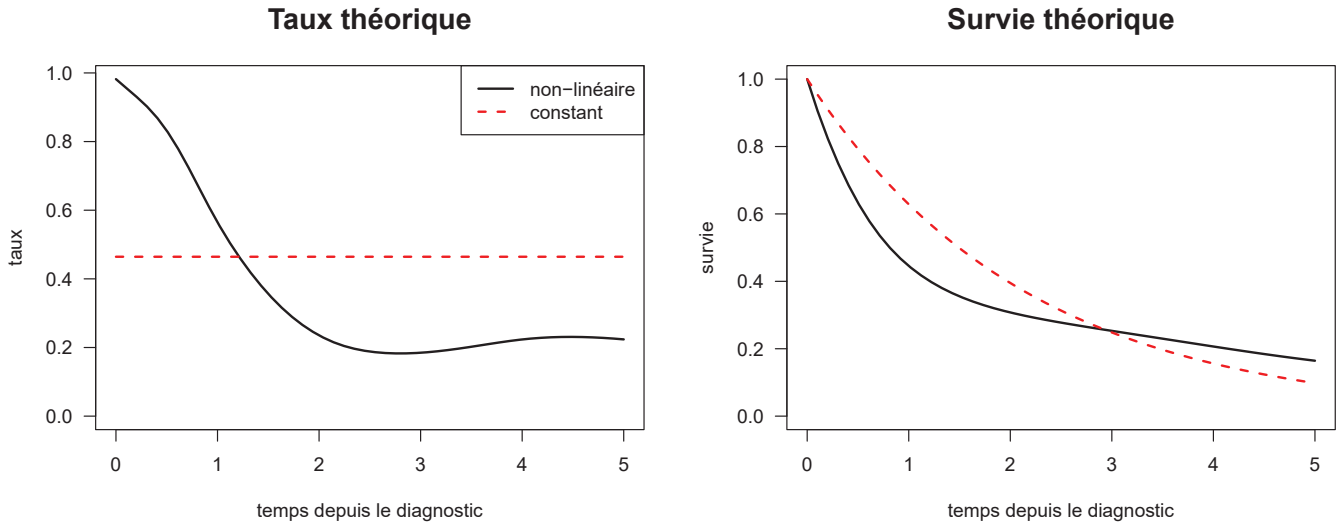

Figure 8.1 – Taux et survie théoriques associés aux scénarios constant et non-linéaire

Dans le cadre de cette thèse, nous explorons deux scénarios : un scénario de taux constant pour lequel les deux critères devraient être théoriquement optimisés pour une valeur infinie de paramètre de lissage ; un scénario de taux fortement non-linéaire (construit à partir de données réelles de cancers de la tête et du cou) pour lequel la valeur du paramètre de lissage optimal doit être faible.

Nous utilisons 10 fichiers de 500 individus pour chaque scénario. Pour chacun des 20 fichiers (10 fichiers $\times$ 2 scénarios), nous modélisons l'effet du temps sur le logarithme du taux de mortalité (toutes causes) par une spline de régression cubique pénalisée à quatre nœuds répartis sur les quantiles des temps observés. Chaque fichier est ajusté avec un paramètre de lissage fixé et différent à chaque fois : les valeurs de paramètre de lissage sont au nombre de 50, équitablement réparties entre -5 et 15 (échelle logarithmique). La figure 8.1 montre les taux et les survies théoriques de chaque scénario.

La figure 8.2 montre l'évolution de $log(LCV)$ et $log(-LAML)$ en fonction du logarithme du paramètre de lissage. Les valeurs de $log(LCV)$ et $log(-LAML)$ sont en fait divisées par la valeur maximum observée sur les 50 paramètres de lissage afin de rapprocher graphiquement les 10 courbes présentées.

Dans le scénario constant, le minimum de LCV est atteint, comme attendu, lorsque $\lambda \to \infty$ ($log(\lambda) > 10$) pour 8 fichiers sur les 10. Toutefois, le critère LCV semble un peu plus erratique que LAML ; nous observons en effet deux cas de sous-lissage dans la fenêtre en haut à gauche contre un seul cas de sous-lissage avec LAML dans la fenêtre en haut à droite.
Concernant le scénario non-linéaire, les minima trouvés par LCV sont clairement moins accentués que ceux trouvées par LAML ; autrement dit, le minimum est plus clairement identifié avec LAML.

Cette simple étude de quelques fichiers simulés semble confirmer les comportements relevés par Wood (2011) dans le cadre des GAM. L'estimation des paramètres de lissage via LAML pourrait ainsi être plus efficiente que par LCV. Toutefois, ces résultats ne présument pas du comportement des deux critères en présence de plusieurs paramètres de lissage.

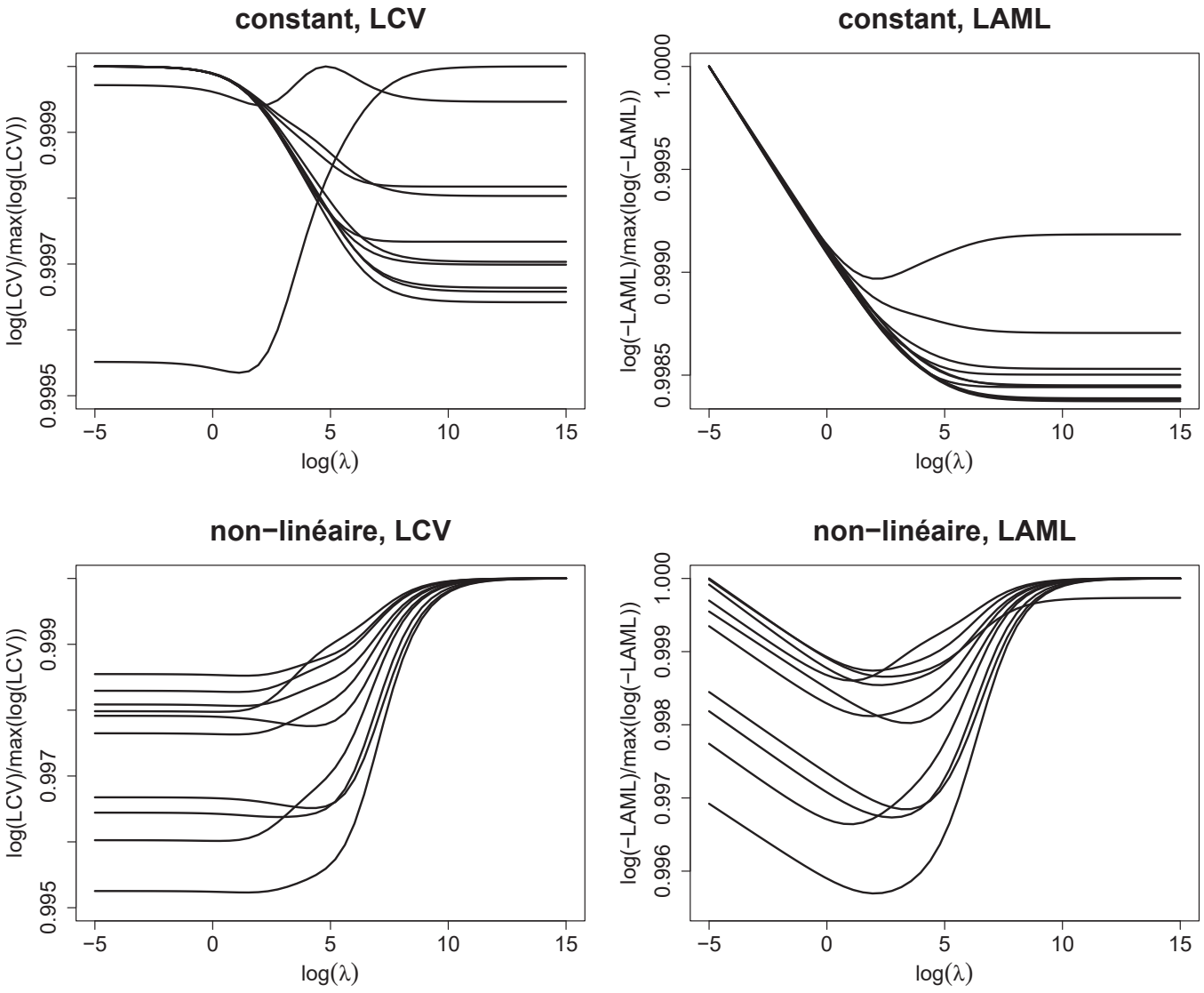

Figure 8.2 – Allures des critères LCV et LAML dans les deux scénarios

8.6 Variance des estimateurs

Dans cette section nous rappelons les trois possibilités pour l'estimation de la variance vues aux sections 5.3 et 6.5.1. Tout d'abord, une approximation Bayésienne asymptotique conduit à la matrice de covariance a posteriori suivante :

$$\boldsymbol{V}_\beta = \boldsymbol{\mathcal{H}}^{-1}$$

Une approche fréquentiste conduit à la matrice de covariance :

$$\boldsymbol{V}_{\hat{\beta}} = \boldsymbol{\mathcal{H}}^{-1}\boldsymbol{H}\boldsymbol{\mathcal{H}}^{-1}$$

Enfin, en plus de l'incertitude sur $\hat{\boldsymbol{\beta}}$, Wood et al. (2016b) ont proposé de prendre en compte l'incertitude provenant de l'estimation des paramètres de lissage. Cela conduit à une troisième option pour le calcul de la variance de $\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{V}'_\beta = \boldsymbol{V}_\beta + \boldsymbol{V'} + \boldsymbol{V''}$$

Cette nouvelle matrice de covariance est une matrice de covariance Bayésienne à laquelle on ajoute deux termes correcteurs (pour plus de détails voir la section 6.5.1).

8.7 Effets aléatoires

Comme vu dans la section 6.4, l'utilisation du critère LAML permet de considérer des effets aléatoires comme des splines pénalisées. Dans notre contexte, considérons le modèle suivant :

$$log\{h(t_i;\boldsymbol{x}_i)\}_{1\leq i\leq n} = \boldsymbol{X}(\boldsymbol{t})\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{t})\boldsymbol{\gamma}$$

avec $\boldsymbol{Z}$ la matrice de design des effets aléatoires et $\boldsymbol{\gamma}$ les paramètres de régression associés. Si l'on note $\boldsymbol{S}$ la matrice de pénalisation associée et λ le paramètre de lissage, l'a priori Gaussien sur $\boldsymbol{\gamma}$ s'écrit

$$f(\boldsymbol{\gamma}) \propto exp\left(-\frac{\lambda}{2}\boldsymbol{\gamma}^T\boldsymbol{S}\boldsymbol{\gamma}\right)$$

Si l'on veut spécifier des effets aléatoires Gaussiens i.i.d, c'est à dire $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ ou $f(\boldsymbol{\gamma}) \propto exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\gamma}^T\boldsymbol{S}\boldsymbol{\gamma}\right)$, on note $\lambda = \frac{1}{\sigma^2}$ et $\boldsymbol{S} = \boldsymbol{I}$. Ainsi, nous pouvons estimer $\boldsymbol{\gamma}$ et λ (et donc σ^2) en considérant une pénalisation Ridge sur $\boldsymbol{\gamma}$ avec une estimation de λ par le critère LAML.

Concernant l'inférence, étant donné que

$$\rho = log(\lambda) = -2log(\sigma)$$

nous pouvons déduire la variance de notre estimateur

$$Var[log(\hat{\sigma})] = \frac{1}{4}\boldsymbol{V}_\rho$$

8.8 Algorithme et techniques numériques

L'implémentation d'un modèle de taux pénalisé requiert une attention particulière quant à la robustesse et à l'efficience des calculs numériques effectués. En effet, la moindre imprécision dans le plus insignifiant des rouages peut mener l'algorithme à un défaut de convergence.

La section suivante a pour but de donner un récapitulatif de l'algorithme développé et des techniques numériques utilisées afin d'assurer sa convergence.

8.8.1 Schéma du double Newton-Raphson

L'optimisation de la vraisemblance pénalisée du modèle de taux présenté dans cette thèse repose sur un algorithme composé de deux algorithmes de Newton-Raphson imbriqués. Le schéma simplifié de ces algorithmes est le suivant :

Choisir $\boldsymbol{\lambda}^0$

while *convergence1 = TRUE* **do**

À partir de $\boldsymbol{\lambda}^k$, choisir $\boldsymbol{\beta}^0_{\boldsymbol{\lambda}^k}$

while *convergence2 = TRUE* **do**

À partir de $\boldsymbol{\beta}^p_{\boldsymbol{\lambda}^k}$, calculer :

$$\mathcal{L}(\boldsymbol{\beta}^p) = l(\boldsymbol{\beta}^p) - \tfrac{1}{2}\boldsymbol{\beta}^{pT}\boldsymbol{S}^{\lambda^k}\boldsymbol{\beta}^p$$

$$\forall l, m,\ \mathcal{G}_l = \frac{\partial \mathcal{L}(\boldsymbol{\beta}^p)}{\partial \beta^p_l} \text{ et } \mathcal{H}_{l,m} = -\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}^p)}{\partial \beta^p_l \partial \beta^p_m}$$

$\boldsymbol{\beta}^{p+1}_{\boldsymbol{\lambda}^k} = \boldsymbol{\beta}^p_{\boldsymbol{\lambda}^k} + \boldsymbol{\mathcal{H}}^{-1}\boldsymbol{\mathcal{G}}$ (pas du Newton-Raphson interne)

end

À partir de $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}^k}$, calculer :

$$critère(\boldsymbol{\lambda}^k) = LCV(\boldsymbol{\lambda}^k) \text{ ou } critère(\boldsymbol{\lambda}^k) = LAML(\boldsymbol{\lambda}^k)$$

$$\forall i, j,\ \mathcal{P}_i = \frac{\partial critère(\boldsymbol{\lambda}^k)}{\partial \lambda^k_i} \text{ et } \mathcal{Q}_{i,j} = -\frac{\partial^2 critère(\boldsymbol{\lambda}^k)}{\partial \lambda^k_i \partial \lambda^k_j}$$

$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \boldsymbol{\mathcal{Q}}^{-1}\boldsymbol{\mathcal{P}}$ (pas du Newton-Raphson externe)

end

Afin de limiter le nombre d'itérations, une fois que $\boldsymbol{\lambda}^k$ est calculé, plutôt que de choisir n'importe quel $\boldsymbol{\beta}^0_{\boldsymbol{\lambda}^k}$, on peut utiliser $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}^{k-1}}$.

8.8.2 Critères de convergence

Dans les schémas de la section 8.8.1, les critères convergence1 et convergence2 sont les suivants.

convergence1

$$\left|critère(\boldsymbol{\lambda}^{k+1}) - critère(\boldsymbol{\lambda}^{k})\right| > tol_{\lambda} \qquad \text{OU} \qquad \exists i \quad \text{tel que} \quad \left|\frac{\partial critère(\boldsymbol{\lambda}^{k+1})}{\partial \lambda_i^{k+1}}\right| > tol_{\lambda}$$

avec $tol_{\lambda} = 10^{-4}$ par défaut.

convergence2

$$\left|\mathcal{L}(\boldsymbol{\beta}^{p+1}) - \mathcal{L}(\boldsymbol{\beta}^{p})\right| > tol_{\beta} \qquad \text{OU} \qquad \exists i \quad \text{tel que} \quad \left|\frac{\beta_i^{p+1} - \beta_i^{p}}{\beta_i^{p}}\right| > tol_{\beta}$$

avec $tol_{\beta} = 10^{-4}$ par défaut.

8.8.3 Complexité algorithmique

La complexité algorithmique du double Newton-Raphson (section 8.8.1) est dominée par le calcul des dérivées troisième et quatrième de la vraisemblance. À savoir les termes

$$\frac{\partial \boldsymbol{H}}{\partial \rho_l}$$

donné par l'équation (8.2) et

$$\frac{\partial^2 \boldsymbol{H}}{\partial \rho_l \partial \rho_m}$$

donné par l'équation (8.3).

En effet, les complexités algorithmiques associées à ces deux termes sont respectivement $O(Mqnp^2)$ et $O(M^2qnp^2)$, avec M le nombre de paramètres de lissage, q le nombre de points associés à la quadrature de Gauss-Legendre, n le nombre d'individus et p le nombre de paramètres de régression.

La complexité algorithmique du processus d'optimisation est alors donnée par le terme dominant $O(M^2qnp^2)$.

Toutefois, si l'on s'intéresse à l'utilisation de la dérivée quatrième $\frac{\partial^2 \boldsymbol{H}}{\partial \rho_l \partial \rho_m}$, on se rend compte qu'elle intervient uniquement dans des opérateurs trace (voir l'équation (8.1) pour LAML et les équations (8.5) et (8.6) pour LCV). Ainsi, dans (8.1), on doit calculer :

$$tr\left(\boldsymbol{\mathcal{H}}^{-1}\frac{\partial^2 \boldsymbol{\mathcal{H}}}{\partial \rho_l \partial \rho_m}\right) \tag{8.7}$$

Pour rappel, le terme $\frac{\partial^2 \boldsymbol{\mathcal{H}}}{\partial \rho_l \partial \rho_m}$ est calculé à partir de $\frac{\partial^2 \boldsymbol{H}}{\partial \rho_l \partial \rho_m}$ (8.4).

La matrice $\boldsymbol{\mathcal{H}}^{-1}$ est symétrique et admet donc une décomposition de la forme :

$$\boldsymbol{\mathcal{H}}^{-1} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^T$$

avec $\boldsymbol{D}$ une matrice diagonale et $\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$.

L'opérateur trace étant indépendant de la base choisie, il vient :

$$tr\left(\boldsymbol{\mathcal{H}}^{-1}\frac{\partial^2\boldsymbol{\mathcal{H}}}{\partial\rho_l\partial\rho_m}\right) = tr\left(\boldsymbol{Q}^T\boldsymbol{\mathcal{H}}^{-1}\frac{\partial^2\boldsymbol{\mathcal{H}}}{\partial\rho_l\partial\rho_m}\boldsymbol{Q}\right)$$

En insérant la matrice identité sous la forme $\boldsymbol{Q}\boldsymbol{Q}^T$, on a :

$$tr\left(\boldsymbol{\mathcal{H}}^{-1}\frac{\partial^2\boldsymbol{\mathcal{H}}}{\partial\rho_l\partial\rho_m}\right) = tr\left(\boldsymbol{Q}^T\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{Q}\boldsymbol{Q}^T\frac{\partial^2\boldsymbol{\mathcal{H}}}{\partial\rho_l\partial\rho_m}\boldsymbol{Q}\right)$$

Et finalement :

$$tr\left(\boldsymbol{\mathcal{H}}^{-1}\frac{\partial^2\boldsymbol{\mathcal{H}}}{\partial\rho_l\partial\rho_m}\right) = tr\left(\boldsymbol{D}\boldsymbol{Q}^T\frac{\partial^2\boldsymbol{\mathcal{H}}}{\partial\rho_l\partial\rho_m}\boldsymbol{Q}\right) \tag{8.8}$$

Le grand intérêt de (8.8) repose sur le fait que $\boldsymbol{D}$ est diagonale. Ainsi, pour calculer la trace, nous avons uniquement besoin des éléments diagonaux de $\boldsymbol{Q}^T\frac{\partial^2\boldsymbol{\mathcal{H}}}{\partial\rho_l\partial\rho_m}\boldsymbol{Q}$ (et même seulement ceux de $\boldsymbol{Q}^T\frac{\partial^2\boldsymbol{H}}{\partial\rho_l\partial\rho_m}\boldsymbol{Q}$). Cette idée a été proposée par Wood (2017, Section 6.6.2) dans le cadre des GAM mais une adaptation est nécessaire pour les modèles de taux.
Dans la cadre des modèles de taux pénalisés, d'après l'équation (8.3), on a :

$$\begin{aligned}\boldsymbol{Q}^T\frac{\partial^2\boldsymbol{H}}{\partial\rho_l\partial\rho_m}\boldsymbol{Q} = \boldsymbol{Q}^T\sum_{k=1}^{q}\left(\boldsymbol{GL}^k\right)^T\Bigg[\boldsymbol{w}^k\circ\boldsymbol{GL}^k\circ\Bigg\{\Bigg(\boldsymbol{GL}^k\circ\Bigg[\left(\boldsymbol{GL}^k\circ exp\left(\boldsymbol{GL}^k\hat{\boldsymbol{\beta}}\right)\right)\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_m}\Bigg]\Bigg)\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_l}\\ + \left(\boldsymbol{GL}^k\circ exp\left(\boldsymbol{GL}^k\hat{\boldsymbol{\beta}}\right)\right)\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_l\partial\rho_m}\Bigg\}\Bigg]\boldsymbol{Q}\end{aligned}$$

$$\begin{aligned}\boldsymbol{Q}^T\frac{\partial^2\boldsymbol{H}}{\partial\rho_l\partial\rho_m}\boldsymbol{Q} = \sum_{k=1}^{q}\left(\boldsymbol{GL}^k\boldsymbol{Q}\right)^T\Bigg[\boldsymbol{w}^k\circ\Bigg\{\Bigg(\boldsymbol{GL}^k\circ\Bigg[\left(\boldsymbol{GL}^k\circ exp\left(\boldsymbol{GL}^k\hat{\boldsymbol{\beta}}\right)\right)\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_m}\Bigg]\Bigg)\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_l}\\ + \left(\boldsymbol{GL}^k\circ exp\left(\boldsymbol{GL}^k\hat{\boldsymbol{\beta}}\right)\right)\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_l\partial\rho_m}\Bigg\}\Bigg]\left(\boldsymbol{GL}^k\boldsymbol{Q}\right)\end{aligned} \tag{8.9}$$

Ainsi, en créant les matrices $\widetilde{\boldsymbol{GL}}^k = \boldsymbol{GL}^k\boldsymbol{Q}$ en amont du processus d'optimisation, on peut économiser beaucoup de temps de calcul afin de calculer (8.9) et (8.7). En effet, (8.9) revient à former la diagonale des produits :

$$(\widetilde{\boldsymbol{GL}}^k)^T\boldsymbol{W}^k(\widetilde{\boldsymbol{GL}}^k) \tag{8.10}$$

où $\boldsymbol{W}^k$ est une matrice diagonale dont les éléments diagonaux sont donnés par le vecteur

$$\boldsymbol{w}^k\circ\Bigg\{\Bigg(\boldsymbol{GL}^k\circ\Bigg[\left(\boldsymbol{GL}^k\circ exp\left(\boldsymbol{GL}^k\hat{\boldsymbol{\beta}}\right)\right)\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_m}\Bigg]\Bigg)\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_l} + \left(\boldsymbol{GL}^k\circ exp\left(\boldsymbol{GL}^k\hat{\boldsymbol{\beta}}\right)\right)\frac{\partial^2\hat{\boldsymbol{\beta}}}{\partial\rho_l\partial\rho_m}\Bigg\}$$

Dans le logiciel R, on peut calculer de manière efficiente la diagonale de (8.10), c'est à dire sans passer par la formation explicite du produit matriciel, en utilisant la commande

$$colSums\left(\widetilde{\boldsymbol{GL}}^k * \widetilde{\boldsymbol{GL}}^k * \boldsymbol{W}^k\right)$$

Le coût du calcul de la dérivée quatrième passe donc de $O(M^2qnp^2)$ à $O(M^2qnp)$. Le coût dominant du calcul des dérivées devient alors le coût associé à la dérivée troisième $O(Mqnp^2)$ (en effet, p est toujours beaucoup plus grand que M).

8.8.4 Perturbation de la hessienne

À une itération donnée, si la matrice $\boldsymbol{\mathcal{H}}$ (respectivement $\boldsymbol{\mathcal{Q}}$) n'est pas définie positive, soit elle n'est pas inversible, soit la direction prise par le pas $\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{\mathcal{G}}$ (respectivement $\boldsymbol{\mathcal{Q}}^{-1}\boldsymbol{\mathcal{P}}$) ne sera pas ascendante. Si c'est le cas, cela signifie que l'itération en cours nous fait perdre du temps en nous envoyant dans la mauvaise direction. Afin d'éviter cela, dès que nécessaire, une perturbation de $\boldsymbol{\mathcal{H}}$ ou de $\boldsymbol{\mathcal{Q}}$ est effectuée. Pour plus de détails, voir Nocedal and Wright (2006, Section 3.4).

8.8.5 Contrôle du pas

Même si le pas nous amène dans la bonne direction, il se peut qu'il nous amène trop loin. Afin d'éviter ce problème et assurer l'optimisation à chaque itération, on peut contrôler la longueur du pas. En pratique, si la vraisemblance pénalisée se met à diminuer d'une itération à l'autre, le pas $\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{\mathcal{G}}$ est divisé par 2 jusqu'à ce qu'elle augmente. En ce qui concerne l'estimation des paramètres de lissage, le pas $\boldsymbol{\mathcal{Q}}^{-1}\boldsymbol{\mathcal{P}}$ est lui divisé par 10 si nécessaire.
En outre, afin de garder chaque ρ_m dans un intervalle de valeurs raisonnables (typiquement $[-15; 15]$), la valeur absolue maximale des éléments de $\boldsymbol{\mathcal{Q}}^{-1}\boldsymbol{\mathcal{P}}$ est contrôlée à chaque itération par un seuil maximum (par défaut, ce seuil est fixé à 5).

8.8.6 Inversion de la matrice de pénalisation

L'un des calculs les plus délicats du processus d'optimisation concerne les dérivées de $log|\boldsymbol{S}^\lambda|_+$. Pour ce faire, nous suivons partiellement Wood et al. (2016b) et diagonalisons

$$\sum_{m=1}^{M} \frac{\boldsymbol{S}^m}{||\boldsymbol{S}^m||_F} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Delta}}\tilde{\boldsymbol{U}}^T$$

avec $||.||_F$ la norme de Frobenius.
Si l'on note $\boldsymbol{U}_+$ les colonnes de $\tilde{\boldsymbol{U}}$ qui correspondent aux valeurs propres strictement positives, on a :

$$\tilde{\boldsymbol{S}}^m = \boldsymbol{U}_+^T\boldsymbol{S}^m\boldsymbol{U}_+$$

Ainsi, $|\boldsymbol{S}^\lambda|_+ = |\sum_{m=1}^{M}\lambda_m\tilde{\boldsymbol{S}}^m|$ avec $\sum_{m=1}^{M}\lambda_m\tilde{\boldsymbol{S}}^m$ une matrice de plein rang.

D'un point de vue informatique, la matrice $\tilde{\boldsymbol{U}}$ est utilisée comme reparamétrisation initiale, avant de débuter la procédure d'optimisation. Cette reparamétrisation est inversée à la convergence. Concernant l'évaluation de $log|\boldsymbol{S}^\lambda|_+$ et de ses dérivées, nous ne suivons pas complètement la procédure décrite dans Wood (2011, Appendix B). En effet, une fois la reparamétrisation effectuée, nous utilisons la décomposition QR suivante :

$$\sum_{m=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^m = \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{R}}$$

$log|\boldsymbol{S}^\lambda|_+$ correspond ainsi à la somme des valeurs absolues des éléments diagonaux de $\tilde{\boldsymbol{R}}$ et les dérivées nécessitent le calcul de $\left(\sum_{m=1}^{M}\lambda_m\tilde{\boldsymbol{S}}^m\right)^{-1}$ qui peut être réalisé de manière stable grâce à cette

décomposition QR (on peut également utiliser une décomposition LU ou encore une décomposition de Cholesky).

8.8.7 Paramètres d'échelle

Dans l'écriture de la log-vraisemblance pénalisée (5.2), la log-vraisemblance est constituée de termes de la forme

$$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$$

où $\boldsymbol{X}$ est la matrice de design associée au modèle et $\boldsymbol{W}$ est une matrice quelconque.

La pénalisation, quant à elle, est constituée des matrices $\boldsymbol{S}^m$.

Pour stabiliser la procédure d'estimation des paramètres de lissage, il est préférable que les différentes matrices de pénalisation $\boldsymbol{S}^m$ soient comparables à la matrice $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ en termes de norme matricielle. On définit alors les facteurs d'échelles suivants :

$$S.scale_m = \frac{norm(\boldsymbol{X})^2}{norm(\boldsymbol{S}^m)}$$

de telle sorte qu'avant le début de l'optimisation, les matrices $\boldsymbol{S}^m$ sont remplacées par :

$$\boldsymbol{S}^m_{scale} = S.scale_m \times \boldsymbol{S}^m$$

Chapitre 9

Proposition d'un modèle de taux en excès pénalisé

Maintenant que nous savons comment construire un modèle de taux pénalisé, il convient d'étendre notre approche aux modèles de taux en excès. En effet, en épidémiologie des maladies chroniques, ce n'est pas tant la mortalité qui nous intéresse mais la mortalité en excès des patients.

9.1 Le modèle

D'après la section 2.9 et l'équation (2.7), dans un modèle de taux en excès, le taux de mortalité observé h_O se décompose de la manière suivante :

$$h_O(t, \boldsymbol{x}, \boldsymbol{z}) = h_E(t, \boldsymbol{x}) + h_P(a + t, \boldsymbol{z})$$

avec h_E est l'excès de mortalité dû à la pathologie étudiée (par exemple le cancer), t est le temps écoulé depuis le diagnostic, a est l'âge au diagnostic, $\boldsymbol{x}$ est un vecteur de covariables ayant un effet sur h_E, et h_P est la mortalité attendue à l'âge $a + t$ dans la population générale avec les caractéristiques démographiques $\boldsymbol{z}$.

En reprenant les notations utilisées dans la section 8.1, le modèle de taux en excès pénalisé s'écrit :

$$log\{h_E(t, \boldsymbol{x})\} = \sum_{j=1}^{J} g_j(t, \boldsymbol{x})$$

où les g_j sont J splines de régression. Si l'on note $\boldsymbol{\beta}$ le vecteur de taille p contenant les paramètres à estimer du modèle, chaque fonction g_j peut être pénalisée ou non et ainsi être associée à aucun, un ou plusieurs termes de pénalité de la forme :

$$\lambda_m \boldsymbol{\beta}^T \boldsymbol{S}^m \boldsymbol{\beta}$$

où les $\boldsymbol{S}^m$ sont des matrices symétriques semi-définies positives connues et les λ_m sont des paramètres de lissage inconnus formant le vecteur $\boldsymbol{\lambda}$.

Une fois que les M matrices de pénalisation $\boldsymbol{S}^m$ ont été spécifiées, la matrice de pénalisation du modèle complet $\boldsymbol{S}^\lambda$ s'écrit

$$\boldsymbol{S}^\lambda = \sum_{m=1}^{M} \lambda_m \boldsymbol{S}^m$$

Comme les modèles de taux pénalisés, les modèles de taux en excès pénalisés proposés dans cette thèse permettent de prendre en compte simultanément :

— les effets non-linéaires des covariables et du temps
— les effets non-proportionnels
— les interactions

L'ajout de termes paramétriques non pénalisés associés à des variables continues ou catégorielles est également possible.

Ces caractéristiques sont essentielles pour répondre aux problématiques épidémiologiques évoquées en introduction.

9.2 Calcul de la vraisemblance et de ses dérivées

Dans le contexte du taux en excès, la contribution d'un individu i à la log-vraisemblance est :

$$l_i(\boldsymbol{\beta}) = \delta_i log\left[h_E(t_i;\boldsymbol{x}_i) + h_P(t_i;\boldsymbol{z}_i)\right] - \int_0^{t_i} h_E(u;\boldsymbol{x}_i)du$$

Par souci de lisibilité, nous écrirons $r_i = h_p(a_i + t_i, \boldsymbol{z}_i)$. La contribution à la log-vraisemblance devient donc :

$$l_i(\boldsymbol{\beta}) = \delta_i log\left[h_E(t_i;\boldsymbol{x}_i) + r_i\right] - \int_0^{t_i} h_E(u;\boldsymbol{x}_i)du$$

La quadrature de Gauss-Legendre nous donne :

$$l_i(\boldsymbol{\beta}) \approx \delta_i log\left[exp\left(\boldsymbol{X}_i\boldsymbol{\beta}\right) + r_i\right] - \sum_{k=1}^{q} w_i^k exp\left(\boldsymbol{GL}_i^k\boldsymbol{\beta}\right)$$

La dérivée première de l_i par rapport à β_l s'écrit :

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_l} = \frac{\delta_i X_{il} exp\left(\boldsymbol{X}_i\boldsymbol{\beta}\right)}{exp\left(\boldsymbol{X}_i\boldsymbol{\beta}\right) + r_i} - \sum_{k=1}^{q} w_i^k GL_{il}^k exp\left(\boldsymbol{GL}_i^k\boldsymbol{\beta}\right)$$

et la dérivée seconde est :

$$\frac{\partial^2 l_i(\boldsymbol{\beta})}{\partial \beta_l \partial \beta_m} = \frac{r_i \delta_i X_{il} X_{im} exp\left(\boldsymbol{X}_i\boldsymbol{\beta}\right)}{\left(exp\left(\boldsymbol{X}_i\boldsymbol{\beta}\right) + r_i\right)^2} - \sum_{k=1}^{q} w_i^k GL_{il}^k GL_{im}^k exp\left(\boldsymbol{GL}_i^k\boldsymbol{\beta}\right)$$

Nota Bene : Dans cette section, nous avons besoin d'utiliser la division de matrices par des vecteurs ainsi que l'exponentiation de vecteurs. Si $\boldsymbol{v}$ est un vecteur de taille n, $\boldsymbol{M}$ une matrice de dimension (n,p), et y un nombre réel : $\left(\frac{\boldsymbol{M}}{\boldsymbol{v}}\right)_{i,j} = \frac{M_{ij}}{v_i}$ et $\left(\boldsymbol{v}^y\right)_i = (v_i)^y$.

Tout comme dans le modèle de taux classique, $\boldsymbol{H}$ peut être exprimée de manière matricielle :

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\boldsymbol{X}^T \left[\frac{\boldsymbol{r} \circ \boldsymbol{\delta} \circ \boldsymbol{X} \circ exp(\boldsymbol{X\beta})}{(exp(\boldsymbol{X\beta}) + \boldsymbol{r})^2} \right] + \sum_{k=1}^{q} \left(\boldsymbol{GL}^k\right)^T \left[\boldsymbol{w}^k \circ \boldsymbol{GL}^k \circ exp\left(\boldsymbol{GL}^k \boldsymbol{\beta}\right) \right]$$

On note $\boldsymbol{H}_1(\boldsymbol{\beta}) = -\boldsymbol{X}^T \left[\frac{\boldsymbol{r} \circ \boldsymbol{\delta} \circ \boldsymbol{X} \circ exp(\boldsymbol{X\beta})}{(exp(\boldsymbol{X\beta}) + \boldsymbol{r})^2} \right]$ et $\boldsymbol{H}_2(\boldsymbol{\beta}) = \sum_{k=1}^{q} \left(\boldsymbol{GL}^k\right)^T \left[\boldsymbol{w}^k \circ \boldsymbol{GL}^k \circ exp\left(\boldsymbol{GL}^k \boldsymbol{\beta}\right) \right]$.

Enfin, pour estimer les paramètres de lissage en survie nette, nous avons besoin des dérivées de $\boldsymbol{H}(\hat{\boldsymbol{\beta}})$ par rapport aux paramètres de lissage. En fait, les dérivées de $\boldsymbol{H}_2(\hat{\boldsymbol{\beta}})$ sont déjà données dans le contexte de la survie brute dans (8.2) et (8.3). Ainsi, seules les dérivées de $\boldsymbol{H}_1(\hat{\boldsymbol{\beta}})$, noté $\boldsymbol{H}_1$ dans la suite, sont nécessaires :

$$\frac{\partial \boldsymbol{H}_1}{\partial \rho_l} = -\boldsymbol{X}^T \left[\boldsymbol{r} \circ \boldsymbol{\delta} \circ \boldsymbol{X} \circ \left\{ \left(\frac{\boldsymbol{X} \circ exp(\boldsymbol{X\hat{\beta}}) \circ \left(\boldsymbol{r} - exp(\boldsymbol{X\hat{\beta}}) \right)}{\left(exp(\boldsymbol{X\hat{\beta}}) + \boldsymbol{r} \right)^3} \right) \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} \right\} \right]$$

et

$$\frac{\partial^2 \boldsymbol{H}_1}{\partial \rho_l \partial \rho_m} = -\boldsymbol{X}^T \left[\boldsymbol{r} \circ \boldsymbol{\delta} \circ \boldsymbol{X} \circ \left\{ \left(\frac{\boldsymbol{X} \circ \boldsymbol{X} \circ exp(\boldsymbol{X\hat{\beta}}) \circ \left[exp(\boldsymbol{X\hat{\beta}}) \circ exp(\boldsymbol{X\hat{\beta}}) - 4\boldsymbol{r} \circ exp(\boldsymbol{X\hat{\beta}}) + \boldsymbol{r}^2 \right]}{\left(exp(\boldsymbol{X\hat{\beta}}) + \boldsymbol{r} \right)^4} \right) \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} + \left(\frac{\boldsymbol{X} \circ exp(\boldsymbol{X\hat{\beta}}) \circ \left(\boldsymbol{r} - exp(\boldsymbol{X\hat{\beta}}) \right)}{\left(exp(\boldsymbol{X\hat{\beta}}) + \boldsymbol{r} \right)^3} \right) \frac{\partial^2 \hat{\boldsymbol{\beta}}}{\partial \rho_l \partial \rho_m} \right\} \right]$$

Enfin, afin de calculer (8.7) de manière efficiente, on calcule les dérivées de $\boldsymbol{H}_1$ selon le même principe que (8.9). C'est à dire :

$$\boldsymbol{Q}^T \frac{\partial^2 \boldsymbol{H}_1}{\partial \rho_l \partial \rho_m} \boldsymbol{Q} = -\left(\boldsymbol{XQ}\right)^T \boldsymbol{W} \left(\boldsymbol{XQ}\right)$$

où $\boldsymbol{W}$ est une matrice diagonale dont les éléments diagonaux sont donnés par le vecteur

$$\boldsymbol{w} = \boldsymbol{r} \circ \boldsymbol{\delta} \circ \left\{ \left(\frac{\boldsymbol{X} \circ \boldsymbol{X} \circ exp(\boldsymbol{X\hat{\beta}}) \circ \left[exp(\boldsymbol{X\hat{\beta}}) \circ exp(\boldsymbol{X\hat{\beta}}) - 4\boldsymbol{r} \circ exp(\boldsymbol{X\hat{\beta}}) + \boldsymbol{r}^2 \right]}{\left(exp(\boldsymbol{X\hat{\beta}}) + \boldsymbol{r} \right)^4} \right) \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} + \left(\frac{\boldsymbol{X} \circ exp(\boldsymbol{X\hat{\beta}}) \circ \left(\boldsymbol{r} - exp(\boldsymbol{X\hat{\beta}}) \right)}{\left(exp(\boldsymbol{X\hat{\beta}}) + \boldsymbol{r} \right)^3} \right) \frac{\partial^2 \hat{\boldsymbol{\beta}}}{\partial \rho_l \partial \rho_m} \right\}$$

9.3 Modèle de taux en excès pénalisé à effets mixtes

La section 8.7 présente les modèles de taux pénalisés à effets mixtes. La généralisation de l'approche aux modèles de taux en excès permet d'inclure très facilement un intercept aléatoire dans un modèle de taux en excès.

Ainsi, les modèles proposés par Charvat et al. (2016) sont directement transposables dans un cadre pénalisé et l'équation (2.7) devient :

$$h_O(t, x_{ij}, z_{ij}) = h_E(t, x_{ij})exp(w_j) + h_P(a_{ij} + t, z_{ij})$$

pour chaque individu i d'un *cluster* j, c'est à dire d'un ensemble regroupant plusieurs individus (par exemple une zone géographique comme le département). Le terme w_j est un effet aléatoire au niveau du *cluster* et $exp(w_j)$ est appelée la fragilité (*frailty*).

L'inconvénient des splines pénalisées par rapport à un cadre classique de modèles mixtes réside dans la nécessité d'estimer les w_j (lors de l'optimisation) et pas seulement leur variance. Ainsi, lorsque le nombre de clusters est très important, l'approche proposée devient difficilement praticable.
De plus, cette approche fondée sur la pénalisation Ridge ne permet pas d'intégrer tout type de structures d'effets aléatoires. Par exemple, il est possible d'incorporer un effet aléatoire sur l'intercept et un autre effet aléatoire associé à une covariable mais la covariance entre les deux doit alors être nulle.

Toutefois, dès que les données présentent de la variabilité à un autre niveau que le niveau individuel, le modèle de taux en excès pénalisé à effets mixtes présente un grand intérêt épidémiologique.

9.4 Illustration de la pénalisation sur l'échelle du rapport de taux en excès

Comme évoqué dans la section 8.1.2, la pénalisation des différences entre modalités d'une variable catégorielle permet de lisser les rapports de taux entre modalités. L'intérêt de cette approche est illustré dans ce qui suit.
On s'intéresse à la mortalité en excès chez 12 331 individus (9 769 hommes et 2 562 femmes) diagnostiqués en France entre 2005 et 2010 d'un cancer lèvre-bouche-pharynx (données FRANCIM). La faible proportion de femmes implique une plus grande variabilité des estimations chez les femmes que chez les hommes.

On compare les estimations des deux modèles suivants :

$$\text{modèle 1 :} \quad log[h_E(t, age, sexe)] = tensor(t, age)\mathbb{1}_{Hommes} + tensor(t, age)\mathbb{1}_{Femmes}$$

$$\text{modèle 2 :} \quad log[h_E(t, age, sexe)] = tensor(t, age) + tensor(t, age)\mathbb{1}_{Femmes}$$

Le modèle 1 correspond à des pénalisations indépendantes chez les hommes et les femmes et coïncide donc avec ce que l'on obtiendrait en réalisant une analyse séparée chez les hommes et les femmes. Le modèle 2 pénalise la différence d'effets entre les femmes et les hommes (les hommes étant pris en référence). La figure 9.1 compare les dynamiques de taux de mortalité en excès (hommes/femmes) prédites par les modèles 1 et 2 à des taux constants par intervalles. Les taux constants par intervalles sont calculés pour les classes d'âges $]35; 45]$, $]45; 55]$, $]55; 65]$ et $]65; 75]$ tandis que les taux prédits par les modèles 1 et 2 sont les taux prédits à l'âge médian au sein de chaque classe.

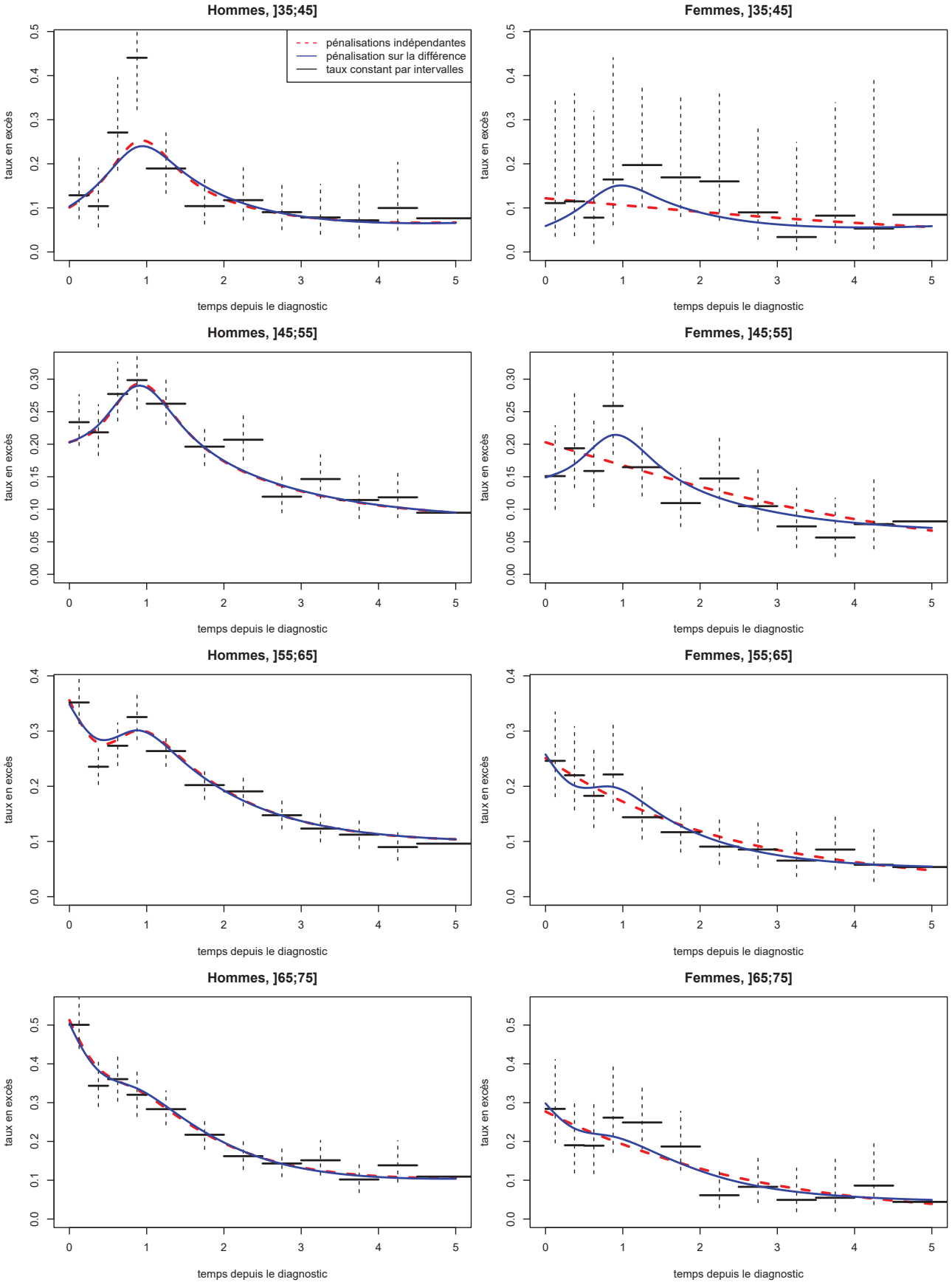

Figure 9.1 – Comparaison des taux en excès Hommes/Femmes chez les patients atteints de cancer lèvre-bouche-pharynx

Notons qu'ici, par simplicité, nous n'avons pas calculé les taux populationnels mais les taux à un âge donné (sans conséquence sur notre démonstration).
On remarque que le type de pénalisation influence la dynamique du taux chez les femmes. En effet, une pénalisation indépendante de celle des hommes engendre une prédiction plus lisse chez les femmes de moins de 55 ans (notamment les deux premières années). Au contraire, une pénalisation sur la différence entre hommes et femmes implique une dynamique du taux chez les femmes calquée sur celle des hommes (ce qui semble conforme aux taux constants par intervalles).
Clairement, le sur-lissage observé chez les femmes avec la pénalisation indépendante vient d'un manque d'information comme le montrent les intervalles de confiance des taux constants par intervalles (par exemple, sur la classe d'âges $]35;45]$ on a 437 hommes pour 109 femmes). Ainsi, les taux constants indiquent des ressemblances entre les deux sexes et seule la pénalisation sur la différence permet de retrouver cette similarité.

Une autre manière de comprendre le phénomène est de s'intéresser aux rapports de taux en excès (hommes/femmes) prédits. La figure 9.2 présente les rapports de taux de mortalité en excès (hommes/femmes) en fonction du temps selon l'âge au diagnostic et le modèle.

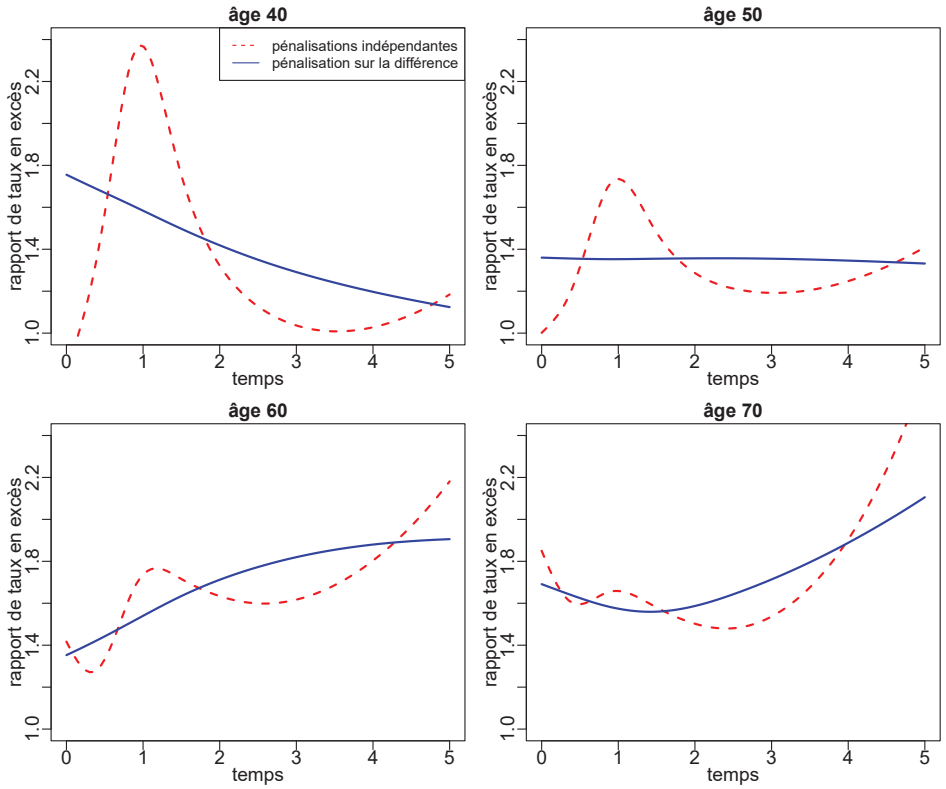

Figure 9.2 – Rapport de taux en excès Hommes/Femmes chez les patients atteints de cancer lèvre-bouche-pharynx

Le modèle 2 propose des estimations bien plus convaincantes que le modèle 1. Le fait d'utiliser les hommes comme référence permet de stabiliser énormément les estimations chez les femmes et ainsi d'obtenir des rapports de taux en excès beaucoup plus faciles à interpréter.

La pénalisation sur la différence (ou sur le rapport de taux) est particulièrement attrayante lorsqu'il faut étudier des facteurs pronostiques catégoriels. Pour le stade au diagnostic par exemple, il est raisonnable de penser que les dynamiques du taux pour deux stades successifs présentent des similarités.

Chapitre 10

Étude par simulation des propriétés statistiques de la méthode proposée

Les tendances de la survie nette en fonction de l'année de diagnostic sont très importantes en épidémiologie des cancers. L'analyse de ces tendances tient compte de l'âge au diagnostic car celui-ci demeure un facteur pronostique incontournable de la survie nette. Afin d'évaluer les performances des splines multidimensionnelles pénalisées dans ce contexte, nous avons simulé des données de survie dans lesquelles le taux de mortalité en excès est une fonction du temps écoulé depuis le diagnostic, de l'âge au diagnostic et de l'année de diagnostic.

10.1 Design

Les données simulées décrites dans cette thèse sont celles utilisées par Remontet et al. (2019). L'étude de simulation considère deux scénarios : un provenant du cancer de l'œsophage (pas d'effet de l'année) et un provenant du cancer du col de l'utérus (tendance complexe).
Premièrement, deux modèles paramétriques (non pénalisés) prenant en compte les effets du temps, de l'âge et de l'année ont été ajustés sur des données réelles françaises (Cowppli-Bony et al., 2017) et les paramètres de régression de ces modèles ont été utilisés pour générer des données simulées. Afin d'éviter tout risque de surestimation des performances du tensor, nous avons utilisé des polynômes fractionnaires (Royston and Sauerbrei, 2008) au lieu de splines de régression. La stratégie de Sauerbrei et al. (2007) a ensuite été adaptée afin de choisir les puissances des polynômes fractionnaires.

Chaque scénario s'accompagne de trois tailles d'échantillon (N = 2 000, 5 000, et 10 000 patients). les scénarios avec 2 000 et 5 000 patients sont associés à 1 000 jeux de données (D=1000), tandis que le scénario avec 10 000 patients est associé à 200 jeux de données (D=200) afin de limiter le temps de calcul. Dans chaque scénario, la distribution d'âge utilisée est équivalente à celle de patients atteints du cancer considéré dans les données réelles.
L'année de diagnostic a été tirée aléatoirement dans une loi uniforme comprise entre 1990 et 2010. Les patients sont censurés au bout de cinq ans de suivi ou en 2013.
Le temps jusqu'au décès (T) est le minimum entre le temps jusqu'au décès dû au cancer (T_E) et le temps jusqu'au décès dû à une autre cause (T_P). T_P est généré à partir d'une loi exponentielle par morceaux (Danieli et al., 2012). Une fois les paramètres de régression estimés, la fonction de répartition de T_E est calculée et T_E est généré par la méthode de la transformée inverse.

10.2 Description des tendances théoriques

En notant t le temps depuis le diagnostic, a l'âge au diagnostic et y l'année de diagnostic, les modèles théoriques sont :

$$log[h_E^{\text{œsophage}}(t,a,y)] = FP_0(t) + FP_1(t) \times a + \beta_7 \times a + \beta_8 \times \frac{1}{\sqrt{a}} + \beta_9 \times log(a)$$

et

$$\begin{aligned} log[h_E^{col}(t,a,y)] = FP_0(t) + FP_1(t) \times a + \beta_7 \times a + \beta_8 \times a^2 + \beta_9 \times log(a) \\ +\beta_{10} \times y + \beta_{11} \times y \times t + \beta_{12} \times a^2 \times y + \beta_{13} \times a^2 \times y \times t \\ +\beta_{14} \times a^3 \times log(a) \times y + \beta_{15} \times a^3 \times log(a) \times y \times t \end{aligned}$$

avec

$$FP_0(t) = \beta_0 + \beta_1 \times \frac{1}{1+t} + \beta_2 \times log(1+t) + \beta_3 \times t$$

et

$$FP_1(t) = \beta_4 \times \frac{1}{1+t} + \beta_5 \times log(1+t) + \beta_6 \times t$$

Dans les deux scénarios, l'effet de l'âge est non-linéaire et non-proportionnel.

Le scénario col de l'utérus inclut un effet linéaire et non-proportionnel de l'année, avec une forte interaction avec l'effet de l'âge, ainsi qu'une triple interaction entre les effets du temps, de l'âge et de l'année.

Au contraire, le scénario œsophage est associé à une absence d'effet de l'année. Cette hypothèse a pour but d'étudier le comportement du tensor lorsqu'un fort lissage est nécessaire selon une direction (en l'occurrence l'année) et qu'un très faible lissage est nécessaire selon une autre direction (le temps). En effet, l'une des caractéristiques intéressantes de ce scénario est la non-monotonie du taux en excès en fonction du temps : le taux de mortalité en excès augmente jusqu'à un an après le diagnostic puis décroît fortement (voir la figure 10.1).

10.3 Modèles ajustés

Les effets conjoints du temps écoulé depuis le diagnostic, de l'âge au diagnostic et de l'année de diagnostic sur le logarithme du taux en excès sont modélisés par un tensor à trois dimensions.
Les bases marginales du temps, de l'âge et de l'année sont des splines cubiques naturelles respectivement associées à six, cinq et quatre nœuds. Ces nœuds sont positionnés suivant les percentiles (parmi les décès) suivants ; 0, 0,20, 0,40, 0,60, 0,80, et 1 pour le temps ; 0, 0,25, 0,50, 0,75, et 1 pour l'âge, et 0, 0,33, 0,66, et 1 pour l'année.
Le nombre total de paramètres de régression s'élève à $6 \times 5 \times 4 = 120$ tandis que la pénalisation est contrôlée par trois paramètres de lissage. Deux modèles sont ajustés : un tensor dont les paramètres de lissage sont estimés par LAML et un tensor dont les paramètres de lissage sont estimés par LCV.

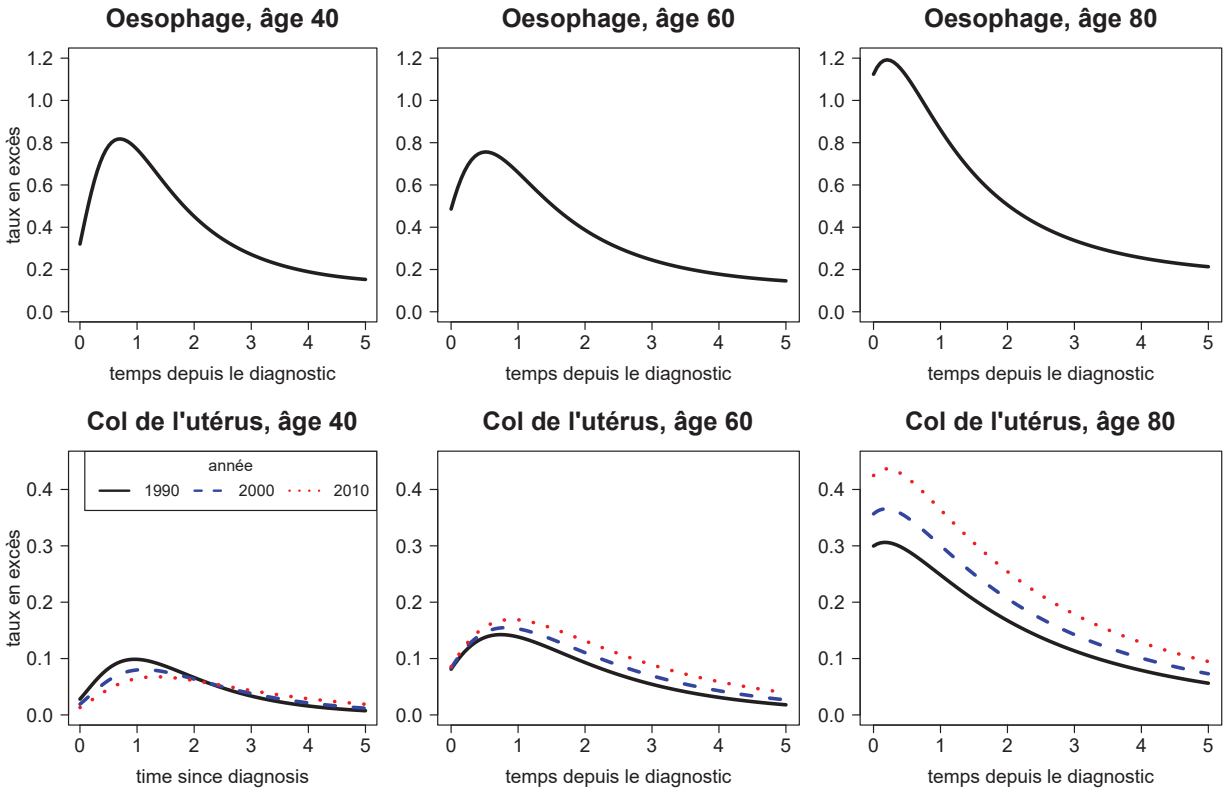

Figure 10.1 – Simulation : tendances théoriques

10.4 Évaluation de la performance de l'approche

Les performances statistiques de l'approche proposée ont été évaluées par l'estimation du taux en excès et de la survie nette à différentes combinaisons de temps écoulé depuis le diagnostic (0,5, 1, 2, 3, 4, et 5 ans), d'âges au diagnostic (tous les 5 ans, de 30 à 85 ans pour le col de l'utérus et de 45 à 85 ans pour l'œsophage), et d'années de diagnostic (tous les 5 ans de 1990 à 2010).

La simulation prend en compte 12 configurations (2 scénarios x 3 taille d'échantillon x 2 critères pour l'estimation des paramètres de lissage).

Pour un temps donné t et un vecteur de covariables $\boldsymbol{x}$, soient

$$h_E(t, \boldsymbol{x}) \quad \text{et} \quad SN(t, \boldsymbol{x})$$

les valeurs théoriques du taux en excès et de la survie nette respectivement. Soient

$$\hat{h}_E^d(t, \boldsymbol{x}) \quad \text{et} \quad \widehat{SN}^d(t, \boldsymbol{x})$$

leurs valeurs estimées respectives à partir du d^e jeu de données simulées.

Dans chacune des 12 configurations possibles, nous estimons : i) la *Root Mean Integrated Squared Error* (RMISE, voir par exemple Rondeau et al. 2007) du taux en excès,

$$\sqrt{1/D \sum_{d=1}^{D} \left[\int \left(\hat{h}_E^d(t, \boldsymbol{x}) - h_E(t, \boldsymbol{x}) \right)^2 \mathrm{d}t \right]}$$

ii) le biais dans l'estimation de la survie nette,

$$1/D \sum_{d=1}^{D} \left[\widehat{SN}^d(t, \boldsymbol{x}) - SN(t, \boldsymbol{x}) \right]$$

iii) la *Root Mean Squared Error* (RMSE) dans l'estimation de la survie nette,

$$\sqrt{1/D \sum_{d=1}^{D} \left[\widehat{SN}^d(t, \boldsymbol{x}) - SN(t, \boldsymbol{x}) \right]^2}$$

iv) les probabilités de couverture définies comme la proportion des intervalles de confiance à 95% qui contiennent la valeur théorique.
La RMISE est approchée par une quadrature de Gauss-Legendre avec 20 valeurs de temps. Les probabilités de couverture sont estimées à partir des matrices de covariance fréquentiste $\boldsymbol{V}_{\hat{\beta}}$, Bayésienne $\boldsymbol{V}_{\beta}$, et Bayésienne corrigée $\boldsymbol{V}'_{\beta}$.

Les degrés de liberté effectifs (6.1) sont utilisés afin de mesurer la complexité des modèles ajustés. Afin de mettre en perspective les performances du tensor et de valider le processus de génération des données simulées, le vrai modèle (i.e., celui qui a servi à générer les données) a également été ajusté.

10.5 Résultats

La figure 10.2 montre les performances de l'approche en termes de RMISE sur le taux en excès et la figure 10.3 montre les performances en termes de biais et de RMSE sur la survie nette. Pour chaque scénario et taille d'échantillon, les performances du vrai modèle, du tensor LCV et du tensor LAML sont comparées. Les boxplots de la RMISE sont calculés à partir de combinaisons d'âge (par tranches de 5 ans, de 45 à 85 ans pour œsophage et de 30 à 85 ans pour le col de l'utérus) et d'année de diagnostic (par tranches de 5 ans, de 1990 à 2010). Pour les boxplots de la RMSE, ces combinaisons âge-année sont répétées pour 0,5, 1, 2, 3, 4 et 5 ans de suivi.

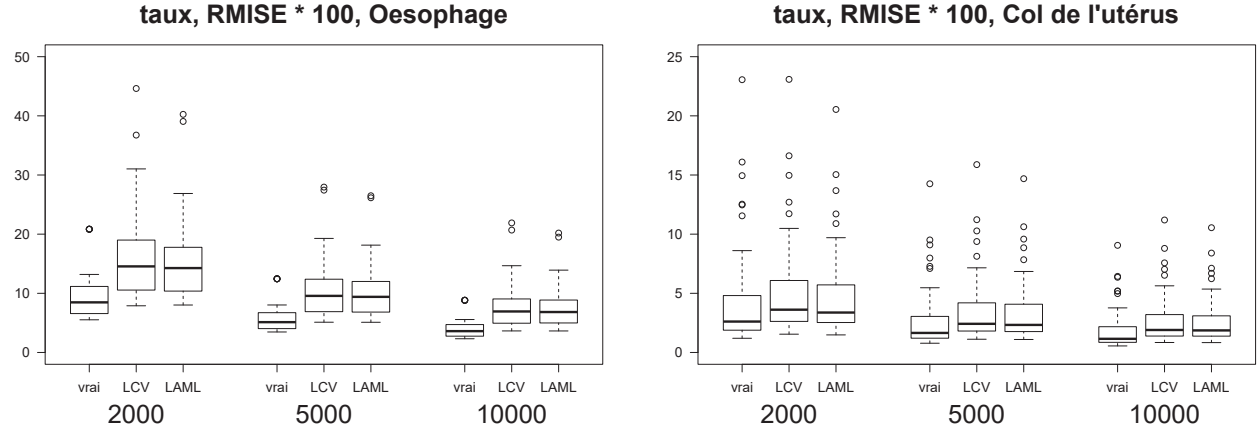

Figure 10.2 – Boxplots de la RMISE (multipliés par 100) sur le taux en excès pour chaque scénario et taille d'échantillon (45 combinaisons pour œsophage et 60 pour col).

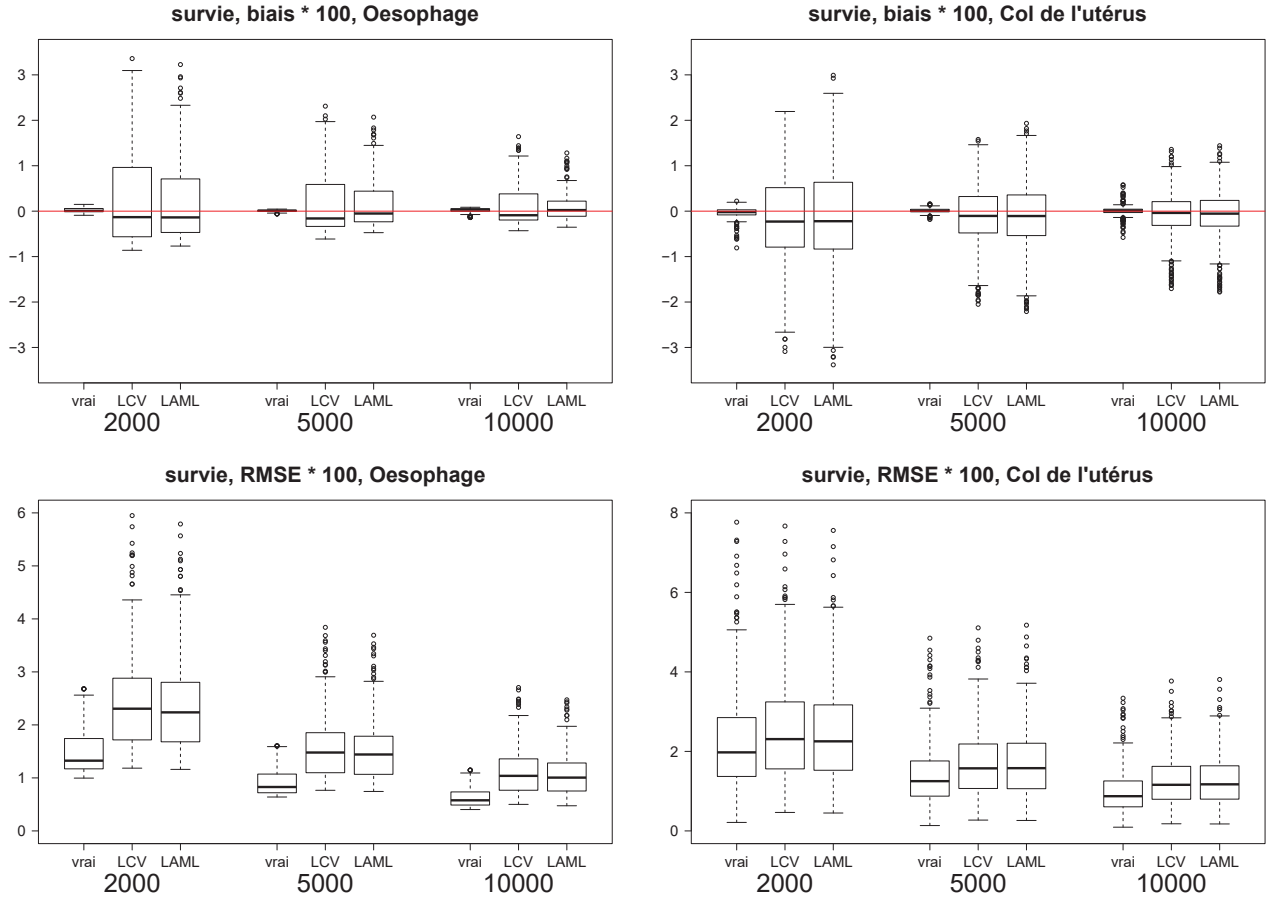

Figure 10.3 – Boxplots du biais et de la RMSE (multipliés par 100) sur la survie nette pour chaque scénario et taille d'échantillon (270 combinaisons pour œsophage et 360 pour col).

Dans le scénario œsophage, comme attendu, la RMISE du tensor est plus importante que celle du vrai modèle. Le tensor est en effet trop flexible pour un scénario aussi simple, notamment car il tient compte de l'effet de l'année alors qu'il est absent du modèle ayant généré les données : cela implique une plus grande variabilité.

Nota Bene : Toutefois, en comparaison avec *rstpm2* (voir la figure 12.1 de la section 12.2), les performances en termes de RMISE de notre approche sont bien supérieures. De même, concernant la survie (voir la figure 12.2 de la section 12.2), notre approche est supérieure en termes de RMSE par rapport à *rstpm2* bien que le biais engendré par ce dernier soit légèrement plus faible. La RMSE est un indicateur à privilégier, notamment dans un cadre pénalisé où le biais est inévitable, car il tient compte du biais et de la variance.

Quels que soient le scénario ou la taille d'échantillon, le critère LAML semble offrir des résultats légèrement meilleurs que ceux du critère LCV en termes de valeurs médianes de RMISE et RMSE (e.g., RMSE médian*100, scénario œsophage et N=2000 : 2,31 avec LCV vs. 2,24 avec LAML).

De plus, LCV semble être légèrement plus volatile que LAML : les boxplots associés à LCV sont en général plus dispersés que ceux associés à LAML. Ces résultats sont en accord avec ceux de Ruppert et al. (2003) qui montrent que le critère REML est plus susceptible d'avoir une variance inférieure

à celle du critère GCV. Les tableaux 10.1 et 10.2 donnent les probabilités de couverture médianes obtenues à partir de chaque matrice de covariance dans l'estimation du taux en excès et de la survie nette.

			$\boldsymbol{V}_{\hat{\beta}}$		$\boldsymbol{V}_{\beta}$		$\boldsymbol{V}'_{\beta}$
Scénario	Taille d'échantillon	Vrai modèle	LCV	LAML	LCV	LAML	LAML
Oesophage	2000	94,5	92,3	91,9	95,4	95,7	97,5
	5000	94,8	93,3	93,5	96,0	96,2	97,4
	10000	96,0	94,0	94,5	96,5	96,5	97,5
Col de l'utérus	2000	95,1	92,0	92,4	96,1	96,3	98,2
	5000	94,6	92,0	92,0	96,0	96,1	97,8
	10000	95,0	92,5	92,5	96,0	96,5	97,5

Tableau 10.1 – Médianes des probabilités de couverture (en %) dans l'estimation du taux en excès en fonction du type d'estimation de variance

			$\boldsymbol{V}_{\hat{\beta}}$		$\boldsymbol{V}_{\beta}$		$\boldsymbol{V}'_{\beta}$
Scénario	Taille d'échantillon	Vrai modèle	LCV	LAML	LCV	LAML	LAML
Oesophage	2000	94,8	91,5	92,0	94,1	94,8	96,9
	5000	95,2	92,2	93,1	94,9	95,6	96,9
	10000	96,0	93,5	94,5	96,0	96,5	97,5
Col de l'utérus	2000	95,1	90,8	90,6	93,4	93,3	96,3
	5000	95,0	91,1	90,4	93,9	93,5	96,1
	10000	95,0	92,0	91,5	94,5	94,0	96,0

Tableau 10.2 – Médianes des probabilités de couverture (en %) dans l'estimation de la survie nette en fonction du type d'estimation de variance

Les probabilités de couverture obtenues par la matrice de covariance fréquentiste $\boldsymbol{V}_{\hat{\beta}}$ (avec LCV ou LAML) sont assez basses, notamment dans le scénario col de l'utérus. Cette piètre performance était attendue car la matrice fréquentiste ne tient pas compte du biais induit par la pénalisation (Wood, 2006a).
Au contraire, les probabilités de couverture obtenues par la matrice de covariance Bayésienne $\boldsymbol{V}_{\beta}$ sont proches de 95% et ce même dans le scénario œsophage dans lequel le paramètre de lissage associé à l'année de diagnostic est infini. Ainsi, la matrice de covariance Bayésienne propose des performances satisfaisantes sans prendre en compte l'incertitude sur les paramètres de lissage.
Par construction, la matrice de covariance Bayésienne corrigée $\boldsymbol{V}'_{\beta}$ produit des estimations de variance plus importantes par rapport à $\boldsymbol{V}_{\beta}$. Dans l'estimation du taux en excès, les probabilités de couverture dépassent assez largement la valeur cible de 95%.

Le tableau 10.3 donne les valeurs médianes des degrés de liberté effectifs obtenus par le tensor LCV et LAML pour chaque scénario et taille d'échantillon. Le critère LAML semble ainsi produire des estimations un peu plus lisses que le critère LCV.

Scénario	Taille d'échantillon	edf avec LCV	edf avec LAML
Oesophage	2000	22,8	21,5
	5000	27,5	26,7
	10000	31,5	31,0
Col de l'utérus	2000	19,5	17,8
	5000	26,2	24,8
	10000	32,6	30,3

Tableau 10.3 – Médianes des degrés de liberté effectifs parmi tous les modèles ajustés

Parmi les 8 800 modèles ajustés (4 400 avec LAML et 4 400 avec LCV), seul un tensor LCV n'a pas atteint les critères de convergence. Les propriétés de convergence du tensor sont donc extrêmement satisfaisantes.
Les temps d'exécution médians avec LCV et LAML sont comparables pour N=2 000 et 5 000 (autour de 60s et 160s par modèle, respectivement, sur un ordinateur de bureau utilisant un processeur Intel i5-6600 cadencé à 3,30 GHz et doté de 16 Go de mémoire vive). Pour N=10 000, l'optimisation via LAML est légèrement plus rapide que celle via LCV (430s vs. 460s). En outre, la variance des temps d'exécution est plus importante avec LCV.

En règle générale, le critère LCV donne des probabilités de couverture un peu plus faibles que LAML. Toutefois, les performances des deux critères à partir de la matrice de covariance Bayésienne deviennent comparables avec l'augmentation de la taille d'échantillon. L'optimisation via LAML semble légèrement plus performante que celle fondée sur le LCV en termes de RMISE, RMSE, et probabilités de couverture. Cependant, les différences sont assez faibles et les deux critères demeurent utiles en pratique.

Enfin, notons que les simulations exposées ici ne considèrent pas de très petits échantillons (entre 200 et 1 000 individus). Toutefois, Remontet et al. (2019) ont montré que les performances du tensor pénalisé ne se dégradaient pas dans de telles configurations.

Chapitre 11

Le package *survPen*

11.1 Motivation

Lorsque l'on souhaite promouvoir une nouvelle méthode d'analyse, il est essentiel de la rendre utilisable par d'autres. C'est la raison pour laquelle j'ai choisi d'écrire le package *survPen*. En outre, l'écriture de ce package est indissociable de mon travail de thèse car il représente la synthèse de mon appropriation du sujet. Le package a fait l'objet d'un article scientifique (Fauvernier et al., 2019a). La version publiée est disponible à la fin de cette thèse (voir section D).

Le package *survPen* est disponible sur le site du CRAN à l'adresse suivante :
https ://CRAN.R-project.org/package=survPen

La version de développement est disponible à l'adresse suivante :
https ://github.com/fauvernierma/survPen

Notons enfin que ce chapitre est une traduction française de la vignette du package disponible via les liens ci-dessus.

11.2 datCancer

Le package *survPen* dispose d'un jeu de données simulées (*datCancer*) portant sur 2000 patientes diagnostiquées d'un cancer du col de l'utérus entre 1990 et 2010. La fin du suivi est fixée au 30 Juin 2013. Le jeu de données comprend 6 variables :

- begin : début du suivi, de 0 à 1 an. Pour illustrations sur la troncature gauche uniquement.
- fu : temps de suivi en années, de 0 à 5 ans.
- age : âge au diagnostic en années, de 21,39 à 99,33 ans.
- yod : année de diagnostic, de 1990,023 à 2010,999.
- dead : indicatrice de décès (1 si décès, 0 sinon).
- rate : taux de mortalité attendu, de 0 à 0,38 décès par personne-année.

Les dix premières lignes sont données ci-dessous :

```
data(datCancer)
```

Tableau 11.1 – *survPen* - jeu de données datCancer

begin	fu	age	yod	dead	rate
0.26	0.74	35.86	1990.62	1	0.00
0.20	0.77	43.52	1990.19	1	0.00
0.74	0.88	46.04	1990.16	1	0.00
0.55	0.76	49.97	1990.06	1	0.00
0.27	0.88	49.18	1990.31	1	0.00
0.45	0.77	52.53	1990.22	1	0.00
0.57	0.94	53.26	1990.74	1	0.00
0.04	0.05	55.25	1990.12	1	0.00
0.00	0.03	66.30	1990.30	1	0.01
0.15	0.18	73.87	1990.31	1	0.02

11.3 Débuter avec *survPen*

La spécification du modèle devrait paraître naturelle aux utilisateurs de la fonction *glm* car le prédicteur linéaire est intégralement spécifié par la formule du modèle. En plus du temps de suivi (argument *t1*) et de l'indicatrice d'événement (argument *event*), l'utilisateur ne doit fournir qu'une formule commençant par le symbole "∼" suivi des formes fonctionnelles des covariables et du temps. Rien n'est spécifié à la gauche de la formule car le prédicteur est implicite (log du taux ou log du taux en excès).

Supposons que l'on ne soit intéressé que par l'effet du temps sur le taux de mortalité.
Voici quelques exemples de modèles ajustés sur le log du taux.

11.3.1 Taux constant

$$log[h(t)] = \beta_0$$

```
f.cst <- ~1
mod.cst <- survPen(f.cst, data = datCancer, t1 = fu, event = dead)
```

11.3.2 Taux constant par intervalles

$$log[h(t)] = \sum_{k=1}^{p} \beta_k I_k(t)$$

où $I_k(t) = 1$ si t appartient au k^e intervalle et 0 sinon.

```
f.pwcst <- ~cut(fu, breaks = seq(0, 5, by = 0.5), include.lowest = TRUE)
mod.pwcst <- survPen(f.pwcst, data = datCancer, t1 = fu, event = dead, n.legendre = 200)
```

Ici, on augmente le nombre de nœuds pour la quadrature de Gauss-Legendre afin que le taux cumulé soit bien approximé.

11.3.3 Taux log-linéaire

$$log[h(t)] = \beta_0 + \beta_1 \times t$$

```
f.lin <- ~fu
mod.lin <- survPen(f.lin, data = datCancer, t1 = fu, event = dead)
```

11.3.4 Splines cubiques restreintes

$$log[h(t)] = f(t)$$

où f est une spline cubique restreinte (linéaire au-delà des nœuds extérieurs) avec les nœuds intérieurs 0,25, 0,5, 1, 2 et 4 et les nœuds extérieurs 0 et 5. Ici, il faut noter que f n'est pas pénalisée.

On utilise le package *splines*

```
library(splines)

f.rcs <- ~ns(fu, knots = c(0.25, 0.5, 1, 2, 4), Boundary.knots = c(0, 5))

mod.rcs <- survPen(f.rcs, data = datCancer, t1 = fu, event = dead)
```

11.3.5 Splines cubiques restreintes pénalisées

Nous utilisons le même modèle que précédemment à ceci près que l'on ajoute un terme de pénalité qui contrôle la régularité de la courbe prédite

$$log[h(t)] = s(t)$$

où s est une spline cubique restreinte **pénalisée** avec les nœuds intérieurs 0,25, 0,5, 1, 2 et 4 et les nœuds extérieurs 0 et 5.

On utilise le constructeur *smf* (pour *smooth function*) au sein du package *survPen*

```
f.pen <- ~smf(fu, knots = c(0, 0.25, 0.5, 1, 2, 4, 5))
# attention, les noeuds exterieurs sont inclus ici
mod.pen <- survPen(f.pen, data = datCancer, t1 = fu, event = dead)
```

Nota Bene : la version non pénalisée de ce modèle pouvait également être spécifiée en forçant le paramètre de lissage à zéro :

```
mod.unpen <- survPen(f.pen, data = datCancer, t1 = fu, event = dead, lambda = 0)
```

mod.unpen correspond ainsi à mod.rcs.

11.4 Prédictions et sorties du modèle

11.4.1 Prédictions standards

```
new.time <- seq(0, 5, length = 100)
pred.cst <- predict(mod.cst, data.frame(fu = new.time))
pred.pwcst <- predict(mod.pwcst, data.frame(fu = new.time))
pred.lin <- predict(mod.lin, data.frame(fu = new.time))
pred.rcs <- predict(mod.rcs, data.frame(fu = new.time))
pred.pen <- predict(mod.pen, data.frame(fu = new.time))

lwd1 <- 2

par(mfrow = c(1, 1))
plot(new.time, pred.cst$haz, type = "l", ylim = c(0, 0.2), main = "Dynamique du taux",
    xlab = "temps écoulé depuis le diagnostic (années)", ylab = "taux de mortalité",
    col = "black", lwd = lwd1)
segments(x0 = new.time[1:99], x1 = new.time[2:100], y0 = pred.pwcst$haz[1:99],
    col = "blue3", lwd = lwd1)
lines(new.time, pred.lin$haz, col = "green3", lwd = lwd1)
lines(new.time, pred.rcs$haz, col = "orange", lwd = lwd1)
lines(new.time, pred.pen$haz, col = "red", lwd = lwd1)

legend("topright", legend = c("constant", "constant par morceaux", "log-linéaire",
    "spline cubique restreinte", "spline cubique restreinte pénalisée"), col = c("black",
    "blue3", "green3", "orange", "red"), lty = rep(1, 5), lwd = rep(lwd1, 5))
```

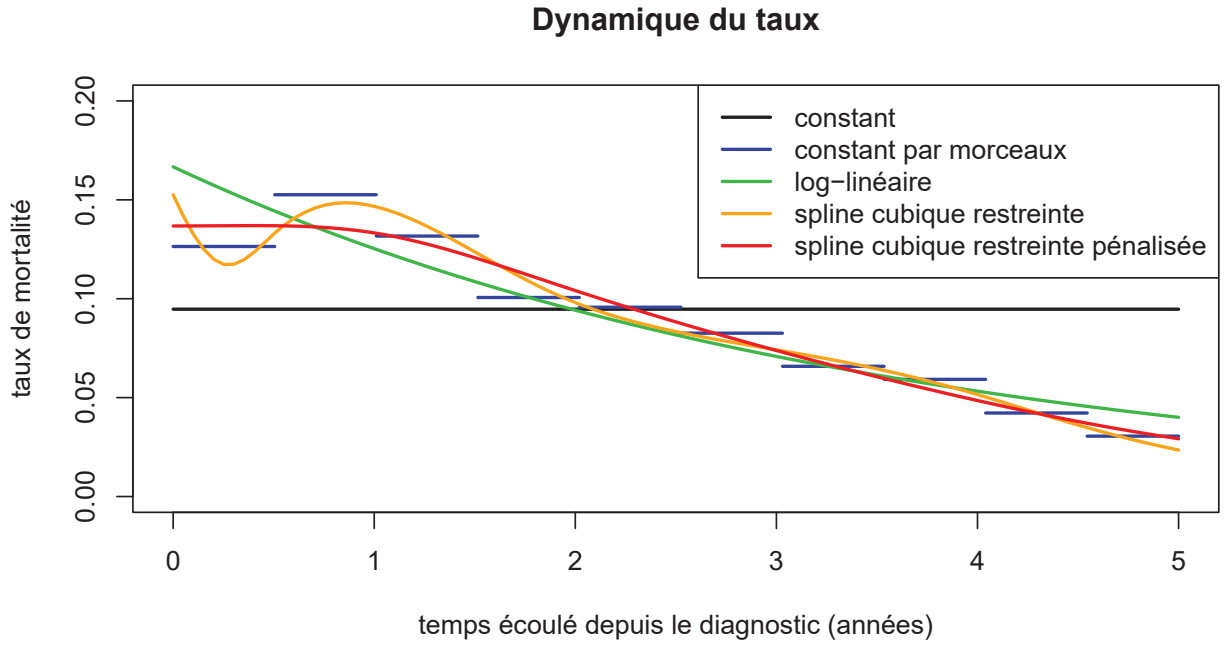

Figure 11.1 – *survPen* - Comparaison de différents modèles

La figure 11.1 montre que le modèle pénalisé donne un taux plus lisse que la version non pénalisée. De manière générale, l'estimation réalisée à partir d'un modèle pénalisé présente ainsi un biais légèrement

plus important mais est moins propice au sur-ajustement.

Les prédictions du taux et de la survie peuvent également être accompagnées de leurs intervalles de confiance (voir la figure 11.2)

```
par(mfrow = c(1, 2))
plot(new.time, pred.pen$haz, type = "l", ylim = c(0, 0.2), main = "Taux prédit par mod.pen",
    xlab = "temps", ylab = "taux", col = "red", lwd = lwd1)
lines(new.time, pred.pen$haz.inf, lty = 2)
lines(new.time, pred.pen$haz.sup, lty = 2)

plot(new.time, pred.pen$surv, type = "l", ylim = c(0, 1), main = "Survie prédite par mod.pen",
    xlab = "temps", ylab = "survie", col = "red", lwd = lwd1)
lines(new.time, pred.pen$surv.inf, lty = 2)
lines(new.time, pred.pen$surv.sup, lty = 2)
```

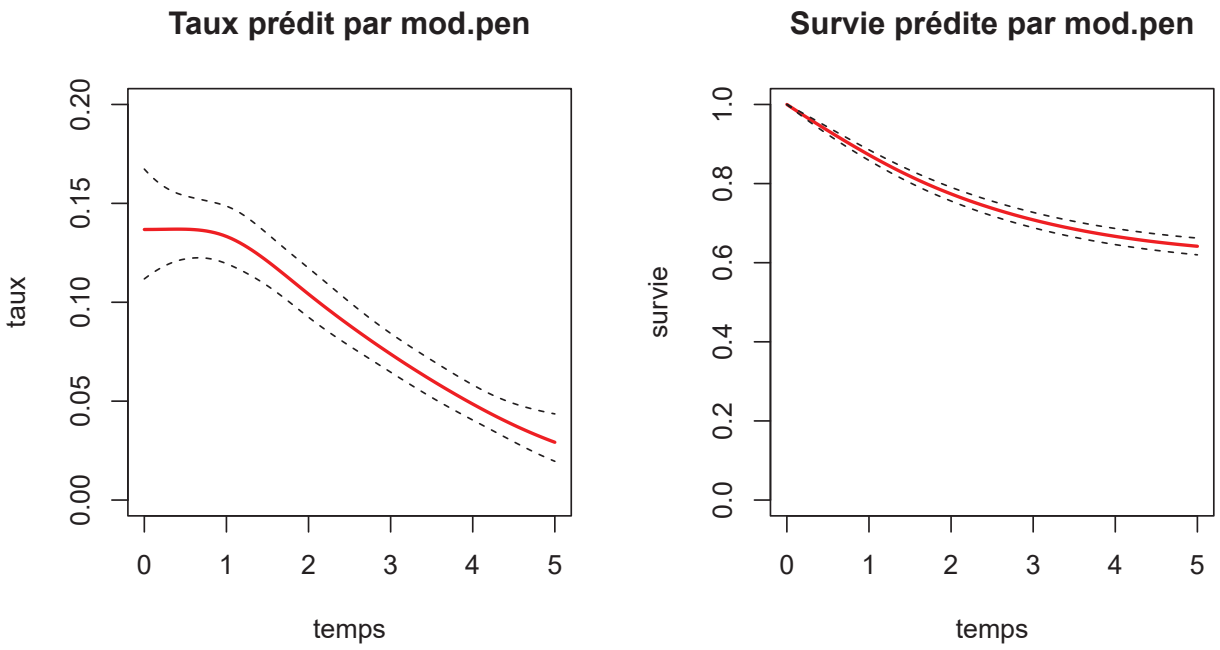

Figure 11.2 – *survPen* - Prédictions avec intervalles de confiance

11.4.2 Prédictions à la carte

En plus des traditionnelles estimations du taux et de la survie (et leurs intervalles de confiance), l'utilisateur peut récupérer directement la matrice de design associée aux nouvelles données de prédiction. Cette fonctionnalité est disponible avec la méthode *predict* via l'argument *type = "lpmatrix"*. Cette fonctionnalité est particulièrement intéressante si l'utilisateur veut réaliser des prédictions sur une échelle arbitraire (en dehors du taux, taux cumulé et de la survie).

Recalculons par exemple le taux à la main et comparons avec la prédiction par défaut

```
haz.pen <- pred.pen$haz
X.pen <- predict(mod.pen, data.frame(fu = new.time), type = "lpmatrix")
haz.pen.lpmatrix <- as.numeric(exp(X.pen %*% mod.pen$coefficients))
```

```
summary(haz.pen.lpmatrix - haz.pen)

#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>       0       0       0       0       0       0
```

Les intervalles de confiance à 95% peuvent être calculés ainsi

```
# erreurs standards a partir de la matrice de covariance bayesienne Vp
std <- sqrt(rowSums((X.pen %*% mod.pen$Vp) * X.pen))

qt.norm <- stats::qnorm(1 - (1 - 0.95)/2)
haz.inf <- as.vector(exp(X.pen %*% mod.pen$coefficients - qt.norm * std))
haz.sup <- as.vector(exp(X.pen %*% mod.pen$coefficients + qt.norm * std))

summary(haz.inf - pred.pen$haz.inf)

#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>       0       0       0       0       0       0

summary(haz.sup - pred.pen$haz.sup)

#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>       0       0       0       0       0       0
```

11.4.3 Résumé du modèle

Regardons le résumé de *mod.pen*

```
summary(mod.pen)

#> penalized hazard model
#>
#> Call:
#> survPen(formula = f.pen, data = datCancer, t1 = fu, event = dead)
#>
#> Coefficients:
#>             Estimate Std. Error z value  Pr(>|z|)
#> (Intercept) -3.04226    0.12344 -24.645 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> likelihood= -2320.85 , penalized likelihood= -2321.66
#> Number of parameters= 7 , effective degrees of freedom= 3.6549
#> LAML = 2326.26
#>
#> Smoothing parameter(s):
#>  smf(fu)
#> 1690.498
#>
#> edf of smooth terms:
#>  smf(fu)
```

```
#> 2.654904
#>
#> converged= TRUE
```

La fonction *summary* affiche :
- la log-vraisemblance
- la log-vraisemblance pénalisée
- le nombre de paramètres de régression
- le nombre de degrés de liberté effectifs
- **l'opposé** du critère LAML à convergence
- les paramètres de lissage estimés
- les degrés de liberté effectifs par splines pénalisées

Toutes ces valeurs peuvent être également récupérées ainsi

```
mod.pen$ll.unpen

#> [1] -2320.848

mod.pen$ll

#> [1] -2321.659

mod.pen$p

#> [1] 7

sum(mod.pen$edf)

#> [1] 3.654904

mod.pen$LAML

#> [1] 2326.262

mod.pen$lambda

#>  smf(fu)
#> 1690.498

summary(mod.pen)$edf.per.smooth

#>  smf(fu)
#> 2.654904
```

11.4.4 Sélection de modèle

L'AIC du modèle est donné par

```
mod.pen$aic

#> [1] 4649.005
```

Les degrés de liberté effectifs (edf) utilisés pour définir l'AIC sont

```
sum(mod.pen$edf)

#> [1] 3.654904
```

L'AIC corrigé pour tenir compte de l'incertitude sur les paramètres de lissage est

```
mod.pen$aic2

#> [1] 4650.505
```

Les degrés de liberté effectifs corrigés associés sont

```
sum(mod.pen$edf2)

#> [1] 4.405014
```

11.5 Estimation des paramètres de lissage

Le package *survPen* propose deux critères afin d'estimer les paramètres de lissage : LCV pour *Likelihood Cross Validation* et LAML pour *Laplace Approximate Marginal Likelihood* (voir le chapitre 6).

```
f1 <- ~smf(fu)

mod.LCV <- survPen(f1, data = datCancer, t1 = fu, event = dead, expected = NULL,
    method = "LCV")
mod.LCV$lambda

#>  smf(fu)
#> 3346.303



mod.LAML <- survPen(f1, data = datCancer, t1 = fu, event = dead, expected = NULL,
    method = "LAML")
mod.LAML$lambda

#>  smf(fu)
#> 3682.498
```

En général, choisir l'un ou l'autre n'impactera quasiment pas les prédictions dans le sens où les paramètres de lissage estimés seront les mêmes (voir la figure 11.3).

```
new.time <- seq(0, 5, length = 100)
pred.LCV <- predict(mod.LCV, data.frame(fu = new.time))
pred.LAML <- predict(mod.LAML, data.frame(fu = new.time))

par(mfrow = c(1, 1))
plot(new.time, pred.LCV$haz, type = "l", ylim = c(0, 0.2), main = "LCV vs LAML",
    xlab = "temps écoulé depuis le diagnostic (années)", ylab = "taux", col = "black",
    lwd = lwd1)
lines(new.time, pred.LAML$haz, col = "red", lwd = lwd1, lty = 2)
legend("topright", legend = c("LCV", "LAML"), col = c("black", "red"), lty = c(1,
    2), lwd = rep(lwd1, 2))
```

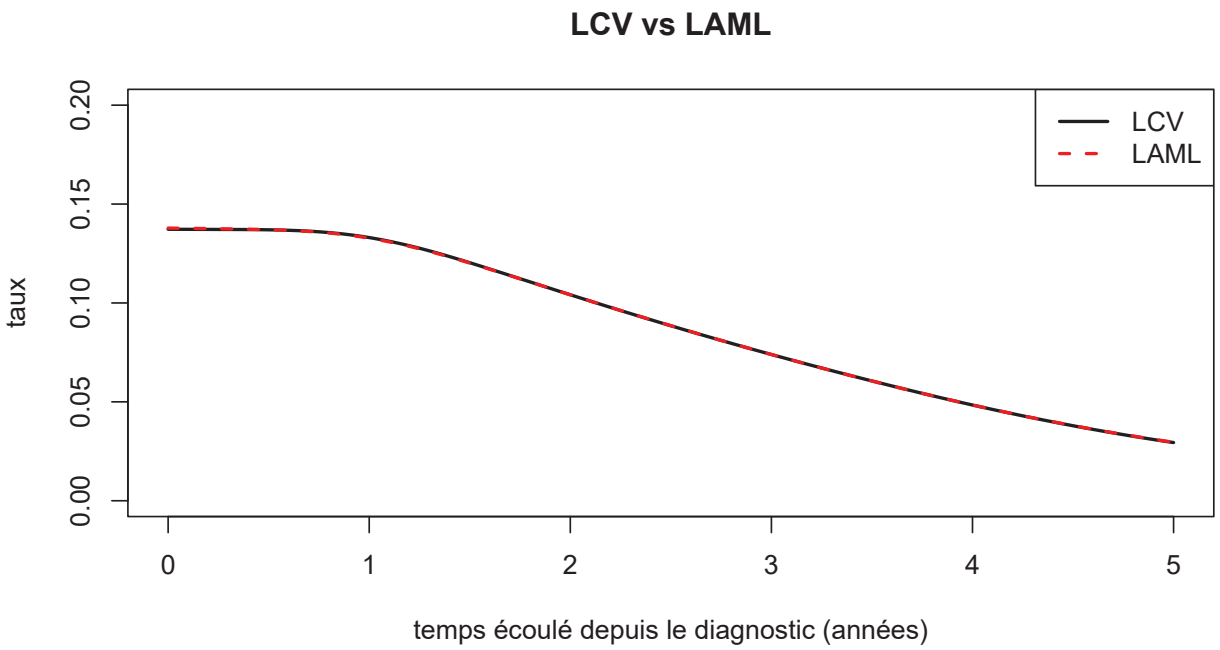

Figure 11.3 – *survPen* - LCV vs LAML avec 10 paramètres de régression

Pour mieux comprendre le processus d'estimation, affichons les critères LCV et LAML comme fonctions des (log-) paramètres de lissage (figure 11.4).

```
rho.vec <- seq(-1, 15, length = 50)
LCV <- rep(0, 50)
LAML <- rep(0, 50)

for (i in 1:50) {
    mod <- survPen(f1, data = datCancer, t1 = fu, event = dead, lambda = exp(rho.vec[i]))
    LCV[i] <- mod$LCV
    LAML[i] <- mod$LAML
}

rho.LCV <- rho.vec[which(LCV == min(LCV))]
rho.LAML <- rho.vec[which(LAML == min(LAML))]
```

```
par(mfrow = c(1, 2), mar = c(3, 3, 1.5, 0.5), mgp = c(1.5, 0.5, 0))

plot(rho.vec, LCV, type = "l", main = expression(paste("LCV vs ", log(lambda))),
    ylab = "LCV", xlab = expression(log(lambda)), lwd = lwd1)
abline(v = rho.LCV, lty = 2, col = "red")
text(rho.LCV, mean(LCV), bquote(log(lambda) == .(round(rho.LCV, 2))))

plot(rho.vec, LAML, type = "l", main = expression(paste("LAML vs ", log(lambda))),
    ylab = "-LAML", xlab = expression(log(lambda)), lwd = lwd1)
abline(v = rho.LAML, lty = 2, col = "red")
text(rho.LAML, mean(LAML), bquote(log(lambda) == .(round(rho.LAML, 2))))
```

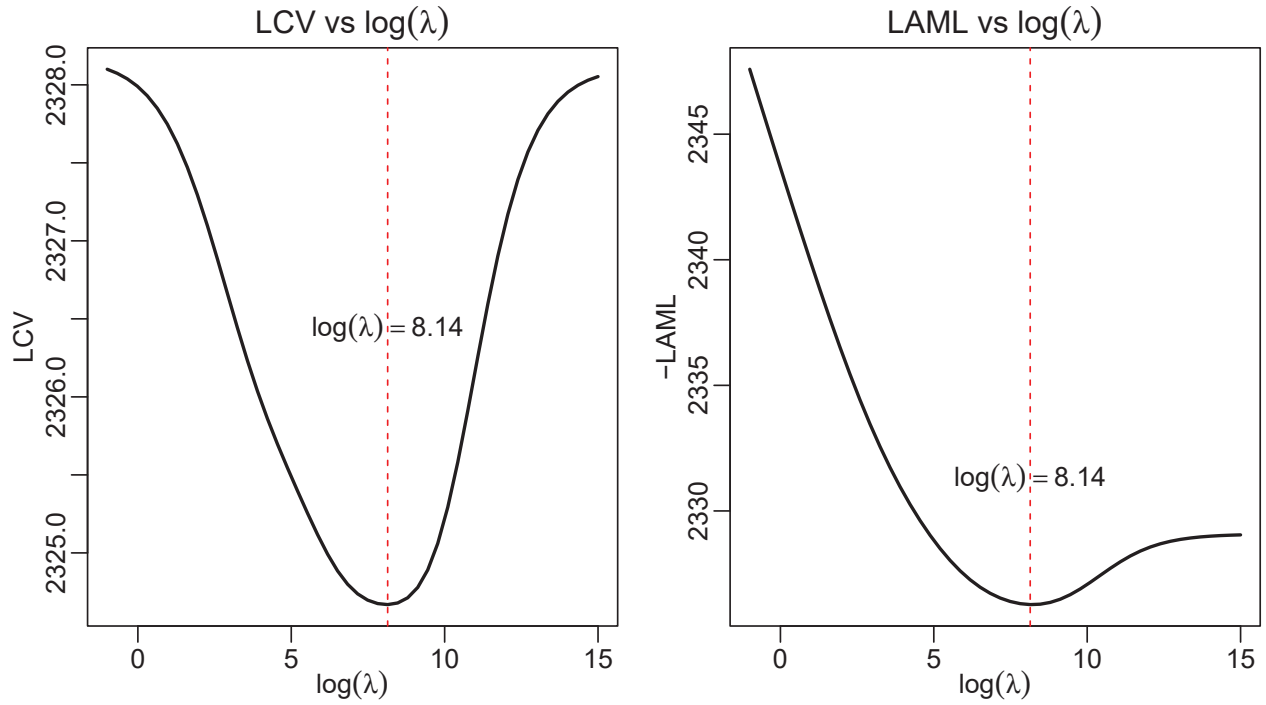

Figure 11.4 – *survPen* - LCV et LAML comme fonctions des paramètres de lissage

D'après la figure 11.4, les deux fonctions conduisent au même paramètre de lissage.

11.6 Position des noeuds

Exemple d'une spline pénalisée du temps avec 5 nœuds

```
f1 <- ~smf(fu, df = 5)
```

Quand les nœuds ne sont pas spécifiés, ils sont calculés à partir des quantiles de la covariable : par exemple, pour le terme *smf(x,df=df1)*, les nœuds seront : *quantile(unique(x),seq(0,1,length=df1))*.

```
df1 <- 5
quantile(unique(datCancer$fu), seq(0, 1, length = df1))


#>          0%         25%         50%         75%        100%
#> 0.001455958 0.732075578 1.425060350 2.570586727 5.000000000
```

Les noeuds sont également accessibles via l'objet *survPen*

```
mod1 <- survPen(f1, data = datCancer, t1 = fu, event = dead)

mod1$list.smf

#> [[1]]
#> $term
#> [1] "fu"
#>
#> $dim
#> [1] 1
#>
#> $knots
#> $knots[[1]]
#>          0%         25%         50%         75%        100%
#> 0.001455958 0.732075578 1.425060350 2.570586727 5.000000000
#>
#>
#> $df
#> [1] 5
#>
#> $by
#> [1] "NULL"
#>
#> $same.rho
#> [1] FALSE
#>
#> $name
#> [1] "smf(fu)"
#>
#> attr(,"class")
#> [1] "smf.smooth.spec"
```

et les noeuds peuvent être spécifiés par l'utilisateur

```
# f1 <- ~smf(fu,knots=c(0,1,3,6,8))
```

11.7 Taux en excès

L'une des caractéristiques très importantes du package *survPen* est la possibilité d'ajuster des modèles de taux en excès pénalisés (voir le chapitre 9). Ajustons un modèle de taux et un modèle de taux en excès.

```
mod.total <- survPen(f1, data = datCancer, t1 = fu, event = dead, method = "LAML")
mod.excess <- survPen(f1, data = datCancer, t1 = fu, event = dead, expected = rate,
    method = "LAML")
```

La spécification des taux attendus passe par l'argument *expected*. Comparons les prédictions des deux modèles dans la figure 11.5 :

```r
new.time <- seq(0, 5, length = 100)
pred.total <- predict(mod.total, data.frame(fu = new.time))
pred.excess <- predict(mod.excess, data.frame(fu = new.time))

# taux vs taux en exces
par(mfrow = c(1, 2))
plot(new.time, pred.total$haz, type = "l", ylim = c(0, 0.2), main = "taux vs taux en excès",
    xlab = "temps", ylab = "taux")
lines(new.time, pred.excess$haz, col = "green3", lty = 2)
legend("topright", legend = c("taux brut", "taux en excès"), col = c("black",
    "green3"), lty = c(1, 2))

# survie vs survie nette
plot(new.time, pred.total$surv, type = "l", ylim = c(0, 1), main = "survie vs survie nette",
    xlab = "temps", ylab = "survie")
lines(new.time, pred.excess$surv, col = "green3", lty = 2)
legend("bottomleft", legend = c("survie brute", "survie nette"), col = c("black",
    "green3"), lty = c(1, 2))
```

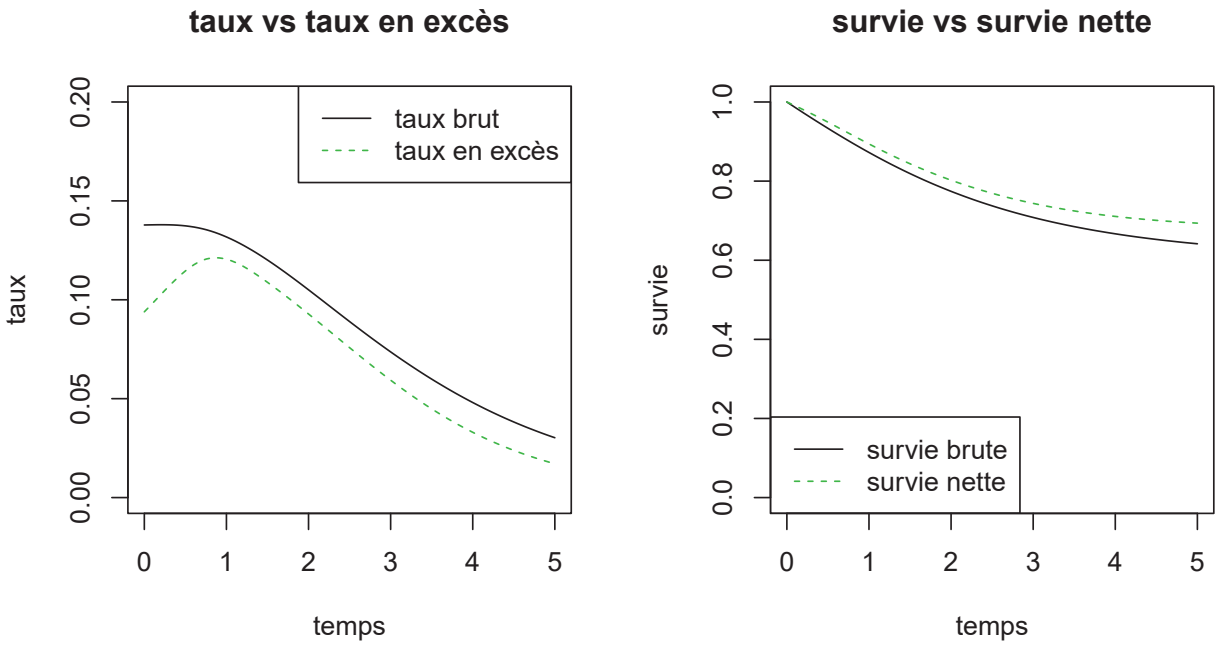

Figure 11.5 – *survPen* - Taux brut vs taux en excès

11.8 Produit tensoriel

Les produits tensoriels représentent une fonctionnalité majeure du package *survPen* (voir la section 4.4.3). En effet, ils permettent de modéliser conjointement la non-linéarité, la non-proportionnalité ainsi que les interactions.

Deux constructeurs sont disponibles

- *tensor* : le nombre de paramètres de lissage est égal au nombre de covariables. Ce constructeur est similaire à *te* du package *mgcv*.
- *tint* : implique le même design qu'avec *tensor* mais décompose les termes de pénalité entre les effets propres d'une part et les interactions de l'autre (*ANOVA decomposition*). Ce constructeur est similaire à *ti* du package *mgcv*.

Le constructeur *tensor* permet de spécifier des modèles tels que

$$log[h(t, age)] = f(t, age)$$

où f est un produit tensoriel associé à deux paramètres de lissage, un pour chaque direction. Cependant, cette construction fait l'hypothèse que l'effet propre de chaque covariable est de même complexité que son effet associé dans le terme d'interaction.

Le constructeur *tint* lève cette hypothèse. En effet, le modèle devient :

$$log[h(t, age)] = f(t) + g(age) + k(t, age)$$

où f est associée à un paramètre de lissage, g également et k est associée à deux paramètres de lissage. Nous obtenons quatre paramètres de lissage au total mais le design est le même que précédemment.

Bien sûr, l'approche *tint* atteint rapidement ses limites en termes de complexité quand le nombre de covariables augmente. En effet, avec trois covariables par exemple, alors que le *tensor* est associé à trois paramètres de lissage, la version intégralement décomposée via *tint* atteint les douze paramètres de lissage à estimer.

11.8.1 Deux dimensions

Nous comparons les approches *tensor* et *tint*. On s'intéresse au temps de suivi et à l'âge au diagnostic. Les paramètres de lissage sont estimés par LAML.

```
f.tensor <- ~tensor(fu, age, df = c(5, 5))
f.tint <- ~tint(fu, df = 5) + tint(age, df = 5) + tint(fu, age, df = c(5, 5))

mod.tensor <- survPen(f.tensor, data = datCancer, t1 = fu, event = dead)
summary(mod.tensor)


#> penalized hazard model
#>
#> Call:
#> survPen(formula = f.tensor, data = datCancer, t1 = fu, event = dead)
#>
#> Coefficients:
#>             Estimate Std. Error z value  Pr(>|z|)
```

```
#> (Intercept) -3.31334    0.17612 -18.813 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> likelihood= -2106.15 , penalized likelihood= -2110
#> Number of parameters= 25 , effective degrees of freedom= 11.6904
#> LAML = 2121.7
#>
#> Smoothing parameter(s):
#> tensor(fu,age).1 tensor(fu,age).2
#>        0.7792689       21.6699606
#>
#> edf of smooth terms:
#> tensor(fu,age)
#>       10.69041
#>
#> converged= TRUE


mod.tint <- survPen(f.tint, data = datCancer, t1 = fu, event = dead)
summary(mod.tint)

#> penalized hazard model
#>
#> Call:
#> survPen(formula = f.tint, data = datCancer, t1 = fu, event = dead)
#>
#> Coefficients:
#>             Estimate Std. Error z value  Pr(>|z|)
#> (Intercept) -3.23237    0.15164 -21.316 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> likelihood= -2106.39 , penalized likelihood= -2109.6
#> Number of parameters= 25 , effective degrees of freedom= 10.462
#> LAML = 2122.7
#>
#> Smoothing parameter(s):
#>       tint(fu)      tint(age) tint(fu,age).1 tint(fu,age).2
#>   9.405365e-01   6.451661e+00   1.830722e-01   1.279995e+05
#>
#> edf of smooth terms:
#>     tint(fu)    tint(age) tint(fu,age)
#>     3.566385     2.519234     3.376361
#>
#> converged= TRUE
```

Le modèle *tensor* est associé à deux paramètres de lissage et le modèle *tint* à quatre.
Dans le modèle *tint*, le paramètre de lissage associé à l'âge dans le terme d'interaction (tint(fu,age).2) est beaucoup plus important que celui associé à l'effet propre de l'âge (tint(age)). Ce comportement est impossible à reproduire à partir d'une approche *tensor*.

Malgré cette différence, les écarts de prédictions sont imperceptibles dans la figure 11.6.

```
# predictions
new.age <- seq(50, 90, length = 50)
new.time <- seq(0, 7, length = 50)
Z.tensor <- outer(new.time, new.age, function(t, a) predict(mod.tensor, data.frame(fu = t,
    age = a))$haz)
Z.tint <- outer(new.time, new.age, function(t, a) predict(mod.tint, data.frame(fu = t,
    age = a))$haz)

# couleurs
col.pal <- colorRampPalette(c("white", "red"))
colors <- col.pal(100)
facet <- function(z) {
    facet.center <- (z[-1, -1] + z[-1, -ncol(z)] + z[-nrow(z), -1] + z[-nrow(z),
        -ncol(z)])/4
    cut(facet.center, 100)

}

par(mfrow = c(1, 2), mar = c(3, 3, 1.5, 0.5), mgp = c(1.5, 0.5, 0))
persp(new.time, new.age, Z.tensor, col = colors[facet(Z.tensor)], main = "tensor",
    theta = 30, xlab = "\n temps", ylab = "\n âge", zlab = "\n taux", ticktype = "detailed",
    zlim = c(0, 1.1))
persp(new.time, new.age, Z.tint, col = colors[facet(Z.tint)], main = "tint",
    theta = 30, xlab = "\n temps", ylab = "\n âge", zlab = "\n taux", ticktype = "detailed",
    zlim = c(0, 1.1))
```

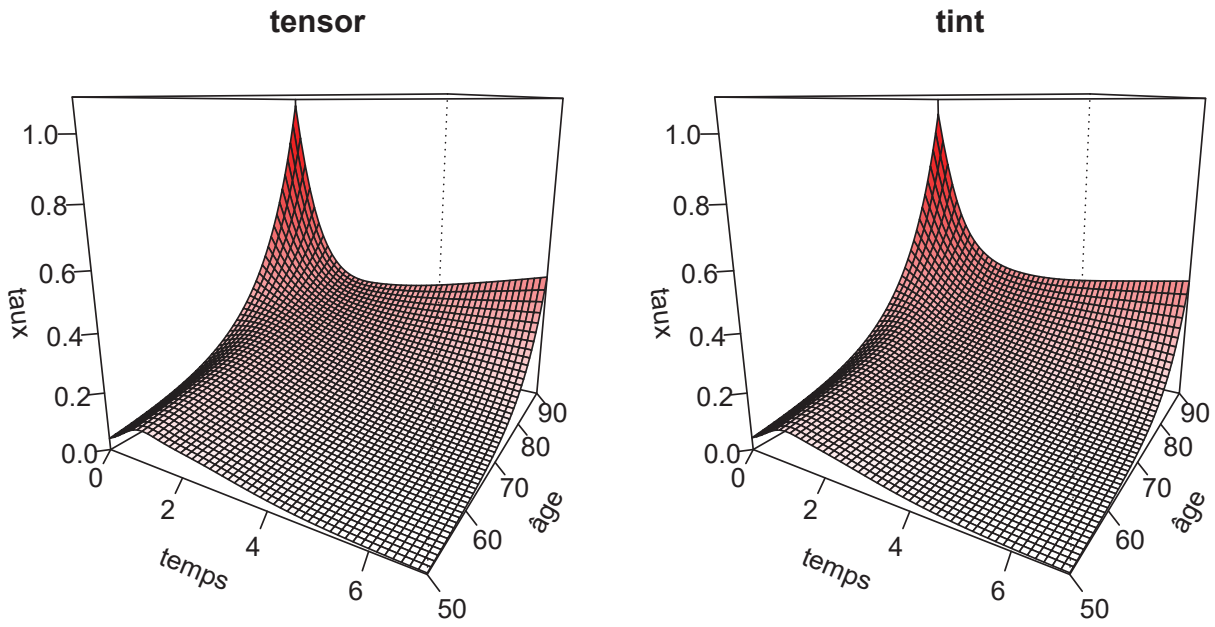

Figure 11.6 – *survPen* - Surfaces du taux, tensor vs tint

En pratique, il est conseillé d'utiliser l'approche *tint* lorsque l'on s'attend à une structure d'interaction plus simple ou plus complexe que les effets propres. Illustrons les différences entre les deux approches à partir du jeu de données suivant :

```
set.seed(18)
subdata <- datCancer[sample(1:2000, 50), ]
```

Maintenant, on ajuste les mêmes modèles que précédemment

```
mod.tensor.sub <- survPen(f.tensor, data = subdata, t1 = fu, event = dead)

mod.tint.sub <- survPen(f.tint, data = subdata, t1 = fu, event = dead)
```

Les paramètres de lissage estimés et les degrés de liberté effectifs (edf) par termes pénalisés sont donnés ci-dessous

```
# tensor
mod.tensor.sub$lambda

#> tensor(fu,age).1 tensor(fu,age).2
#>         6195.994        55611.684

summary(mod.tensor.sub)$edf.per.smooth

#> tensor(fu,age)
#>       3.001796



# tint
mod.tint.sub$lambda

#>       tint(fu)      tint(age) tint(fu,age).1 tint(fu,age).2
#>   2.042933e+05   1.165421e+05   5.238644e-01   1.373263e+05

summary(mod.tint.sub)$edf.per.smooth

#>     tint(fu)    tint(age) tint(fu,age)
#>     1.000031     1.000013     1.822548
```

L'approche *tint* réduit les edf des effets propres presque au minimum (1) ce qui correspond à des effets propres linéaires. Au contraire, l'interaction est assez complexe.
On retrouve ce résultat en regardant les paramètres de lissage : alors que le *tensor* a fortement pénalisé les deux directions, l'approche *tint* a fortement pénalisé les effets propres mais faiblement pénalisé le temps dans son interaction avec l'âge (tint(fu,age).1).

```
new.age <- seq(quantile(subdata$age, 0.1), quantile(subdata$age, 0.9), length = 50)

new.time <- seq(0, max(subdata$fu), length = 50)

Z.tensor.sub <- outer(new.time, new.age, function(t, a) predict(mod.tensor.sub,
    data.frame(fu = t, age = a))$haz)

Z.tint.sub <- outer(new.time, new.age, function(t, a) predict(mod.tint.sub, data.frame(fu = t,
    age = a))$haz)

par(mfrow = c(1, 2), mar = c(3, 3, 1.5, 0.5), mgp = c(1.5, 0.5, 0))

persp(new.time, new.age, Z.tensor.sub, col = colors[facet(Z.tensor.sub)], main = "tensor",
    theta = 30, xlab = "\n temps", ylab = "\n âge", zlab = "\n taux", ticktype = "detailed",
    zlim = c(0, 0.7))

persp(new.time, new.age, Z.tint.sub, col = colors[facet(Z.tint.sub)], main = "tint",
    theta = 30, xlab = "\n temps", ylab = "\n âge", zlab = "\n taux", ticktype = "detailed",
    zlim = c(0, 0.7))
```

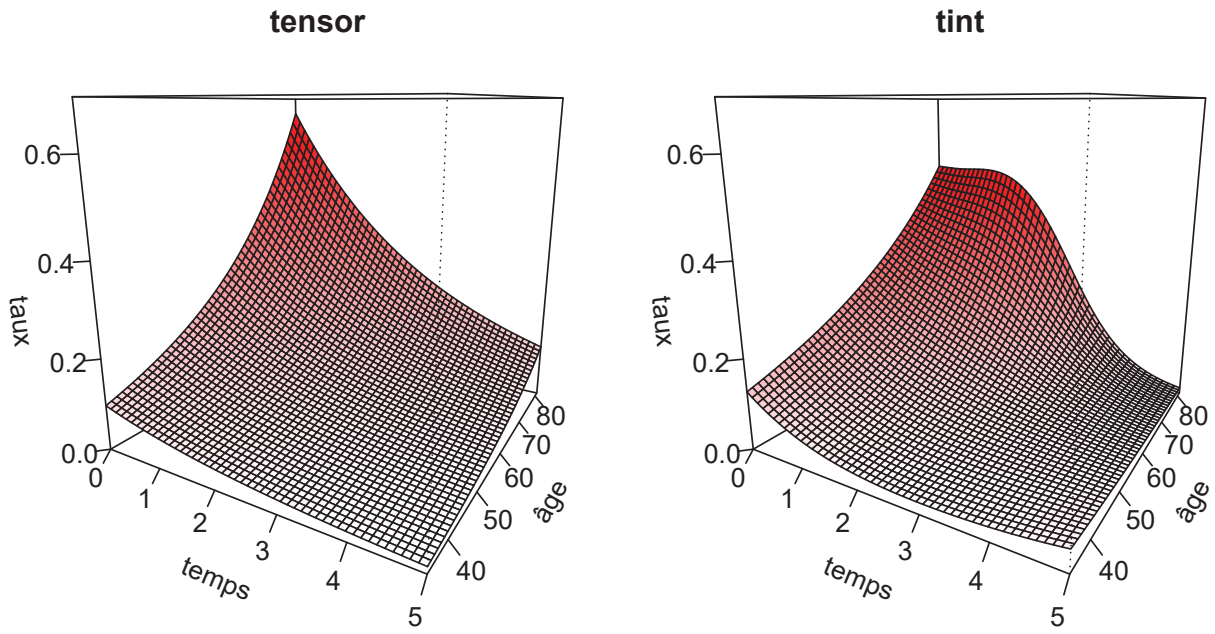

Figure 11.7 – *survPen* - Surfaces du taux, tensor vs tint, deuxième exemple

Les prédictions présentées dans la figure 11.7 confirment que l'interaction âge × temps est beaucoup plus importante d'après l'approche *tint*. Les différences entre les deux approches sont manifestes en début de suivi chez les patients les plus âgés.

Afin de mieux distinguer les différences, regardons des coupes 2D.

```
data2D <- expand.grid(fu = new.time, age = c(50, 60, 70, 80))
data2D$haz.tensor <- predict(mod.tensor.sub, data2D)$haz
data2D$haz.tint <- predict(mod.tint.sub, data2D)$haz

par(mfrow = c(2, 2), mar = c(3, 3, 1.5, 0.5), mgp = c(1.5, 0.5, 0))
plot(new.time, data2D[data2D$age == 50, ]$haz.tensor, type = "l", ylim = c(0,
    0.7), main = "âge 50", xlab = "temps", ylab = "taux", lwd = lwd1)
lines(new.time, data2D[data2D$age == 50, ]$haz.tint, col = "red", lty = 2, lwd = lwd1)
legend("topright", c("tensor", "tint"), lty = c(1, 2), col = c("black", "red"),
    lwd = rep(lwd1, 2))
for (i in c(60, 70, 80)) {
    plot(new.time, data2D[data2D$age == i, ]$haz.tensor, type = "l", ylim = c(0,
        0.7), main = paste("âge", i), xlab = "temps", ylab = "taux", lwd = lwd1)
    lines(new.time, data2D[data2D$age == i, ]$haz.tint, col = "red", lty = 2,
        lwd = lwd1)
}
```

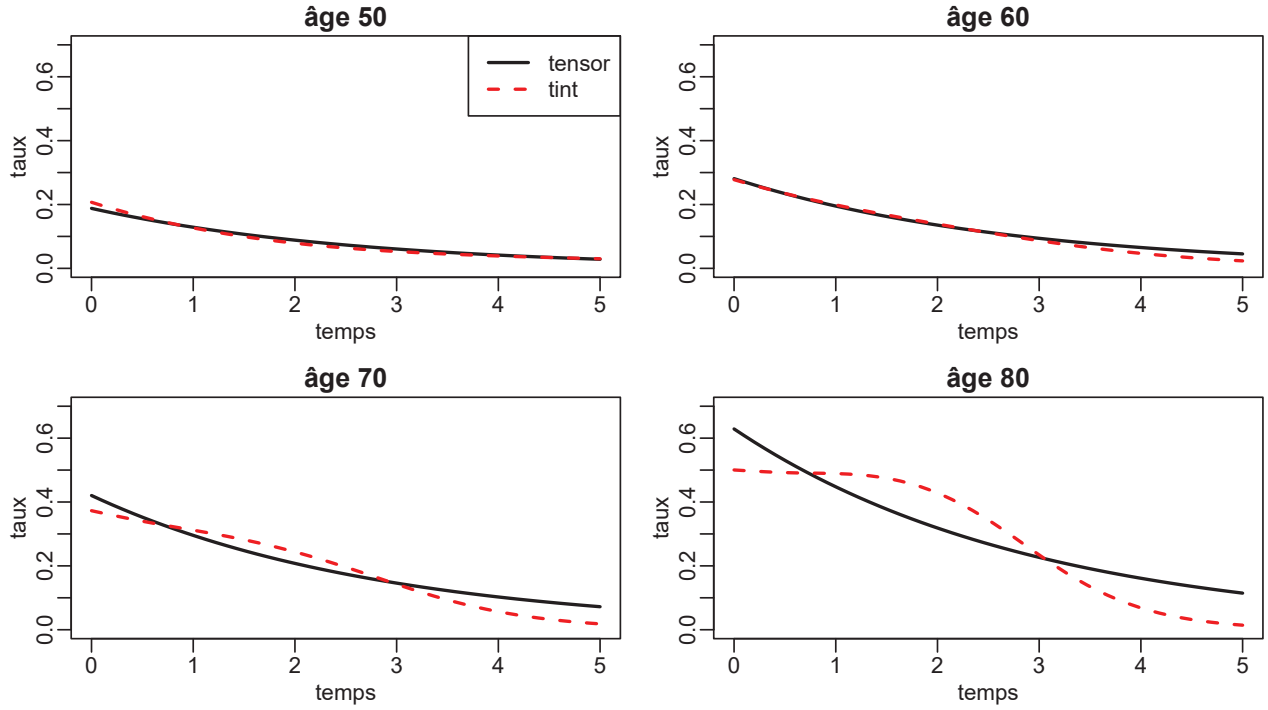

Figure 11.8 – *survPen* - Coupes 2D, tensor vs tint

La figure 11.8 présente les dynamiques de taux aux âges 50, 60, 70 et 80 avec les deux modèles. Afin de choisir parmi les deux modèles, nous pouvons utiliser l'AIC corrigé.

```
mod.tensor.sub$aic2

#> [1] 117.6845

mod.tint.sub$aic2

#> [1] 118.499
```

En l'occurrence, le modèle *tensor* est à privilégier (sur la base du critère de l'AIC corrigé).

11.8.2 Trois dimensions

Le modèle présenté ici est un tensor du temps, de l'âge et de l'année de diagnostic (yod) sur l'échelle du log du taux **en excès**.

```
f.tensor3 <- ~tensor(fu, age, yod, df = c(5, 5, 5))
# modele de taux en exces
mod.tensor3 <- survPen(f.tensor3, data = datCancer, t1 = fu, event = dead, expected = rate)
summary(mod.tensor3)

#> penalized excess hazard model
#>
#> Call:
#> survPen(formula = f.tensor3, data = datCancer, t1 = fu, event = dead,
#>     expected = rate)
#>
#> Coefficients:
#>             Estimate Std. Error z value  Pr(>|z|)
#> (Intercept) -3.43068    0.19226 -17.844 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> likelihood= -2035.52 , penalized likelihood= -2040.37
#> Number of parameters= 125 , effective degrees of freedom= 17.789
#> LAML = 2046.54
#>
#> Smoothing parameter(s):
#> tensor(fu,age,yod).1 tensor(fu,age,yod).2 tensor(fu,age,yod).3
#>            0.2725013           14.1288872           82.3996449
#>
#> edf of smooth terms:
#> tensor(fu,age,yod)
#>           16.78903
#>
#> converged= TRUE
```

```
# predictions
new.age <- seq(50, 90, length = 50)
new.time <- seq(0, 5, length = 50)

Z_1990 <- outer(new.time, new.age, function(t, a) predict(mod.tensor3, data.frame(fu = t,
    yod = 1990, age = a))$haz)
Z_1997 <- outer(new.time, new.age, function(t, a) predict(mod.tensor3, data.frame(fu = t,
    yod = 1997, age = a))$haz)
Z_2003 <- outer(new.time, new.age, function(t, a) predict(mod.tensor3, data.frame(fu = t,
    yod = 2003, age = a))$haz)
Z_2010 <- outer(new.time, new.age, function(t, a) predict(mod.tensor3, data.frame(fu = t,
    yod = 2010, age = a))$haz)
```

La figure 11.9 illustre les capacités de lissage du *tensor* en trois dimensions. Rien n'empêche l'utilisateur d'ajuster un *tensor* à quatre dimensions ou plus mais en pratique cela peut s'avérer extrêmement coûteux en temps et en mémoire. Si la situation s'y prête, on peut essayer avec quatre covariables si l'on fait attention à ne pas spécifier trop de nœuds sur chaque base marginale.

```
par(mfrow = c(1, 2), mar = c(3, 3, 1.5, 0.5), mgp = c(1.5, 0.5, 0))
lbx = "\n temps"
lby = "\n âge"
lbz = "\n taux en excès"
tk1 = "detailed"
persp(new.time, new.age, Z_1990, col = colors[facet(Z_1990)], main = "1990", theta = 20,
    xlab = lbx, ylab = lby, zlab = lbz, ticktype = tk1, zlim = c(0, 1))
persp(new.time, new.age, Z_1997, col = colors[facet(Z_1997)], main = "1997", theta = 20,
    xlab = lbx, ylab = lby, zlab = lbz, ticktype = tk1, zlim = c(0, 1))
par(mfrow = c(1, 2), mar = c(3, 3, 1.5, 0.5), mgp = c(1.5, 0.5, 0))
persp(new.time, new.age, Z_2003, col = colors[facet(Z_2003)], main = "2003", theta = 20,
    xlab = lbx, ylab = lby, zlab = lbz, ticktype = tk1, zlim = c(0, 1))
persp(new.time, new.age, Z_2010, col = colors[facet(Z_2010)], main = "2010", theta = 20,
    xlab = lbx, ylab = lby, zlab = lbz, ticktype = tk1, zlim = c(0, 1))
```

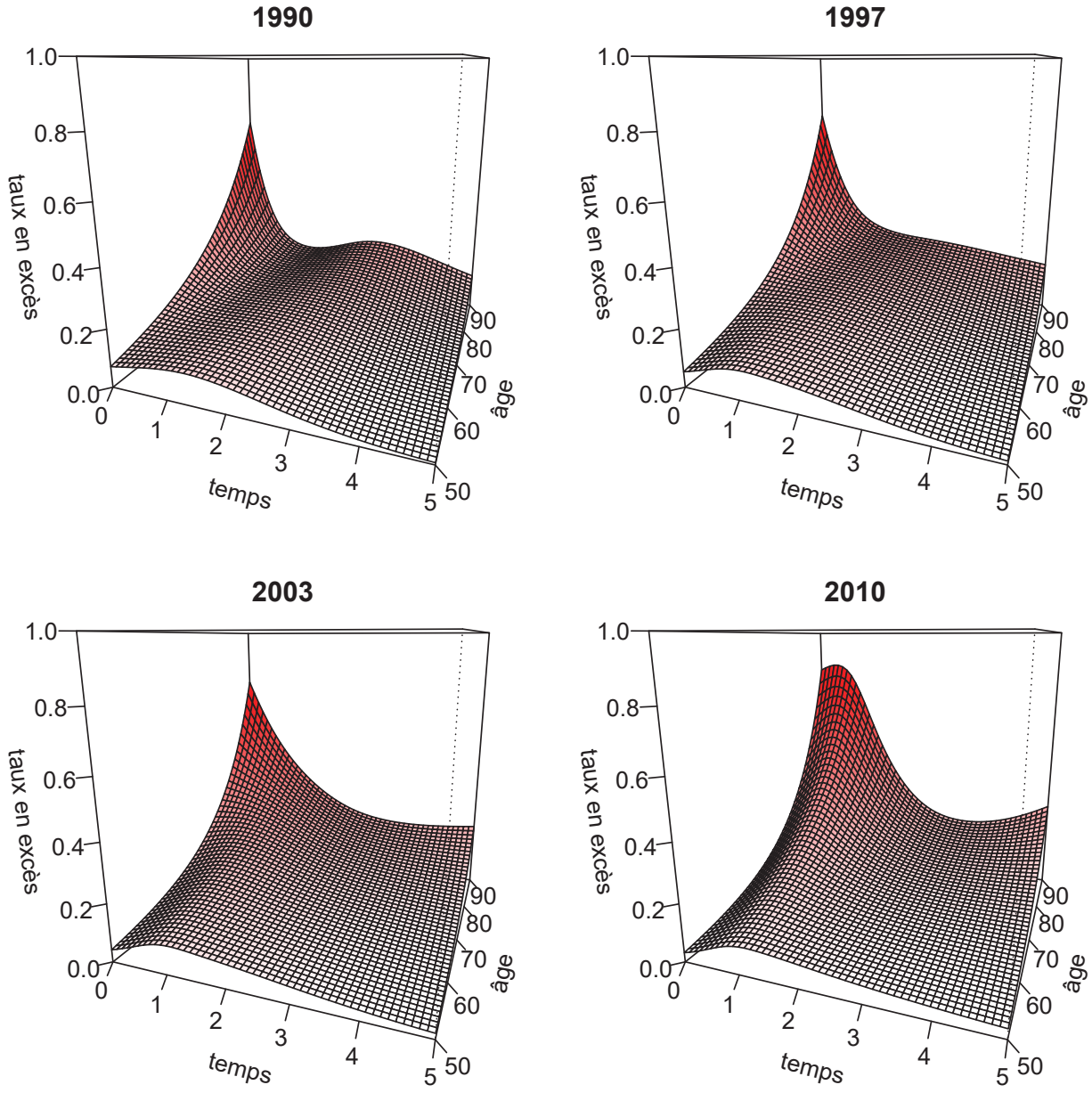

Figure 11.9 – *survPen* - Surfaces prédites par le tensor à trois dimensions

11.9 Interactions entre des splines pénalisées et des termes paramétriques

Les constructeurs *smf*, *tensor* et *tint* acceptent un argument *by* qui permet de créer des interactions avec une variable catégorielle ou une variable continue (voir les sections 4.4.1 et 4.4.2).

11.9.1 Spécification avec des variables continues

Pour les variables continues, une interaction linéaire avec une spline pénalisée peut être spécifiée comme suit (ici la variable continue utilisée est l'âge) :

$$log[h(t, age)] = f(t) + \beta \times age + g(t) \times age$$

avec f et g des splines pénalisées.

Dans *survPen*, ce modèle est spécifié via la formule *smf(t) + smf(t,by=age)*.
L'effet propre de l'âge est inclus dans le terme *smf(t,by=age)*. Si l'on ne souhaite pas l'intégrer, il faut utiliser *tint(t,by=age)*. Cette fonctionnalité est utile lorsque l'on veut ajuster un modèle comme celui-ci :

$$log[h(t, age)] = f(t) + f_2(age) + g(t) \times age$$

où f_2 est une spline pénalisée. Ce modèle est spécifié via *smf(t) + smf(age) + tint(t,by=age)* ou encore *tint(t) + tint(age) + tint(t,by=age)*.

D'un point de vue technique, si une variable *by* est quantitative, alors son i^e élément multiplie la i^e ligne de la matrice de design de la spline associée.

11.9.2 Illustration des variables *by* continues

Les variables *by* continues permettent de produire des interactions entre une spline pénalisée et un effet linéaire d'une autre covariable.

Note importante : en statistiques, il est toujours conseillé de centrer les variables continues avant ajustement. Avec les variables *by* continues, l'application de ce principe est presque toujours indispensable.

```
datCancer$agec <- datCancer$age - 50
```

Spline cubique pénalisée du temps en interaction linéaire avec l'âge :

$$log[h(t, age)] = f(t) + \beta \times age + g(t) \times age$$

```
m <- survPen(~smf(fu) + smf(fu, by = agec), data = datCancer, t1 = fu, event = dead)
m$ll

#> [1] -2112.848
```

Une autre manière d'ajuster le même modèle

```
m.bis <- survPen(~smf(fu) + agec + tint(fu, by = agec, df = 10), data = datCancer,
    t1 = fu, event = dead)
m.bis$ll  # meme log-vraisemblance que m

#> [1] -2112.848
```

Spline cubique pénalisée du temps en interaction linéaire avec l'âge et spline cubique pénalisée de l'âge :

$$log[h(t, age)] = f(t) + g(age) + k(t) \times age$$

```
m2 <- survPen(~smf(fu, df = 10) + smf(agec, df = 10) + tint(fu, by = agec, df = 10),
    data = datCancer, t1 = fu, event = dead)
m2$ll

#> [1] -2110.94
```

ou, de manière équivalente

```
m2.bis <- survPen(~tint(fu, df = 10) + tint(agec, df = 10) + tint(fu, by = agec,
    df = 10), data = datCancer, t1 = fu, event = dead)
m2.bis$ll

#> [1] -2110.94
```

Dans le modèle *m*, l'effet de l'âge est inclus dans le terme *smf(fu,by=agec)*. Dans le modèle *m.bis*, le terme *tint(fu,by=agec,df=10)* est soumis à une contrainte de centrage et l'effet de l'âge n'est pas inclus (on doit donc l'inclure comme un terme paramétrique).
Le constructeur *tint* est particulièrement intéressant lorsque plusieurs splines pénalisées contiennent la même variable *by*.
Notons enfin qu'il faut être prudent lorsque l'on utilise *tint* à la place de *smf* car ils n'ont pas le même nombre de nœuds par défaut (5 vs 10 respectivement).

11.9.3 Spécification avec des variables catégorielles

Les variables *by* catégorielles permettent de spécifier trois types de modèles :

- analyse séparée (ou stratifiée) : la spline pénalisée est répétée autant de fois qu'il y a de modalités
- design séparé avec paramètres de lissage communs : le design est le même que précédemment mais les paramètres de lissage sont communs à toutes les modalités
- pénalisation de la différence : si l'on a k modalités décomposées en 1 modalité de référence et $k-1$ autres modalités. Le modèle contient alors une spline pénalisée commune à toutes les modalités et $k-1$ splines pénalisées représentant les différences entre la modalité de référence et les autres modalités.

Le modèle suivant est un exemple d'analyse séparée :

$$log[h(t, sexe)] = f_{femmes}(t) + f_{hommes}(t)$$

où f_{femmes} et f_{hommes} sont des splines pénalisées correspondant aux taux de base chez les femmes et les hommes respectivement.
Dans ce modèle, les paramètres de régression associés aux hommes sont complètement indépendants de ceux des femmes. Les paramètres de lissage des hommes et des femmes (λ_{hommes} and λ_{femmes}) sont également estimés de manière indépendante. Le modèle est donc bien équivalent à l'ajustement d'un modèle chez les hommes et un autre chez les femmes. Ce modèle est spécifié par la formule *sexe + smf(t,by=sexe)*. Attention ici, contrairement au cas continu, *smf(t,by=sex)* est soumis à une contrainte de centrage et ne contient donc pas l'effet du sexe.

Le design séparé avec paramètres de lissage communs imposerait $\lambda_{hommes} = \lambda_{femmes}$. Cela est utile si l'on pense que les taux de base des hommes et des femmes sont de complexités comparables. Dans ce cas, la formule devient *sexe + smf(t,by=sexe,same.rho=TRUE)* et le modèle n'estime qu'un unique paramètre de lissage. Pour l'analyse séparée, la formule par défaut est en réalité *sexe + smf(t,by=sexe,same.rho=FALSE)*.

La pénalisation sur la différence conduit au modèle

$$log[h(t, sex)] = f(t) + f_{diffhommes}(t)$$

avec f commune aux hommes et aux femmes alors que $f_{diffhommes}$ représente ce que l'on doit ajouter à f afin d'obtenir le taux de base des hommes. En d'autres termes, $f_{diffhommes}$ est une spline sur la différence entre hommes et femmes. Étant donné que cette différence est définie sur l'échelle du logarithme du taux, $f_{diffhommes}$ correspond au logarithme du rapport de taux (*log-hazard ratio*) entre les hommes et les femmes. Ainsi, l'avantage de la pénalisation sur la différence est de permettre de placer la pénalisation non plus sur le log du taux mais sur le log du *hazard ratio*. Ce modèle s'obtient par la formule *sexe + smf(t) + smf(t,by=sexe)* où *sexe* a été transformé en un facteur ordonné (*ordered factor*). La modalité de référence est alors la première modalité du facteur ordonné (ici ce sont les femmes).

11.9.4 Illustration des variables *by* catégorielles

Dans ce qui suit, nous nous intéressons à l'effet d'une variable sexe. Nous simulons des temps de décès à partir d'une distribution de Weibull. Les paramètres de la distribution dépendent du sexe de chaque individu (effet proportionnel).

```
n <- 10000
don <- data.frame(num = 1:n)

shape_men <- 0.9  # premier parametre Weibull
shape_women <- 0.9
scale_men <- 0.6  # second parametre Weibull
scale_women <- 0.7
prop_men <- 0.5  # proportion d'hommes
```

```
set.seed(50)
don$sex <- factor(sample(c("men", "women"), n, replace = TRUE, prob = c(prop_men,
    1 - prop_men)))
don$sex.order <- factor(don$sex, levels = c("women", "men"), ordered = TRUE)
don$shape <- ifelse(don$sex == "men", shape_men, shape_women)
don$scale <- ifelse(don$sex == "men", scale_men, scale_women)
don$fu <- rweibull(n, shape = don$shape, scale = don$scale)
don$dead <- 1  # pas de censure
```

La figure 11.10 donne les taux et les rapports de taux théoriques.

```
hazard <- function(x, shape, scale) {
    exp(dweibull(x, shape = shape, scale = scale, log = TRUE) - pweibull(x, shape = shape,
        scale = scale, log.p = TRUE, lower.tail = FALSE))
}
```

```
nt <- seq(0.01, 5, by = 0.1)

par(mfrow = c(1, 2), mar = c(3, 3, 1.5, 0.5), mgp = c(1.5, 0.5, 0))
plot(nt, hazard(nt, shape_women, scale_women), type = "l", xlab = "temps", ylab = "taux",
    lwd = lwd1, main = "Taux théoriques", ylim = c(0, max(hazard(nt, shape_women,
        scale_women), hazard(nt, shape_men, scale_men))))
lines(nt, hazard(nt, shape_men, scale_men), col = "red", lwd = lwd1, lty = 2)
legend("bottomleft", c("femmes", "hommes"), lty = c(1, 2), lwd = rep(lwd1, 2),
    col = c("black", "red"))

plot(nt, hazard(nt, shape_men, scale_men)/hazard(nt, shape_women, scale_women),
    type = "l", xlab = "temps", ylab = "rapport de taux", lwd = lwd1, ylim = c(0,
        2), main = "HR théorique Hommes / Femmes")
```

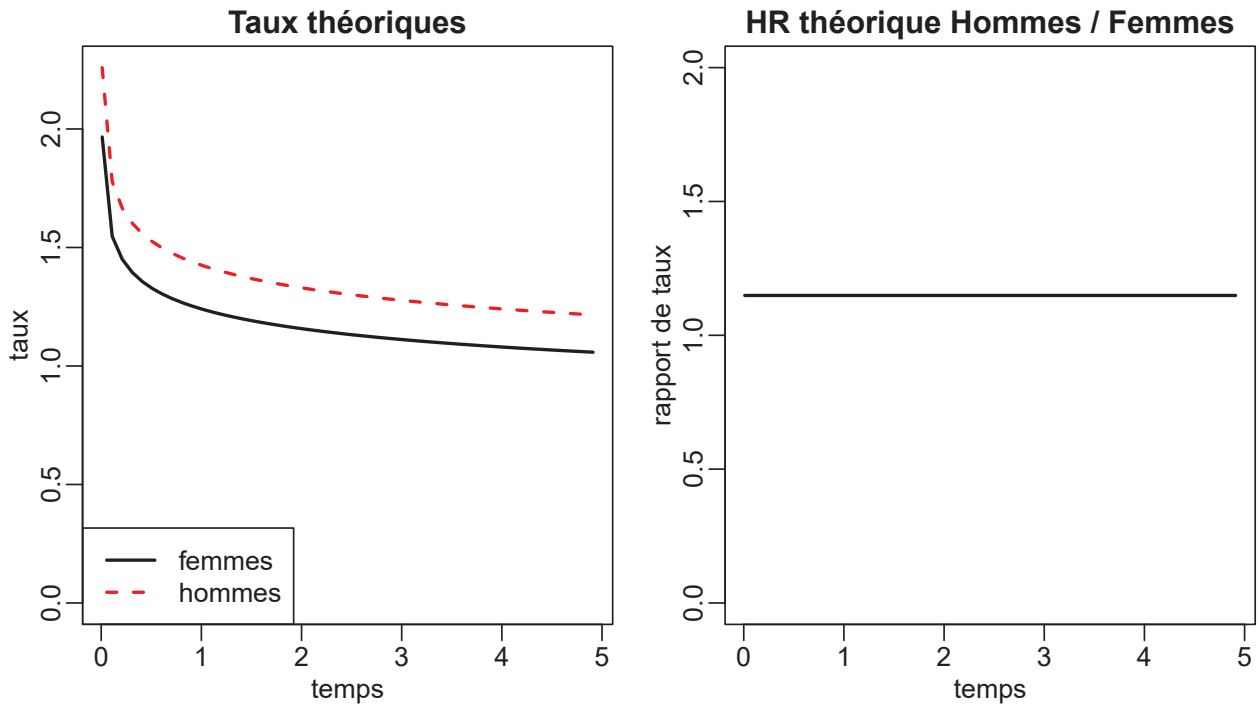

Figure 11.10 – *survPen* - taux et rapport de taux théoriques

Nous allons comparer quatre approches :

— analyse stratifiée via deux modèles disjoints. Nous ajustons un modèle chez les hommes et un autre chez les femmes
— analyse stratifiée via une variable *by* et l'option same.rho=FALSE. Nous ajustons un unique modèle avec une spline pénalisée pour les hommes et une autre pour les femmes
— design séparé avec paramètre de lissage commun. Même chose que précédemment mais en imposant un paramètre de lissage commun aux deux sexes
— pénalisation sur la différence. Nous pénalisons le log-hazard ratio entre hommes et femmes

```
# noeuds
knots.t <- quantile(don$fu, seq(0, 1, length = 10))

# analyse stratifiee avec deux modeles
m.men <- survPen(~smf(fu, knots = knots.t), t1 = fu, event = dead, data = don[don$sex ==
    "men", ])
m.women <- survPen(~smf(fu, knots = knots.t), t1 = fu, event = dead, data = don[don$sex ==
    "women", ])

# by variable avec same.rho = FALSE
m.FALSE <- survPen(~sex + smf(fu, by = sex, same.rho = FALSE, knots = knots.t),
    t1 = fu, event = dead, data = don)

# by variable avec same.rho = TRUE
m.TRUE <- survPen(~sex + smf(fu, by = sex, same.rho = TRUE, knots = knots.t),
    t1 = fu, event = dead, data = don)

# difference
m.difference <- survPen(~sex.order + smf(fu, knots = knots.t) + smf(fu, by = sex.order,
    same.rho = FALSE, knots = knots.t), t1 = fu, event = dead, data = don)
```

Comparons maintenant les prédictions de taux données par la figure 11.11.

```
newt <- seq(0, 5, by = 0.1)

data.pred <- expand.grid(fu = newt, sex = c("women", "men"))
data.pred$men <- ifelse(data.pred$sex == "men", 1, 0)
data.pred$women <- ifelse(data.pred$sex == "women", 1, 0)
data.pred$sex.order <- data.pred$sex  # pas besoin d'ordonner ici
# car le modele enregistre la structure des donnees
data.pred$haz.men <- predict(m.men, data.pred)$haz
data.pred$haz.women <- predict(m.women, data.pred)$haz
data.pred$haz.FALSE <- predict(m.FALSE, data.pred)$haz
data.pred$haz.TRUE <- predict(m.TRUE, data.pred)$haz
data.pred$haz.difference <- predict(m.difference, data.pred)$haz
```

```
par(mfrow = c(1, 2), mar = c(3, 3, 1.5, 0.5), mgp = c(1.5, 0.5, 0))
plot(newt, data.pred[data.pred$sex == "men", ]$haz.men, type = "l", main = "Hommes",
    lwd = lwd1, ylim = c(0, 2.2), xlab = "temps", ylab = "taux")
lines(newt, data.pred[data.pred$sex == "men", ]$haz.FALSE, col = "red", lwd = lwd1,
    lty = 2)
lines(newt, data.pred[data.pred$sex == "men", ]$haz.TRUE, col = "green3", lwd = lwd1,
    lty = 4)
lines(newt, data.pred[data.pred$sex == "men", ]$haz.difference, col = "orange",
    lwd = lwd1, lty = 5)
lines(nt, hazard(nt, shape_men, scale_men), col = "blue3", lty = 3)
legend("bottomleft", c("stratifié", "same.rho=FALSE", "same.rho=TRUE", "différence",
    "vrai"), lty = c(1, 2, 4, 5, 3), col = c("black", "red", "green3", "orange",
    "blue3"), lwd = c(rep(lwd1, 4), 1))

plot(newt, data.pred[data.pred$sex == "women", ]$haz.women, type = "l", main = "Femmes",
    lwd = lwd1, ylim = c(0, 2.2), xlab = "temps", ylab = "taux")
lines(newt, data.pred[data.pred$sex == "women", ]$haz.FALSE, col = "red", lwd = lwd1,
    lty = 2)
lines(newt, data.pred[data.pred$sex == "women", ]$haz.TRUE, col = "green3", lwd = lwd1,
    lty = 4)
lines(newt, data.pred[data.pred$sex == "women", ]$haz.difference, col = "orange",
    lwd = lwd1, lty = 5)
lines(nt, hazard(nt, shape_women, scale_women), col = "blue3", lty = 3)
```

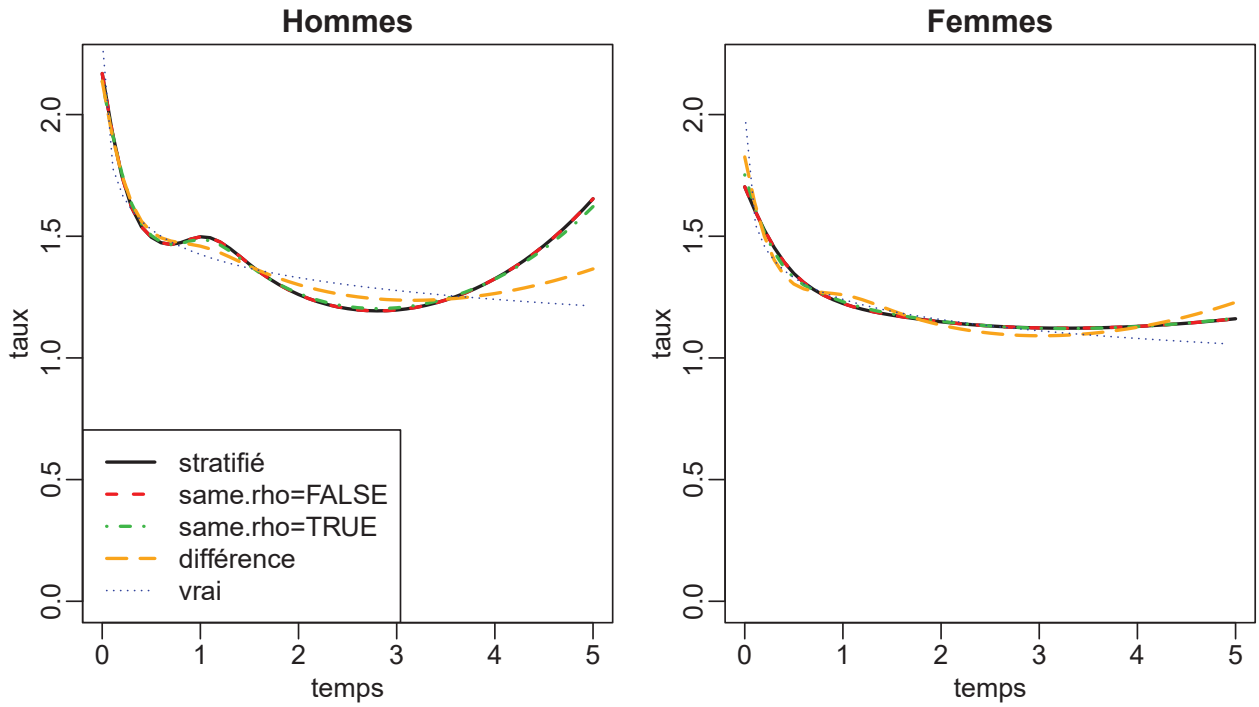

Figure 11.11 – *survPen* - comparaison des taux, variables *by* catégorielles

Comme attendu, les taux issus des modèles *stratifié* et *same.rho=FALSE* sont équivalents.

L'approche *same.rho=TRUE* donnent des prédictions assez similaires par rapport à *same.rho=FALSE*. Quant à la pénalisation sur la différence, elle donne des taux plus lisses chez les hommes (et plus proches de la courbe théorique) et un peu plus variables chez les femmes.

Comparons maintenant les rapports de taux prédits sur la figure 11.12.

```
HR.stratified <- data.pred[data.pred$sex == "men", ]$haz.men/data.pred[data.pred$sex ==
    "women", ]$haz.women
HR.FALSE <- data.pred[data.pred$sex == "men", ]$haz.FALSE/data.pred[data.pred$sex ==
    "women", ]$haz.FALSE
HR.TRUE <- data.pred[data.pred$sex == "men", ]$haz.TRUE/data.pred[data.pred$sex ==
    "women", ]$haz.TRUE
HR.diff <- data.pred[data.pred$sex == "men", ]$haz.difference/data.pred[data.pred$sex ==
    "women", ]$haz.difference

par(mfrow = c(1, 1))
plot(newt, HR.stratified, type = "l", main = "HR, Hommes/Femmes", lwd = lwd1,
    ylim = c(0, 2), xlab = "temps", ylab = "rapport de taux")
lines(newt, HR.FALSE, col = "red", lwd = lwd1, lty = 2)
lines(newt, HR.TRUE, col = "green3", lwd = lwd1, lty = 4)
lines(newt, HR.diff, col = "orange", lwd = lwd1, lty = 5)
abline(h = hazard(nt, shape_men, scale_men)/hazard(nt, shape_women, scale_women),
    lty = 3, col = "blue3")
legend("bottomright", c("stratifié", "same.rho=FALSE", "same.rho=TRUE", "différence",
    "vrai"), lty = c(1, 2, 4, 5, 3), col = c("black", "red", "green3", "orange",
    "blue3"), lwd = c(rep(lwd1, 4), 1))
```

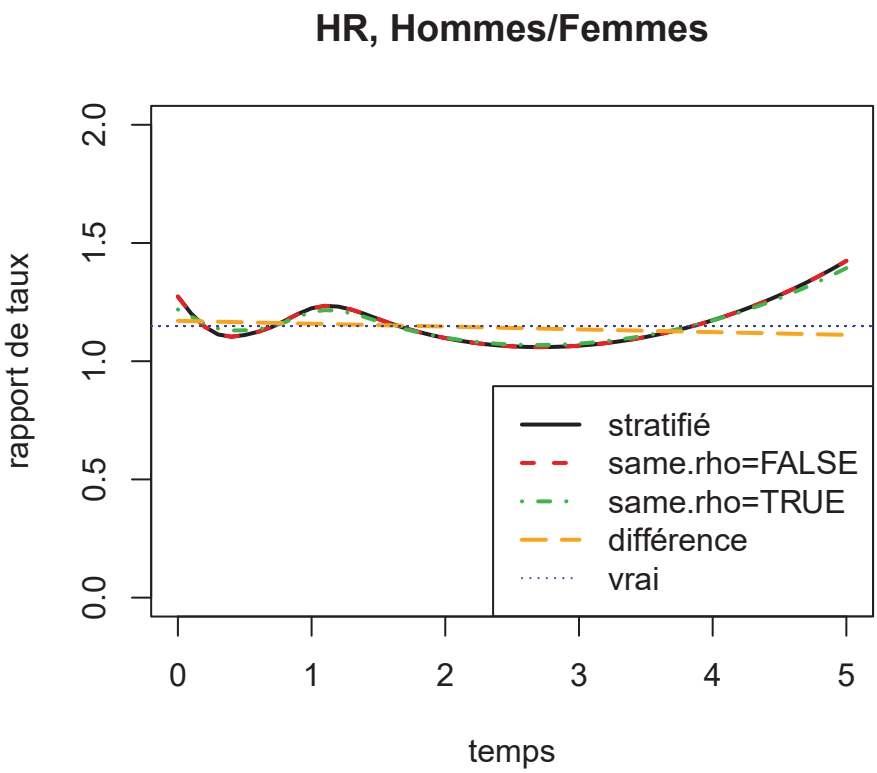

Figure 11.12 – *survPen* - comparaison des rapports de taux, variables *by* catégorielles

Encore une fois, les modèles *stratifié* et *same.rho=FALSE* sont identiques. Ils donnent le même rap-

port de taux presque sinusoïdal qui est assez difficile à justifier.

L'approche *same.rho=TRUE* produit une courbe un peu plus lisse alors que la pénalisation sur la différence donne une ligne droite très proche de la vraie valeur.

Dans ce genre de situation, pénaliser la différence grâce aux variables *by* catégorielles ordonnées est clairement avantageux.

11.10 Effets aléatoires

survPen permet d'inclure des effets aléatoires gaussiens et indépendants en les considérant comme des splines pénalisées par une pénalité Ridge (voir la section 6.4).

La spécification d'effets aléatoires s'effectue à l'aide du constructeur *rd*. Par exemple, si g est une variable catégorielle, alors $rd(g)$ produit un paramètre aléatoire pour chaque modalité de g, les paramètres étant gaussiens et i.i.d.
Si g est une variable catégorielle et x est continue, alors $rd(g, x)$ produit une pente aléatoire gaussienne reliant x à chaque modalité de g.

Ainsi, inclure des effets aléatoires à l'aide de splines pénalisées permet de spécifier des modèles de taux (en excès) à fragilité partagée (*frailty models*, Charvat et al. 2016). À chaque individu i du cluster (en général une zone géographique) j, un modèle possible serait :

$$log[h(t, x_{ij1}, \ldots, x_{ijm})] = \sum_k g_k(t, x_{ij1}, \ldots, x_{ijm}) + w_j$$

où w_j suit une loi normale de moyenne 0. $u_j = exp(w_j)$ est le terme de fragilité. L'effet aléatoire associé à la variable cluster (intercept aléatoire) est spécifié à l'aide du terme *rd(cluster)*. Nous pourrions également spécifier par exemple un effet aléatoire dépendant de l'âge (pente aléatoire) avec *rd(cluster,age)* (w_j deviendrait alors $w_j \times age_{ij}$ dans la formule ci-dessus).
Il faut noter que seuls des effets aléatoires indépendants peuvent être spécifiés pour l'instant. Par exemple, les termes $rd(cluster) + rd(cluster, age)$ créent un intercept aléatoire et une pente aléatoire de l'âge mais il n'est pas possible d'estimer des paramètres de covariance entre les deux.

D'un point de vue technique, le terme *rd(cluster)* produit les paramètres de régression w_j supposés gaussiens et i.i.d, avec une variance inconnue σ^2 à estimer.
Cette hypothèse est équivalente à associer une matrice identité de pénalisation (i.e. une pénalisation Ridge) aux paramètres de régression.
Le paramètre de lissage λ associé au terme *rd(cluster)* est directement lié à σ^2 :

$$\sigma^2 = \frac{1}{\lambda \times S.scale}$$

avec *S.scale* le facteur d'échelle associé à λ (voir la section 8.8.7).

Le logarithme de l'écart-type de l'effet aléatoire est estimé par :

$$log(\hat{\sigma}) = -0.5 \times log(\hat{\lambda}) - 0.5 \times log(S.scale)$$

Et la variance estimée du logarithme de l'écart-type est :

$$Var[log(\hat{\sigma})] = 0.25 \times Var[log(\hat{\lambda})] = 0.25 \times inv.Hess.rho$$

Afin d'illustrer le constructeur *rd*, nous proposons la simulation suivante :

— Pour l'individu i dans le cluster j, le taux théorique s'écrit :

$$h(t) = h_0(t)exp(w_j)$$

où $w_j \sim \mathcal{N}(0, 0.1^2)$ et $h_0(t) = b^{-a} \times a \times t^{a-1}$.

Le taux de base correspond à une distribution de Weibull de paramètres $a = 0.9$ (*shape*) et $b = 2$ (*scale*).

— Nous simulons 50 jeux de données de 2000 individus

— Chaque individu appartient à l'un des 20 clusters créés

```
set.seed(1)

# parameteres Weibull
shape <- 0.9
scale <- 2

# Nombre de jeux de donnees
NFile <- 50

# Nombre d'individus
n <- 2000

# Nombre de clusters
NCluster <- 20


data.rd <- data.frame(cluster = seq(1:NCluster))
cluster <- sample(rep(1:NCluster, each = n/NCluster))
don <- data.frame(num = 1:n, cluster = factor(cluster))  # cluster doit être un facteur
don <- merge(don, data.rd, by = "cluster")[, union(names(don), names(data.rd))]
don <- don[order(don$num), ]
rownames(don) <- NULL

# ecart-type theorique
sd1 <- 0.1

# vecteur des log ecarts-types estimés
log.sd.vec <- rep(as.numeric(NA), NFile)

# temps de suivi maximum
max.time <- 5
```

Pour chaque jeu de données simulées, le modèle suivant est ajusté (pour l'individu i du cluster j) :

$$log[h(t)] = spline(t) + cluster_j$$

```
for (file in 1:NFile) {

    wj <- rnorm(NCluster, mean = 0, sd = sd1)

    don$wj <- wj[don$cluster]

    # temps simules
    u <- runif(n)
    don$fu <- exp(1/shape * (log(-log(1 - u)) - don$wj) + log(scale))

    # censure
    don$dead <- ifelse(don$fu <= max.time, 1, 0)
    don$fu <- pmin(don$fu, max.time)

    # ajustement
    mod.frailty <- survPen(~smf(fu) + rd(cluster), data = don, t1 = fu, event = dead)

    # log ecarts-types estimes
    log.sd.vec[file] <- summary(mod.frailty)$random.effects[, "Estimate"]

}

# Biais relatif en pourcentage dans l'estimation de sd1
100 * (mean(exp(log.sd.vec)) - sd1)/sd1

#> [1] 3.581211
```

L'estimateur de l'écart-type présente un biais relatif très satisfaisant.

Regardons le résumé du dernier modèle ajusté

```
summary(mod.frailty)

#> penalized hazard model
#>
#> Call:
#> survPen(formula = ~smf(fu) + rd(cluster), data = don, t1 = fu,
#>     event = dead)
#>
#> Coefficients:
#>              Estimate Std. Error z value  Pr(>|z|)
#> (Intercept) -0.669393   0.042603 -15.712 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Random effects (log(sd)):
#>              Estimate Std. Error
#> rd(cluster) -1.882433  0.2414823
#>
#> likelihood= -3000.17 , penalized likelihood= -3007.16
#> Number of parameters= 30 , effective degrees of freedom= 15.9556
#> LAML = 3023.06
#>
#> Smoothing parameter(s):
```

```
#>      smf(fu)  rd(cluster)
#> 24007.809301     1.754218
#>
#> edf of smooth terms:
#>     smf(fu) rd(cluster)
#>    2.080618   12.874943
#>
#> converged= TRUE
```

Nous avons $sd(w_j) = \exp(-1.8824335) = 0.1522192$. Cette valeur peut être obtenue à partir du modèle

```
exp(summary(mod.frailty)$random.effects)[1]


#> [1] 0.1522192
```

ou bien recalculée ainsi

```
exp(-0.5 * log(mod.frailty$lambda) - 0.5 * log(mod.frailty$S.scale))[2]


#> rd(cluster)
#>   0.1522192
```

Prédictions pour des clusters spécifiques (*Best Linear Unbiased Prediction*)

```
# Survie a 1 an pour un individu du cluster 6
predict(mod.frailty, data.frame(fu = 1, cluster = 6))$surv


#> [1] 0.5045151



# Survie a 1 an pour un individu du cluster 10
predict(mod.frailty, data.frame(fu = 1, cluster = 10))$surv


#> [1] 0.6102538
```

Prédictions en considérant l'effet aléatoire comme nul (nous devons toujours spécifier un cluster pour la prédiction mais il sera ignoré)

```
# 1-year survival for a patient when random effect is set to zero
predict(mod.frailty, data.frame(fu = 1, cluster = 10), exclude.random = TRUE)$surv


#> [1] 0.5710545
```

11.11 Troncature à gauche

L'argument *t0* permet de spécifier des temps de début de suivi dans le cas où les données seraient tronquées à gauche.

Nota Bene : La variable *begin* a été simulée à des fins d'illustration et n'est pas représentative des données de cancer en général.

```
# predictions
new.time <- seq(0, 5, length = 100)

pred.trunc <- predict(mod.trunc, data.frame(fu = new.time))

par(mfrow = c(1, 2))
plot(new.time, pred.trunc$haz, type = "l", ylim = c(0, 0.2), main = "troncature, taux",
    xlab = "temps", ylab = "taux")

plot(new.time, pred.trunc$surv, type = "l", ylim = c(0, 1), main = "troncature, survie",
    xlab = "temps", ylab = "survie")
```

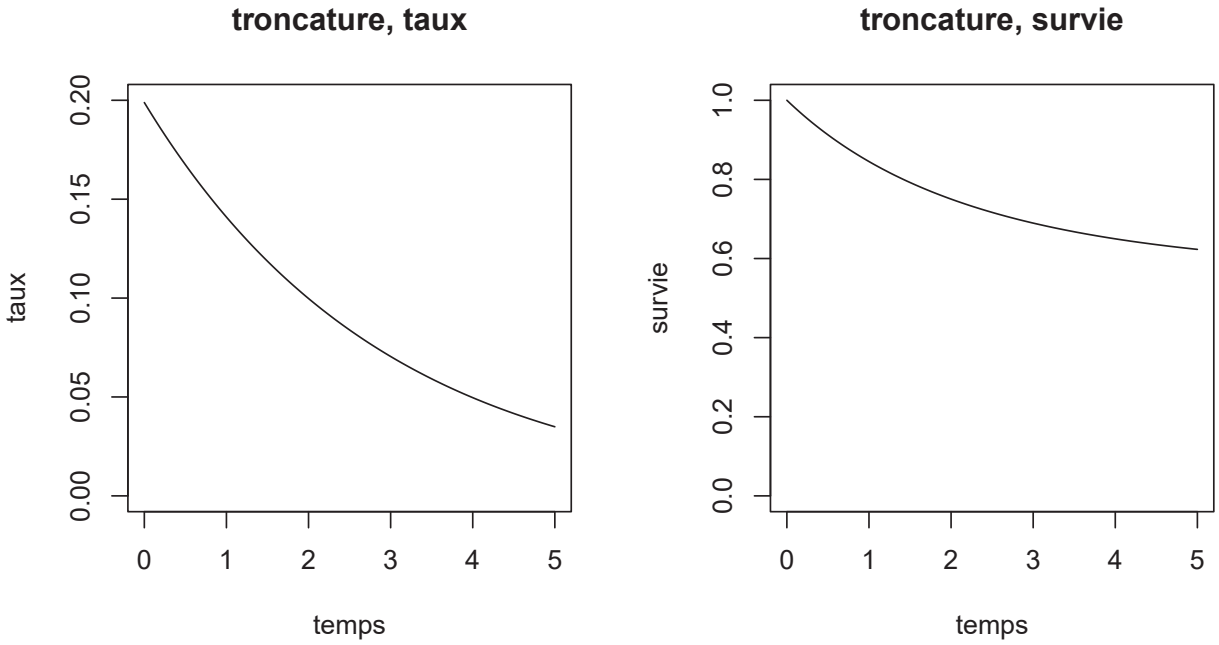

Figure 11.13 – *survPen* - troncature gauche

```
f1 <- ~smf(fu)

mod.trunc <- survPen(f1, data = datCancer, t0 = begin, t1 = fu, event = dead,
    expected = NULL, method = "LAML")
```

La figure 11.13 illustre les prédictions issues d'un modèle prenant en compte la troncature à gauche.

11.12 Autres fonctionnalités utiles

11.12.1 lambda

Le package *survPen* permet d'estimer un ou plusieurs paramètres de lissage via LCV ou LAML. Toutefois, il peut être intéressant d'étudier l'effet d'un paramètre de lissage sur le taux prédit ou la survie prédite. L'argument *lambda* permet de choisir un ou plusieurs paramètres de lissage personnalisés. La figure 11.14 illustre l'effet du paramètre de lissage sur la prédiction.

```
f.pen <- ~smf(fu)

vec.lambda <- c(0, 1000, 10^6)
new.time <- seq(0, 5, length = 100)
```

```
par(mfrow = c(1, 3), mar = c(3, 3, 3, 0.5), mgp = c(1.5, 0.5, 0))

for (i in (1:3)) {

    mod.pen <- survPen(f.pen, data = datCancer, t1 = fu, event = dead, lambda = vec.lambda[i])
    pred.pen <- predict(mod.pen, data.frame(fu = new.time))

    plot(new.time, pred.pen$haz, type = "l", ylim = c(0, 0.2), main = paste0("lambda = ",
        vec.lambda[i]), xlab = "temps", ylab = "taux", col = "black", lwd = lwd1)

}
```

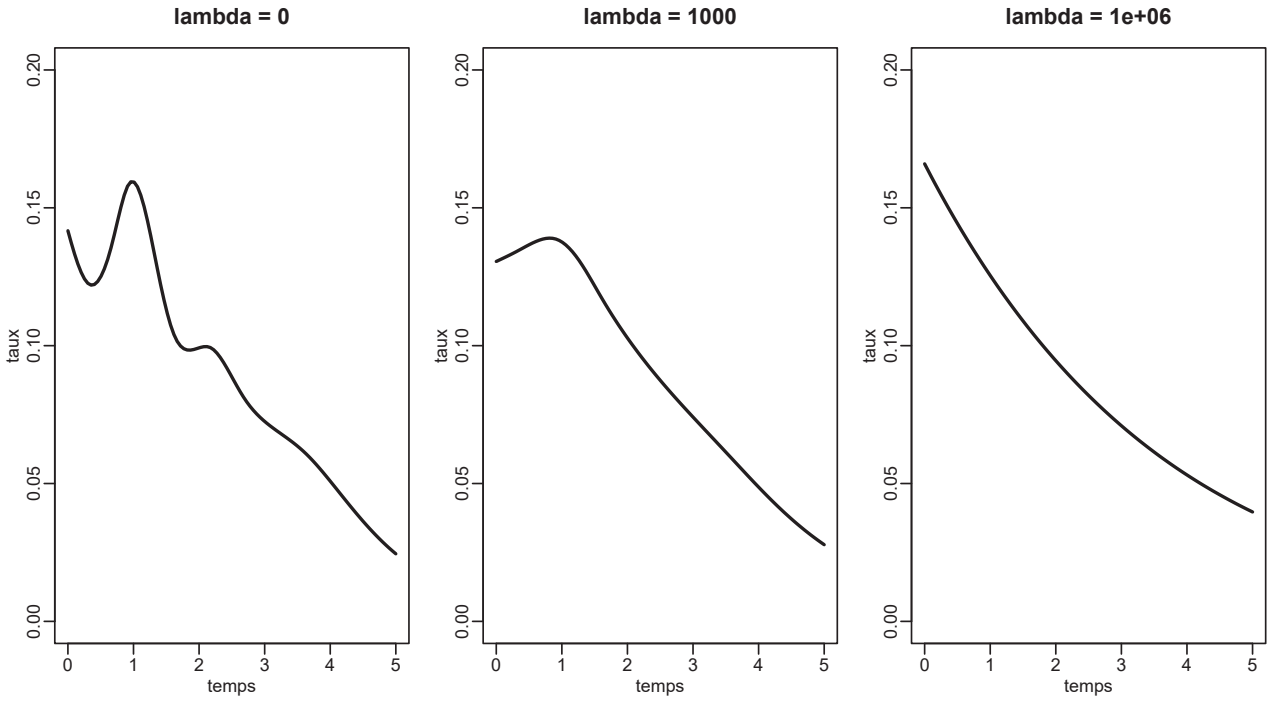

Figure 11.14 – *survPen* - Effet du paramètre de lissage sur la prédiction

11.12.2 beta.ini et rho.ini

Si l'utilisateur souhaite régler un problème de convergence, la première option possible est de spécifier des valeurs initiales différentes de celles par défaut. L'argument *beta.ini* permet de spécifier les paramètres de régression initiaux. Cette option est à explorer notamment si votre modèle de taux en excès ne converge pas. En effet, dans ce cas, vous pouvez essayer d'ajuster le modèle de taux correspondant et d'utiliser ses paramètres de régression estimés comme valeurs initiales pour le modèle de taux en excès. L'argument *rho.ini* permet de spécifier le logarithme des paramètres de lissage initiaux.

```
mod.pen <- survPen(f.pen, data = datCancer, t1 = fu, event = dead, rho.ini = 5)

mod.excess.pen <- survPen(f.pen, data = datCancer, t1 = fu, event = dead, expected = rate,
    rho.ini = 5, beta.ini = mod.pen$coef)
```

11.12.3 detail.rho et detail.beta

Si un problème de convergence survient, il est utile d'identifier ce qu'il se passe à l'intérieur du processus d'optimisation. Les arguments *detail.rho* et *detail.beta* sont là pour rendre la vie plus facile à l'utilisateur dans de telles circonstances.

L'exemple ci-dessous illustre le comportement de *detail.rho=TRUE*

```
mod.pen <- survPen(~smf(fu, df = 5), data = datCancer, t1 = fu, event = dead,
    detail.rho = TRUE)


#> ------------------------------------------------------------------------------------------
#>
#>  Beginning smoothing parameter estimation via  LAML  optimization
#>  ------------------------------------------------------------------------------------------
#>
#> --------------------
#>   Initial calculation
#>  ------------------
#>
#>
#>
#>  new step =  42.3
#> new step corrected =  5
#>
#>
#>  Smoothing parameter selection, iteration  1
#>
#> ------------------------------------------------------------------------------------------
#>
#>  iter  LAML :  1
#>  rho.old=  -1
#>  rho=  4
#>  val.old=  2332.217
#>  val=  2326.77
#>  val-val.old=  -5.4474
#>  gradient=  -0.41
#>
#> ------------------------------------------------------------------------------------------
#>
#>
#>
#>
#>  Smoothing parameter selection, iteration  2
#>
#> ------------------------------------------------------------------------------------------
#>
#>  iter  LAML :  2
```

```
#>  rho.old=  4
#>  rho=  5.0428
#>  val.old=  2326.77
#>  val=  2326.585
#>  val-val.old=  -0.18446
#>  gradient=  0.078
#>
#> ------------------------------------------------------------------------------------
#>
#>
#>
#>
#>  Smoothing parameter selection, iteration  3
#>
#> ------------------------------------------------------------------------------------
#>
#>  iter  LAML :  3
#>  rho.old=  5.0428
#>  rho=  4.8967
#>  val.old=  2326.585
#>  val=  2326.579
#>  val-val.old=  -0.00578
#>  gradient=  0.0012
#>
#> ------------------------------------------------------------------------------------
#>
#>
#>
#>
#>  Smoothing parameter selection, iteration  4
#>
#> ------------------------------------------------------------------------------------
#>
#>  iter  LAML :  4
#>  rho.old=  4.8967
#>  rho=  4.8944
#>  val.old=  2326.579
#>  val=  2326.579
#>  val-val.old=  0
#>  gradient=  3.4e-07
#>
#> ------------------------------------------------------------------------------------
#>
#>
#>
#> Smoothing parameter(s) selection via  LAML  ok,  4 iterations
#>  ------------------------------------------------------------------------------------
```

À chaque itération, on a :

- rho.old = anciennes valeurs des log-paramètres de lissage
- rho = valeurs actuelles des log-paramètres de lissage
- val.old = ancienne valeur de LCV ou LAML (en réalité il s'agit de l'opposé du LAML)
- val = valeurs actuelles de LCV ou LAML
- val-val.old = différence entre les deux valeurs ci-dessus

— gradient = valeur actuelle du gradient de LCV ou LAML en fonction des log-paramètres de lissage

Nous pouvons voir que le premier pas du Newton-Raphson est très important. Quand cela se produit, l'algorithme interdit au pas de dépasser l'argument *step.max* (5 par défaut)

L'exemple ci-dessous illustre le comportement conjoint de *detail.rho=TRUE* et *detail.beta=TRUE*

```
mod.pen <- survPen(~smf(fu, df = 5), data = datCancer, t1 = fu, event = dead,
    detail.rho = TRUE, detail.beta = TRUE)


#> ------------------------------------------------------------------------------------------
#>
#>  Beginning smoothing parameter estimation via  LAML  optimization
#>  ------------------------------------------------------------------------------------------
#>
#> --------------------
#>   Initial calculation
#>  --------------------
#>
#> ------------------------------------------------------------------------------------------
#>  Beginning regression parameter estimation
#>
#> iter beta:  1
#>  betaold=  -2.3569 0 0 0 0
#>  beta=  -2.0665 0.0813 0.0755 -0.2314 -0.9262
#>  abs((beta-betaold)/betaold)=  0.12322 Inf Inf Inf Inf
#>  llold=  -11033.6
#>  ll=  -5032.272
#>  ll-llold=  6001.326
#>
#> iter beta:  2
#>  betaold=  -2.0665 0.0813 0.0755 -0.2314 -0.9262
#>  beta=  -1.4925 0.2144 0.2182 -0.6905 -1.7589
#>  abs((beta-betaold)/betaold)=  0.27774 1.63572 1.88875 1.98346 0.89898
#>  llold=  -5032.272
#>  ll=  -3058.838
#>  ll-llold=  1973.434
#>
#> iter beta:  3
#>  betaold=  -1.4925 0.2144 0.2182 -0.6905 -1.7589
#>  beta=  -0.7311 0.2987 0.3821 -1.2934 -2.4323
#>  abs((beta-betaold)/betaold)=  0.51014 0.39332 0.75127 0.87327 0.38282
#>  llold=  -3058.838
#>  ll=  -2468.365
#>  ll-llold=  590.4724
#>
#> iter beta:  4
#>  betaold=  -0.7311 0.2987 0.3821 -1.2934 -2.4323
#>  beta=  -0.1599 0.239 0.4512 -1.7052 -2.8669
#>  abs((beta-betaold)/betaold)=  0.78136 0.1999 0.18088 0.31832 0.17869
#>  llold=  -2468.365
#>  ll=  -2334.751
#>  ll-llold=  133.6143
#>
#> iter beta:  5
#>  betaold=  -0.1599 0.239 0.4512 -1.7052 -2.8669
```

```
#>  beta=  0.0738 0.1717 0.4377 -1.8332 -3.0324
#>  abs((beta-betaold)/betaold)=  1.46162 0.28173 0.02986 0.07512 0.05774
#>  llold=  -2334.751
#>  ll=  -2320.598
#>  ll-llold=  14.15262
#>
#> iter beta:  6
#>  betaold=  0.0738 0.1717 0.4377 -1.8332 -3.0324
#>  beta=  0.1091 0.1586 0.429 -1.8457 -3.0549
#>  abs((beta-betaold)/betaold)=  0.47877 0.0763 0.0199 0.00678 0.00742
#>  llold=  -2320.598
#>  ll=  -2320.325
#>  ll-llold=  0.27292
#>
#> iter beta:  7
#>  betaold=  0.1091 0.1586 0.429 -1.8457 -3.0549
#>  beta=  0.1099 0.1583 0.4287 -1.8458 -3.0554
#>  abs((beta-betaold)/betaold)=  0.00714 0.00194 0.00069 8e-05 0.00015
#>  llold=  -2320.325
#>  ll=  -2320.325
#>  ll-llold=  0.00014
#>
#> iter beta:  8
#>  betaold=  0.1099 0.1583 0.4287 -1.8458 -3.0554
#>  beta=  0.1099 0.1583 0.4287 -1.8458 -3.0554
#>  abs((beta-betaold)/betaold)=  0 0 0 0 0
#>  llold=  -2320.325
#>  ll=  -2320.325
#>  ll-llold=  0
#>
#>
#>  Beta optimization ok,  8 iterations
#>  ------------------------------------------------------------------------------------
#>
#>
#>  new step =  42.3
#> new step corrected =  5
#>
#>
#>  Smoothing parameter selection, iteration  1
#>
#> ------------------------------------------------------------------------------------
#>  Beginning regression parameter estimation
#>
#> iter beta:  1
#>  betaold=  0.1099 0.1583 0.4287 -1.8458 -3.0554
#>  beta=  0.0221 0.0856 0.3458 -1.8537 -3.0296
#>  abs((beta-betaold)/betaold)=  0.79913 0.4591 0.1934 0.00426 0.00844
#>  llold=  -2323.651
#>  ll=  -2321.422
#>  ll-llold=  2.22931
#>
#> iter beta:  2
#>  betaold=  0.0221 0.0856 0.3458 -1.8537 -3.0296
#>  beta=  0.0222 0.0861 0.3473 -1.8518 -3.0301
```

```
#>  abs((beta-betaold)/betaold)=  0.00608 0.00593 0.00445 0.00103 0.00017
#>  llold=  -2321.422
#>  ll=  -2321.421
#>  ll-llold=  0.00048
#>
#> iter beta:  3
#>  betaold=  0.0222 0.0861 0.3473 -1.8518 -3.0301
#>  beta=  0.0222 0.0861 0.3473 -1.8518 -3.0301
#>  abs((beta-betaold)/betaold)=  0 1e-05 1e-05 0 0
#>  llold=  -2321.421
#>  ll=  -2321.421
#>  ll-llold=  0
#>
#>
#>  Beta optimization ok,  3 iterations
#>  ------------------------------------------------------------------------------------------
#> ------------------------------------------------------------------------------------------
#>
#>  iter  LAML :  1
#>  rho.old=  -1
#>  rho=  4
#>  val.old=  2332.217
#>  val=  2326.77
#>  val-val.old=  -5.4474
#>  gradient=  -0.41
#>
#> ------------------------------------------------------------------------------------------
#>
#>
#>
#>
#>  Smoothing parameter selection, iteration  2
#>
#> ------------------------------------------------------------------------------------------
#>  Beginning regression parameter estimation
#>
#> iter beta:  1
#>  betaold=  0.0222 0.0861 0.3473 -1.8518 -3.0301
#>  beta=  0.0071 0.0533 0.308 -1.8522 -3.02
#>  abs((beta-betaold)/betaold)=  0.68147 0.3815 0.11321 0.00024 0.00332
#>  llold=  -2322.368
#>  ll=  -2322.071
#>  ll-llold=  0.29726
#>
#> iter beta:  2
#>  betaold=  0.0071 0.0533 0.308 -1.8522 -3.02
#>  beta=  0.0071 0.0533 0.3083 -1.852 -3.0201
#>  abs((beta-betaold)/betaold)=  0.00233 0.00126 0.00089 0.00012 4e-05
#>  llold=  -2322.071
#>  ll=  -2322.071
#>  ll-llold=  1e-05
#>
#> iter beta:  3
#>  betaold=  0.0071 0.0533 0.3083 -1.852 -3.0201
#>  beta=  0.0071 0.0533 0.3083 -1.852 -3.0201
```

```
#>  abs((beta-betaold)/betaold)=  0 0 0 0 0
#>  llold=  -2322.071
#>  ll=  -2322.071
#>  ll-llold=  0
#>
#>
#>  Beta optimization ok,  3 iterations
#>  ------------------------------------------------------------------------------------------
#> ------------------------------------------------------------------------------------------
#>
#>  iter  LAML :  2
#>  rho.old=  4
#>  rho=  5.0428
#>  val.old=  2326.77
#>  val=  2326.585
#>  val-val.old=  -0.18446
#>  gradient=  0.078
#>
#> ------------------------------------------------------------------------------------------
#>
#>
#>
#>
#>  Smoothing parameter selection, iteration  3
#>
#> ------------------------------------------------------------------------------------------
#>  Beginning regression parameter estimation
#>
#> iter beta:  1
#>  betaold=  0.0071 0.0533 0.3083 -1.852 -3.0201
#>  beta=  0.0085 0.0577 0.315 -1.8535 -3.0223
#>  abs((beta-betaold)/betaold)=  0.2005 0.08256 0.02171 0.00083 0.00072
#>  llold=  -2321.968
#>  ll=  -2321.963
#>  ll-llold=  0.00454
#>
#> iter beta:  2
#>  betaold=  0.0085 0.0577 0.315 -1.8535 -3.0223
#>  beta=  0.0085 0.0577 0.315 -1.8535 -3.0223
#>  abs((beta-betaold)/betaold)=  7e-05 2e-05 2e-05 0 0
#>  llold=  -2321.963
#>  ll=  -2321.963
#>  ll-llold=  0
#>
#>
#>  Beta optimization ok,  2 iterations
#>  ------------------------------------------------------------------------------------------
#> ------------------------------------------------------------------------------------------
#>
#>  iter  LAML :  3
#>  rho.old=  5.0428
#>  rho=  4.8967
#>  val.old=  2326.585
#>  val=  2326.579
#>  val-val.old=  -0.00578
```

```
#>  gradient=  0.0012
#>
#> ------------------------------------------------------------------------------------------
#>
#>
#>
#>
#>  Smoothing parameter selection, iteration  4
#>
#> ------------------------------------------------------------------------------------------
#>  Beginning regression parameter estimation
#>
#> iter beta:  1
#>  betaold=  0.0085 0.0577 0.315 -1.8535 -3.0223
#>  beta=  0.0085 0.0578 0.3151 -1.8535 -3.0223
#>  abs((beta-betaold)/betaold)=  0.00282 0.00122 0.00032 1e-05 1e-05
#>  llold=  -2321.962
#>  ll=  -2321.962
#>  ll-llold=  0
#>
#> iter beta:  2
#>  betaold=  0.0085 0.0578 0.3151 -1.8535 -3.0223
#>  beta=  0.0085 0.0578 0.3151 -1.8535 -3.0223
#>  abs((beta-betaold)/betaold)=  0 0 0 0 0
#>  llold=  -2321.962
#>  ll=  -2321.962
#>  ll-llold=  0
#>
#>
#>  Beta optimization ok,  2 iterations
#>  ------------------------------------------------------------------------------------------
#> ------------------------------------------------------------------------------------------
#>
#>  iter  LAML :  4
#>  rho.old=  4.8967
#>  rho=  4.8944
#>  val.old=  2326.579
#>  val=  2326.579
#>  val-val.old=  0
#>  gradient=  3.4e-07
#>
#> ------------------------------------------------------------------------------------------
#>
#>
#>
#> Smoothing parameter(s) selection via  LAML  ok,  4 iterations
#>  ------------------------------------------------------------------------------------------
```

Ici, au sein de chaque itération des log-paramètres de lissage, pour chaque itération des paramètres de régression, on a :

- betaold = anciennes valeurs des paramètres de régression
- beta = valeurs actuelles des paramètres de régression
- abs((beta-betaold)/betaold) = valeur absolue de la différence relative entre les valeurs actuelles et anciennes des paramètres de régression (un des critères utilisés pour déclarer la convergence)
- llold = ancienne valeur de la log-vraisemblance pénalisée

— ll = valeur actuelle de la log-vraisemblance pénalisée
— ll-llold = différence entre les deux valeurs ci-dessus

En utilisant *detail.rho* ou *detail.beta*, il est possible qu'un message indiquant qu'une hessienne (de LCV, LAML ou de la vraisemblance pénalisée) a été perturbée s'affiche. Tant que cette perturbation ne se produit pas à la convergence, l'utilisateur ne doit pas s'alarmer. L'algorithme s'assure juste que chaque pas effectué aille dans la bonne direction. Cependant, une perturbation de la hessienne à convergence peut être révélatrice d'un problème de convergence (voir les indicateurs *Hess.beta.modif* et *Hess.rho.modif* renvoyés par le modèle).

Chapitre 12

Comparaison avec des approches existantes

Dans ce chapitre, nous comparons le package *survPen* à différentes approches de modélisation pour l'analyse des temps d'événement.
La première comparaison s'intéresse aux temps d'exécution de *survPen* et d'un certain nombre de méthodes brièvement exposées au chapitre 7.
La seconde et la dernière comparaisons s'intéressent aux propriétés statistiques des méthodes implémentées dans *survPen*, *rstpm2* et *mexhaz*.

Les sections 12.1 et 12.2 sont issues du *Supplementary* de Fauvernier et al. (2019b).

12.1 Temps d'exécution

Cette section compare les temps de calcul des approches implémentant les modèles pénalisés en survie dans le logiciel R : *frailtypack*, *R2BayesX*, *bamlss*, *gss*, *rstpm2*, et *survPen*.
À cette liste, nous ajoutons l'approche « Poisson + *gam* » qui correspond à la méthode détaillée par Remontet et al. (2019). Dans cette approche, un modèle de Poisson est ajusté sur des données augmentées à l'aide de la fonction *gam* du package *mgcv* (la fonction de lien est modifiée afin de prendre en compte les taux attendus).
Le tableau 12.1 présente un bref descriptif des fonctionnalités de chaque approche en termes de splines pénalisées. Les colonnes indiquent si la méthode permet : 1) d'ajuster des splines pénalisées pour décrire l'effet d'une variable continue (colonne « Spline pénalisée ») ; 2) d'ajuster des produits tensoriels pénalisés pour décrire les interactions entre plusieurs covariables continues (« Tensor ») ; 3) d'ajuster des modèles de taux en excès (« Taux en excès ») ; 4) d'ajuster des splines de régression par opposition aux splines de lissage (« splines de régression ») ; et 5) de modéliser directement le taux ou le logarithme du taux (« Taux »).

			Fonctionnalité [1]				
Méthode	package R	Fonction	Spline pénalisée	Tensor	Taux en excès	Splines de régression	Taux
frailtypack	oui	*frailtyPenal*	Seulement pour le temps	non	non	oui	oui
R2BayesX	oui	*bayesx*	oui	non	non	oui	oui
bamlss	oui	*bamlss*	oui	oui	non	oui	oui
gss	oui	*sshzd*	oui	oui	non	non [2]	oui
rstpm2	oui	*pstpm2*	oui	oui	oui	oui	non [3]
Poisson + *gam*	non [4]	-	oui	oui	oui	oui	oui
survPen	oui	*survPen*	oui	oui	oui	oui	oui

[1] Ce tableau compare uniquement les fonctionnalités en prenant celles de *survPen* en référence (splines de régression multidimensionnelles pénalisées pour les modèles de taux et de taux en excès); il ne fournit pas une description exhaustive de chaque package.

[2] Le package *gss* utilise des splines de lissage.

[3] Le package *rstpm2* propose différentes échelles de modélisation, dont le taux cumulé, mais le taux ou le logarithme du taux n'en font pas partie.

[4] L'approche "Poisson + *gam*" n'est pas disponible comme package mais le code est disponible sur github (https://github.com/RocheLHCL/SMMR_Remontet2018/).

Tableau 12.1 – Fonctionnalités de différentes approches en termes de splines pénalisées pour les modèles de survie

Le package *frailtypack* est utile pour ajuster des modèles mixtes hiérarchiques, joints ou à fragilité partagée ; *R2BayesX* et *bamlss* sont utiles pour ajuster des modèles additifs complexes, comme par exemple en modélisant conjointement les paramètres de position, de forme et d'échelle de la distribution de la variable à expliquer (possibilité non restreinte aux modèles de survie) ; *gss* permet d'ajuster des splines de lissage dans différents types de modèles non restreints à l'analyse de survie ; *rstpm2* propose différentes échelles de modélisation dans le cadre des modèles de survie (avec possibilité d'effets aléatoires) ; et, finalement, *mgcv* est le package de référence en termes de splines de régression pénalisées dans le cadre des GAM mais également pour des distributions en dehors de la famille exponentielle (modèle de Cox par exemple).

Pour comparer les temps d'exécution de ces différentes approches, trois modèles de survie et un modèle de survie nette sont considérés (le premier modèle est en réalité décliné en deux versions) :

$$\text{modèle spline} : log[f(t)] = spline(t, df = 10)$$

$$\text{modèle spline (pour frailtypack)} : f(t) = spline(t, df = 10)$$

$$\text{modèle tensor2} : log[f(t, a)] = tensor(t, a, df = c(5, 5))$$

$$\text{modèle tensor3} : log[f(t, a, y)] = tensor(t, a, y, df = c(5, 5, 5))$$

$$\text{modèle tensor3\_net} : log[f_E(t, a, y)] = tensor(t, a, y, df = c(5, 5, 5))$$

Dans ces modèles, t est le temps de suivi, a est l'âge au diagnostic et y est l'année de diagnostic. La fonction f représente le taux instantané (le taux en excès pour f_E) pour toutes les approches excepté *rstpm2* pour lequel f représente le taux cumulé (taux cumulé en excès pour f_E). Le premier modèle comporte 10 paramètres de régression et 1 paramètre de lissage. Le deuxième modèle comporte 25 paramètres de régression pour 2 paramètres de lissage. Les modèles trois et quatre comportent chacun 125 paramètres de régression pour 3 paramètres de lissage. Le nombre de paramètres de lissage est différent pour le package *gss* car il s'appuie sur une décomposition ANOVA (voir la section 4.4.3 de

cette thèse ou la section 5.6.3 de Wood 2017).

Les modèles ont été ajustés sur deux jeux de données simulées (tailles : 2 000 et 20 000) issus des simulations du scénario Col de l'utérus (N=2 000) décrit dans le chapitre 10. Le jeu de données de taille 2 000 est choisi aléatoirement (il correspond au jeu de données du package *survPen* nommé « datCancer ») et le jeu de données de taille 20 000 correspond à la concaténation des 10 premiers jeux de données simulées (chaque jeu comportant 2 000 individus). Au vu des fonctionnalités décrites dans le tableau 12.1, les quatre modèles ne peuvent pas être ajustés par les sept méthodes. Par exemple, le modèle tensor3_net ne peut être ajusté que par *rstpm2*, *survPen* et l'approche « Poisson + *gam* ».

La procédure d'augmentation de données nécessaire à l'utilisation de « Poisson + *gam* » transforme les 2 000 et 20 000 individus en 93 963 et 934 379 pseudo-observations, respectivement.
Concernant *R2BayesX*, nous considérons la version REML (Kneib and Fahrmeir, 2007) et la version MCMC. Pour ce qui est de *survPen*, nous utilisons le critère LCV afin d'être comparable à *rstpm2*.

Nous utilisons la version 3.5.2 de R avec les versions de packages suivantes : *frailtypack_3.0.2.1*, *R2BayesX_1.1-1*, *bamlss_1.0-1*, *gss_2.1-9*, *rstpm2_1.4.5*, *mgcv_1.8-26* et *survPen_1.0.1*. Les tableaux 12.2 et 12.3 présentent les temps d'exécution des différents modèles sur un PC équipé d'un Intel Xeon CPU E3-1245 v5 cadencé à 3.50 GHz et disposant de 16 Go de mémoire vive.

	Modèle			
Méthode	Spline	Tensor2	Tensor3	Tensor_net3
frailtypack	0,7			
R2BayesX REML	1,2			
R2BayesX MCMC	37,0			
bamlss	353,0	1 378,0	10 851,7	
gss	0,0	4,6	5 608,3	
rstpm2	1,2	19,1	562,7	456,2
Poisson + *gam*	1,3	6,8	232,5	463,3
survPen	0,5	3,6	55,4	75,6

Tableau 12.2 – Temps d'exécution (en secondes) pour N = 2 000

	Modèle			
Méthode	Spline	Tensor2	Tensor3	Tensor_net3
frailtypack	19,0			
R2BayesX REML	2,8			
R2BayesX MCMC	525,6			
bamlss	4 828,3	13 751,9	109 708,2	
gss	0,2	16,2	4,712,1	
rstpm2	38,1	37,2	4 270,6	4 197,7
Poisson + *gam*	12,4	91,3	1 755,4	4 270,7
survPen	5,2	23,4	576,3	680,1

Tableau 12.3 – Temps d'exécution (en secondes) pour N = 20 000

Quelle que soit la taille d'échantillon, *survPen* est plus rapide que ses concurrents dans les configurations les plus complexes (tensor3 et tensor3_net) mais *R2BayesX* (version REML) et *gss* sont plus

rapides pour les deux premiers modèles et la taille d'échantillon à 20 000. L'utilisation de survPen dans les configurations les plus complexes apporte un gain absolu de temps non négligeable. *R2BayesX* (version MCMC) et *bamlss*, qui s'appuient sur des algorithmes MCMC, sont logiquement les approches les plus lentes (*bamlss* étant de loin le plus demandeur en temps de calcul).
Malgré l'utilisation d'une approximation efficiente pour l'ajustement des splines de lissage (Du and Gu, 2006; Gu, 2014), les splines de lissage de *gss* sont plus gourmandes en temps de calcul que les splines de régression pénalisées utilisées par *survPen* (dans les configurations les plus complexes). Cependant, les modèles ajustés par *survPen* et *gss* ne sont pas entièrement comparables car *gss* s'appuie sur une décomposition en effets propres + interactions. En outre, les modèles *gss* ont également été ajustés à l'aide de la fonction *sshzd1*, qui est une version plus rapide de *sshzd*, mais la perte en termes de qualité d'ajustement était trop importante (exploration non présentée).

12.2 Estimation de splines pénalisées : comparaison à *rstpm2*

Parmi les nombreux packages R existant, seul le package *rstpm2* (à notre connaissance) permet d'ajuster des tensors pénalisés en survie nette. Nous avons donc décidé de comparer les performances statistiques de *survPen* avec celles de *rstpm2*.
La méthode utilisée dans *rstpm2* est détaillée dans l'article de Liu et al. (2018). Les principales différences entre *rstpm2* et *survPen* sont les suivantes :

- *rstpm2* permet de modéliser l'effet des covariables sur différentes échelles dont celle du logarithme du taux cumulé mais pas sur l'échelle du logarithme du taux comme c'est le cas dans *survPen*
- le calcul du taux cumulé ne nécessite pas d'intégration numérique mais impose une procédure d'optimisation sous contrainte (contrairement à *survPen*)
- *rstpm2* utilise des méthodes numériques pour calculer les dérivées de la vraisemblance et des critères d'estimation des paramètres de lissage tandis que *survPen* utilise les dérivées explicites

À partir des scénarios œsophage et col de l'utérus présentés au chapitre 10, nous avons décidé de reproduire les résultats présentés dans les figures 10.2 et 10.3. Seule la taille d'échantillon 2 000 est utilisée afin de limiter le temps de calcul.

À l'aide de *rstpm2*, nous ajustons le modèle suivant :

$$log[H_E(t, age, annee)] = tensor(log(t), age, annee)$$

où H_E est le taux de mortalité cumulé en excès.

Les mêmes bases de splines que pour *survPen* ont été utilisées (à noter toutefois que *survPen* modélise l'effet du temps t tandis que *rstpm2* modélise l'effet de $log(t)$). Le nombre de nœuds était également identique (6 pour le temps, 5 pour l'âge et 4 pour l'année). Les positions des nœuds correspondent aux valeurs par défaut (quantiles du log du temps, de l'âge et de l'année).

Les résultats sont présentés dans les figures 12.1 et 12.2.

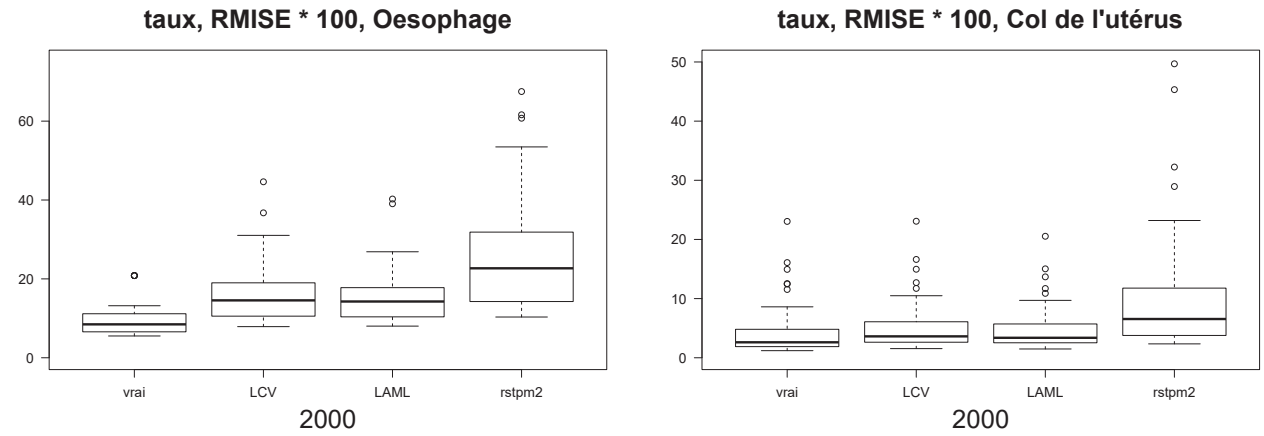

Figure 12.1 – Comparaison *rstpm2*. Boxplots de la RMISE (multipliée par 100) sur le taux en excès pour chaque scénario (taille d'échantillon = 2 000).

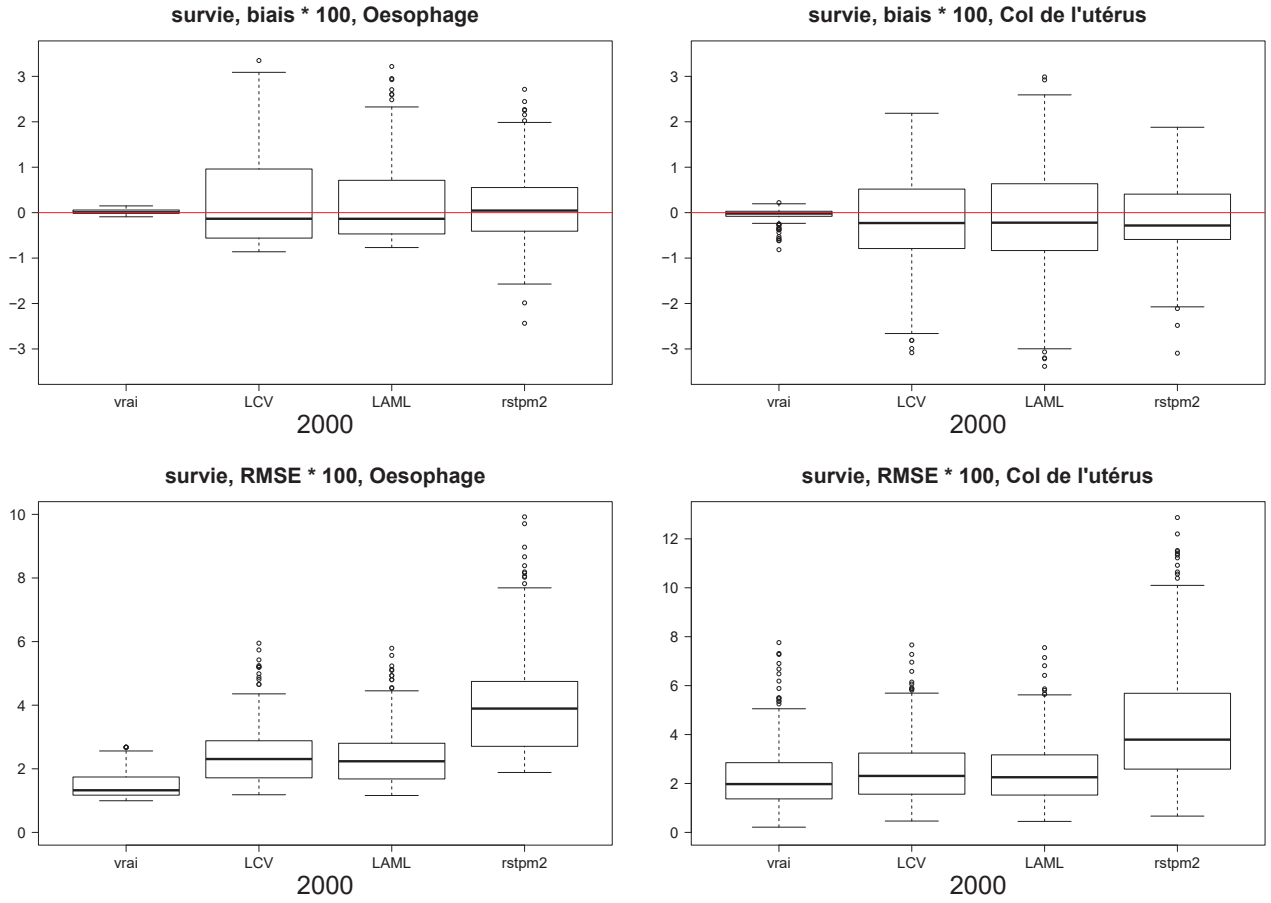

Figure 12.2 – Comparaison *rstpm2*. Boxplots du biais et de la RMSE (multipliés par 100) sur la survie nette pour chaque scénario (taille d'échantillon = 2 000).

Dans le scénario du Col de l'utérus, *rstpm2* présente un biais légèrement inférieur à *survPen*. Cependant, dans les deux scénarios, *survPen* est supérieur à *rstpm2* en termes de RMISE sur le taux en excès et de RMSE sur la survie nette.

Dans le scénario Oesophage, la RMISE de *survPen* LCV est inférieure à celle de *rstpm2* dans 81% des combinaisons d'âges et d'années. Dans le scénario Col, cette proportion atteint 84%. En outre, sur 2 000 modèles ajustés par *rstpm2*, 55 n'ont pas convergé (44 pour l'Oesophage et 11 pour le Col de l'utérus) alors qu'un seul modèle *survPen* n'avait pas réussi à converger (dans le scénario Col avec le critère LCV).
Il semble donc que *survPen* soit meilleur que *rstpm2* en termes de performances statistiques et de stabilité dans les deux scénarios étudiés.

12.3 Estimation d'effets aléatoires : comparaison à *mexhaz*

mexhaz fut le premier package R à proposer des modèles de taux en excès permettant de prendre en compte simultanément la non-linéarité, la non-proportionnalité, les interactions ainsi que les données corrélées. *mexhaz* et *survPen* proposent tous deux l'ajustement de modèles de taux en excès à effets mixtes. Alors que *survPen* utilise une astuce qui considère un effet aléatoire comme une spline pénalisée, *mexhaz* offre un cadre d'inférence plus classique mais aussi plus efficient. L'objectif de cette section est de comparer les performances statistiques des deux approches.

12.3.1 Simulation

Les résultats de simulation présentés s'appuie sur le scénario C détaillé dans Charvat et al. (2016). Ce scénario propose un contexte d'effet non-proportionnel du sexe et d'effets linéaires et proportionnels de l'âge et d'un indice de défavorisation.

Vrai modèle : Pour l'individu i dans le cluster j, on a :

$$h_E(t, age_{ij}, sexe_{ij}, DI_j) = h_{0,sexe_{ij}}(t) \times exp(0.05 \times age_{ij} + 0.02 \times DI_j + w_j)$$

Avec $w_j \sim \mathcal{N}(0, 0.5^2)$ et

$$h_{0,Hommes}(t) = 0.25 \times 0.7 \times t^{0.7-1} \qquad h_{0,Femmes}(t) = 0.18 \times 0.8 \times t^{0.8-1}$$

1000 jeux de données ont été simulés avec une taille de 1000 individus mais en faisant varier le nombre de clusters. Cela entraîne 4 configurations possibles : 10 clusters de 100 individus chacun, 20 clusters de 50 individus, 50 clusters de 20 individus et 100 clusters de 10 individus.

12.3.2 Modèles ajustés

Modèle non pénalisé ajusté par mexhaz et par survPen

$$log[h_E(t, age_{ij}, sexe_{ij}, DI_j)] = bs(t, knots = 1) \times sexe_{ij} + \beta_{age} \times age_{ij} + \beta_{DI} \times DI_j + w_j$$

Avec $w_j \sim \mathcal{N}(0, \sigma^2)$

Attention, le modèle décrit dans les simulations de Charvat et al. (2016) a été réajusté à l'aide du package *mexhaz* (qui n'existait pas à l'époque de l'article). Les résultats présentés ici sont donc différents de ceux publiés dans l'article.
Le modèle ajusté par *mexhaz* comprend 12 paramètres fixes (5 paramètres pour la spline des hommes,

5 pour la spline des femmes, 1 pour l'âge et 1 pour l'indice de défavorisation) et 1 paramètre de variance. En revanche, le modèle ajusté par *survPen* comprend ces mêmes 12 paramètres et autant de paramètres supplémentaires qu'il y a de clusters (voir la section 9.3). Le modèle *survPen* comprend également un paramètre de lissage qui est directement lié au paramètre de variance à estimer.

Modèle pénalisé ajusté par survPen

$$log[h_E(t, age_{ij}, sexe_{ij}, DI_j)] = s(t, df = 10) \times sexe_{ij} + \beta_{age} \times age_{ij} + \beta_{DI} \times DI_j + w_j$$

Où s est une spline pénalisée à 10 nœuds et $w_j \sim \mathcal{N}(0, \sigma^2)$. Le modèle pénalisé comprend donc 1 paramètre de lissage pour le taux de base des hommes, 1 pour celui des femmes et 1 dernier pour l'effet aléatoire.

12.3.3 Résultats

Configuration	Paramètres (vraies valeurs)	Spline mixed (Charvat et al. 2016) réajusté avec mexhaz			survPen avec même modèle que Charvat et al. (2016)			survPen avec une spline pénalisée par sexe		
		Biais[a]	CP[b]	RMSE[c]	Biais[a]	CP[b]	RMSE[c]	Biais[a]	CP[b]	RMSE[c]
Nombre de clusters: 10	β_{age} (0.05)	-0.7	93.9	0.004	-0.3	94.6	0.004	-0.1	94.7	0.004
	β_{DI} (0.02)	7.3	87.2	0.096	4.3	90.8	0.095	5.4	90.7	0.096
Taille moyenne: 100	σ (0.5)	-14	87.8	0.146	-1.7	94.8	0.143	-1.5	94.7	0.143
Nombre de clusters: 20	β_{age} (0.05)	-1.1	95.0	0.004	-0.5	95.0	0.004	-0.3	95.3	0.004
	β_{DI} (0.02)	7.7	92.9	0.1	16.5	94.0	0.098	16.5	94.0	0.098
Taille moyenne: 50	σ (0.5)	-5.6	94.0	0.105	0.9	95.1	0.106	1.26	95.1	0.107
Nombre de clusters: 50	β_{age} (0.05)	-1	94.2	0.005	0.1	94.3	0.004	0.4	94.4	0.004
	β_{DI} (0.02)	7.1	94.5	0.063	5.4	95.6	0.061	6.0	95.5	0.061
Taille moyenne: 20	σ (0.5)	-1.7	96.6	0.079	0.1	96.6	0.079	0.8	96.3	0.080
Nombre de clusters: 100	β_{age} (0.05)	-0.8	93.9	0.005	0.5	94.5	0.004	0.9	94.5	0.005
	β_{DI} (0.02)	5.3	95.4	0.049	-2.0	95.6	0.047	-1.0	95.8	0.047
Taille moyenne: 10	σ (0.5)	-1	96.9	0.077	-2.8	96.9	0.076	-1.7	96.6	0.075

[a]Biais relatif (en %); [b]CP: probabilité de couverture empirique; [c]RMSE: Racine carrée de l'erreur quadratique moyenne

Tableau 12.4 – Comparaison *mexhaz*. Biais, probabilités de couverture et RMSE selon les quatre scénarios et les trois modèles proposés.

Le tableau 12.4 présente les résultats de la simulation en termes de biais, probabilités de couverture et RMSE.
Les principales différences entre *mexhaz* et *survPen* apparaissent dans l'estimation de la variance de l'effet aléatoire. Lorsque le nombre de clusters est limité (10 ou 20), les estimations proposées par *mexhaz* sont légèrement biaisées (-14% et -5,6%) tandis que *survPen* offre des biais toujours faibles quel que soit le nombre de clusters. Cette différence au niveau du biais impacte du même coup les probabilités de couverture : *mexhaz* est en sous-couverture notamment sur la première configuration tandis que *survPen* est toujours proche de la valeur cible à 95%.
Notons enfin que l'ajustement d'un modèle pénalisé ne dégrade pas les bonnes propriétés d'estimation de la variance (par rapport au modèle non pénalisé).

Configuration	Spline mixed (Charvat et al. 2016) réajusté avec mexhaz		survPen avec même modèle que Charvat et al. (2016)		survPen avec une spline pénalisée par sexe	
	Temps d'exécution médian[a]	Nombre de non-convergence[b]	Temps d'exécution médian[a]	Nombre de non-convergence[b]	Temps d'exécution médian[a]	Nombre de non-convergence[b]
10, 100	5.31	0	0.36	0	3.66	2
20, 50	5.35	2	0.80	0	6.32	1
50, 20	5.36	0	2.86	0	16.07	1
100, 10	5.40	1	10.61	0	43.53	1

[a]Temps d'exécution median pour adjuster un modèle (en secondes); [b]Parmi 1 000 modèles ajustés

Tableau 12.5 – Comparaison *mexhaz*. Temps de calcul médian pour ajuster un modèle selon les quatre scénarios et les trois modèles proposés.

Comme attendu, on remarque dans le tableau 12.5 que le temps de calcul via *survPen* augmente avec le nombre de clusters. Ainsi, *survPen* est plus rapide que *mexhaz* dans les trois premières configurations mais prend deux fois plus de temps dans la dernière configuration qui comprend 100 clusters. Les temps d'exécution de *mexhaz* restent constants autour de 5s.

Troisième partie

Applications épidémiologiques

Chapitre 13

Effet de la défavorisation sociale sur la mortalité en excès des patients atteints de cancer

13.1 Contexte et objectifs

Cette application a été effectuée dans le cadre d'un partenariat avec l'unité de recherche interdisciplinaire pour la prévention et le traitement des cancers ANTICIPE (U1086 INSERM-UNICAEN) et le réseau FRANCIM. La question était de savoir si la défavorisation sociale avait un effet sur la mortalité en excès subie par les patients atteints de cancer.
La défavorisation est mesurée au travers de la version française de l'*European Deprivation Index* (EDI, Guillaume et al. 2016).

Une première étape de ce travail a consisté en l'utilisation de l'estimateur non-paramétrique de Pohar-Perme pour mesurer l'impact de la défavorisation socio-économique sur la survie nette (Tron et al., 2019).
Suite à cette publication, la volonté de restituer les effets de la défavorisation de façon plus précise a eu pour conséquence la présente collaboration : celle-ci a pour objectif, en conservant l'EDI en continu, de décrire si l'effet est linéaire ou non-linéaire, proportionnel ou non-proportionnel, en interaction avec l'âge ou non.

Les résultats présentés ici sont des résultats partiels servant à illustrer l'intérêt de la méthode, seul le cancer du col de l'utérus étant analysé.

13.2 Données

La population d'étude compte 1865 patientes atteintes d'un cancer du col de l'utérus et diagnostiquées entre 2006 et 2009 (date de fin de suivi fixée au 30 Juin 2013). Les données proviennent de 14 registres de cancer du réseau FRANCIM. Dans la population d'étude, 689 patientes (37%) sont décédées dans les 5 ans suivant le diagnostic. Les taux de mortalité attendus (h_P) sont issus des taux de mortalité observés dans la population française par sexe, âge, année de décès et département de résidence (données issues de l'Institut National de la Statistique et des Études Économiques). Ces taux attendus ont en-

suite été lissés par le service de biostatistique des HCL grâce à un modèle de Poisson pénalisé incluant un produit tensoriel de l'âge et de l'année de diagnostic (pour chaque sexe et chaque département).

13.3 Méthode

Nous étudions l'effet du temps écoulé depuis le diagnostic (t), de l'âge au diagnostic (a) et de l'EDI sur le taux en excès. L'EDI est une variable continue dont la valeur augmente avec le degré de défavorisation. Nous disposons donc de trois variables continues et proposons d'ajuster le modèle suivant :

$$log\left\{h_E(t,a,EDI)\right\} = tensor(t,a,EDI)$$

Chaque base marginale est une spline cubique restreinte avec 5 nœuds (placés aux quantiles 0, 0,25, 0,50, 0,75 et 1 de l'ensemble des cas). Le modèle conduit donc à estimer 125 paramètres de régression. Les trois paramètres de lissage associés sont estimés via le critère LAML.

13.4 Résultats

La figure 13.1 présente différentes prédictions obtenues à partir du modèle. Dans cette figure, les trois âges et les trois valeurs d'EDI choisis correspondent aux percentiles 10, 50, et 90. Chaque colonne correspond à un âge donné.
La première ligne présente les surfaces de taux en excès en fonction du temps (en années) et de l'EDI.
La deuxième ligne montre le taux en excès en fonction du temps pour trois valeurs d'EDI.
La troisième ligne présente le ratio de taux en excès en fonction de l'EDI (avec pour référence EDI=0) à 4 temps différents (0, 1, 3, et 5 ans).
Enfin, la dernière ligne donne le ratio de taux en excès entre les percentiles d'EDI 90 et 10 en fonction du temps ainsi que les intervalles de confiance à 95%.

La mortalité en excès est plus importante chez les patientes les plus défavorisées (percentile 90 d'EDI vs. percentiles 10 et 50), excepté pendant la première année suivant le diagnostic chez les patientes âgées de 80 ans (deuxième ligne de la figure 13.1). De plus, le ratio de taux en excès entre les patientes les moins et les plus défavorisées est statistiquement différent de 1 aux âges 35 et 51 autour d'un an et demi après le diagnostic (voir les intervalles de confiance donnés dans les graphiques de la ligne 4). Ce ratio décroît ensuite en fonction du temps écoulé depuis le diagnostic. Le modèle ne détecte pas d'effet significatif de l'EDI à l'âge 80.

En ce qui concerne la survie nette, le modèle prédit à l'âge 51, une survie nette à 1 an de 93% [90.7-94.7%] pour le percentile 10 d'EDI et une survie nette de 90.2% [86.7-92.8%] pour le percentile 90. Les survies nettes à 5 ans correspondantes sont respectivement 74.5% [69.9-78.4%] et 61% [53.9-67.2%]. Ainsi, les différences observées dans la dynamique du taux en excès entre 1 et 2 ans après le diagnostic entraînent une différence significative de survie nette à 5 ans. Ces résultats illustrent la nécessité d'interpréter conjointement les estimations de survie avec celle de la dynamique du taux ; cette dernière permettant d'identifier où les différences apparaissent au cours du temps.

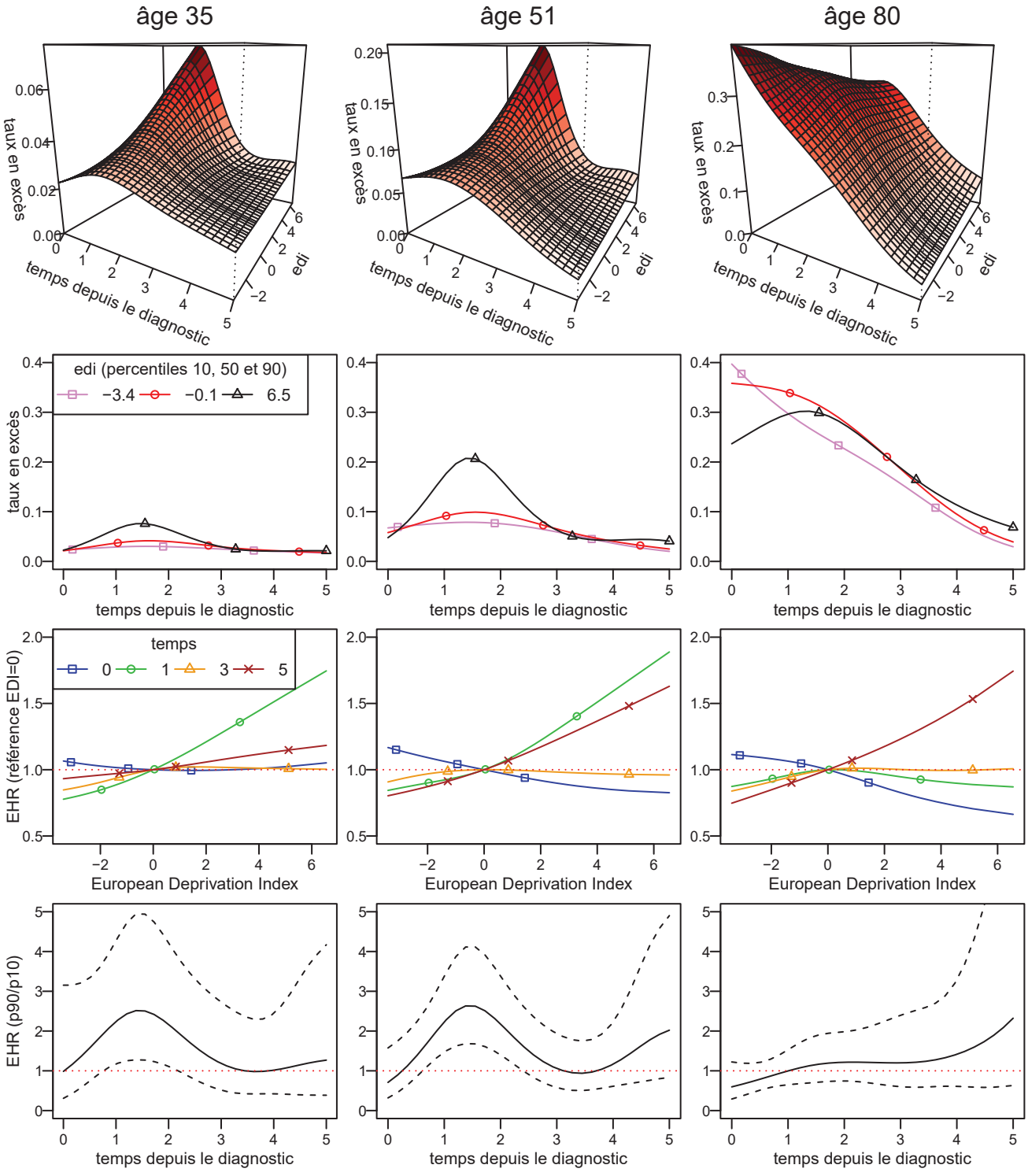

Figure 13.1 – Résultats principaux de l'étude sur l'EDI

13.5 Conclusion

Cette application illustre les aptitudes du modèle à capturer des phénomènes d'interactions complexes tout en assurant des changements lisses dans toutes les directions. Il faut bien entendu que les résultats soient ensuite interprétés sur le plan épidémiologique, et qu'ils puissent être à l'origine d'hypothèses sur les processus conduisant à cet excès de mortalité chez les femmes défavorisées.
Ainsi, le package *survPen* est actuellement utilisé sur la base FRANCIM complète. Il s'agit de la première fois que tous les registres sont analysés avec l'EDI (environ 210 000 tumeurs pour 18 registres) dans le cadre d'un partenariat FRANCIM, ANTICIPE, HCL. Ce travail fera l'objet de plusieurs publications scientifiques par groupes de cancers (digestifs, ORL, hématologiques, ...).

Chapitre 14

Étude de la mortalité en excès des patients atteints de sclérose en plaques

14.1 Contexte

La sclérose en plaques (SEP) est une maladie neurologique chronique caractérisée par une atteinte de la gaine protectrice des neurones : la myéline. La destruction de cette protection naturelle va entraîner des perturbations dans la transmission de l'influx nerveux. Les symptômes de la SEP sont variables d'une personne à l'autre, car ils dépendent de la zone où se produit l'attaque inflammatoire. Ils peuvent correspondre à des troubles de la sensibilité, à des paralysies ou faiblesses musculaires, des troubles de l'équilibre, et autres symptômes. En outre, avec le temps, la reconstruction de la myéline peut s'effectuer de manière imparfaite et la maladie peut ainsi devenir de plus en plus handicapante.

Lorsque de nouveaux symptômes apparaissent, on parle de « poussées ».

Trois modes d'évolution de la SEP existent :

- les formes récurrentes-rémittentes : les plus fréquentes, elles évoluent par poussées successives, entrecoupées de période de rémission.
- les formes secondairement progressives : elles font suite à une phase rémittente et se caractérisent par une progression régulière des symptômes.
- les formes progressives d'emblée : elles représentent entre 10% et 15% des cas et se manifestent par une progression continue des symptômes dès le début de la maladie. Les patients ne passent pas par la phase rémittente et ne connaissent donc pas de poussées.

Les patients atteints de SEP présentent un excès de mortalité par rapport à la population générale (Leray et al., 2015). Toutefois, l'évolution de cet excès de mortalité en fonction de la durée de la maladie (la dynamique du taux) est très mal connue et notons également que la modélisation flexible du taux de mortalité en excès n'a jamais été étudiée dans le cadre d'un modèle de taux additif (2.7). Ce sont ces raisons qui ont poussé à la constitution d'un partenariat entre le service de biostatistique des HCL, le Département d'épidémiologie et de biostatistiques (EPIBIOSTAT) de l'École des Hautes Études en Santé Publique (EHESP) et l'Observatoire Français de la Sclérose en Plaques (OFSEP).
En outre, il est important de mentionner la pertinence du concept de mortalité en excès dans l'étude de la mortalité due à la SEP. En effet, sachant que les patients atteints de SEP peuvent vivre très longtemps avec la maladie, plus le temps s'écoule depuis le diagnostic, plus il est difficile de dire si le décès d'un patient est dû, directement ou indirectement, à la SEP.
Enfin, les résultats présentés ici sont des résultats partiels, avec l'accord des partenaires des HCL, afin d'illustrer l'intérêt de la méthode.

14.2 Objectifs

À la suite des travaux de Leray et al. (2015), nous étudions la dynamique de la mortalité en excès chez les patients atteints de SEP ainsi que l'effet de l'âge à l'entrée dans la maladie. L'étude est stratifiée selon la forme initiale de la maladie.

14.3 Données

Les données utilisées étaient celles de 37 524 patients issus de 18 centres de l'OFSEP situés en France métropolitaine. Les patients sont majoritairement des femmes (à 71%) et ont été diagnostiqués entre 1960 et 2014. La date de fin de suivi est fixée au 1^{er} Janvier 2016 et les temps de suivis ont été censurés à 55 ans.
Pour un patient donné, le début de la maladie correspond à la date d'enregistrement systématique du centre OFSEP. Toutefois, les patients ne sont suivis qu'à partir de leur première évaluation clinique par un neurologue qui correspond donc à la date d'entrée dans l'étude. Ainsi, nous sommes en présence de données tronquées à gauche.
L'âge d'entrée médian est de 31,7 ans.
La forme rémittente de la maladie concerne 33 005 patients contre 4 519 patients pour la forme progressive (formes secondaire et d'emblée confondues).
Le nombre de décès observés est de 2 883 (2 142 chez les rémittents et 741 chez les progressifs), à rapporter aux 340 708 patients-années de suivi.

14.4 Méthode

Nous étudions l'effet du temps écoulé depuis le début de la maladie (t) et de l'âge au début de la maladie ou âge onset (a). Dans la suite, la notion d'âge courant correspond à $a + t$. Nous ajustons le modèle suivant chez les rémittents et les progressifs :

$$log\left\{h_E(t,a)\right\} = tensor(t,a)$$

Les nœuds utilisés sont :
- pour le temps : 0, 10, 20, 30, 40, 50
- pour l'âge : 10, 20, 30, 40, 50, 60, 70

Les deux modèles contiennent donc 42 paramètres de régression et 2 paramètres de lissage estimés par LAML.

14.5 Résultats

La figure 14.1 montre les prédictions de taux de mortalité en excès chez les rémittents et les progressifs. L'évolution du taux en excès est présentée en fonction du temps (première ligne) et en fonction de l'âge courant (deuxième ligne). Afin de mieux comprendre le parallèle entre les deux lignes, prenons un exemple. Si l'on considère la courbe de l'âge onset ($a = 30$) chez les rémittents sur la fenêtre en haut à gauche, on voit que celle-ci affiche un taux en excès de 0,01 décès par personne-année 30 ans après le début de la maladie ($t = 30$). Si l'on regarde maintenant la fenêtre en bas à gauche, nous

retrouvons logiquement cette valeur de 0,01 décès par personne-année lorsque l'âge courant vaut 60 ans ($a + t = 30 + 30 = 60$).

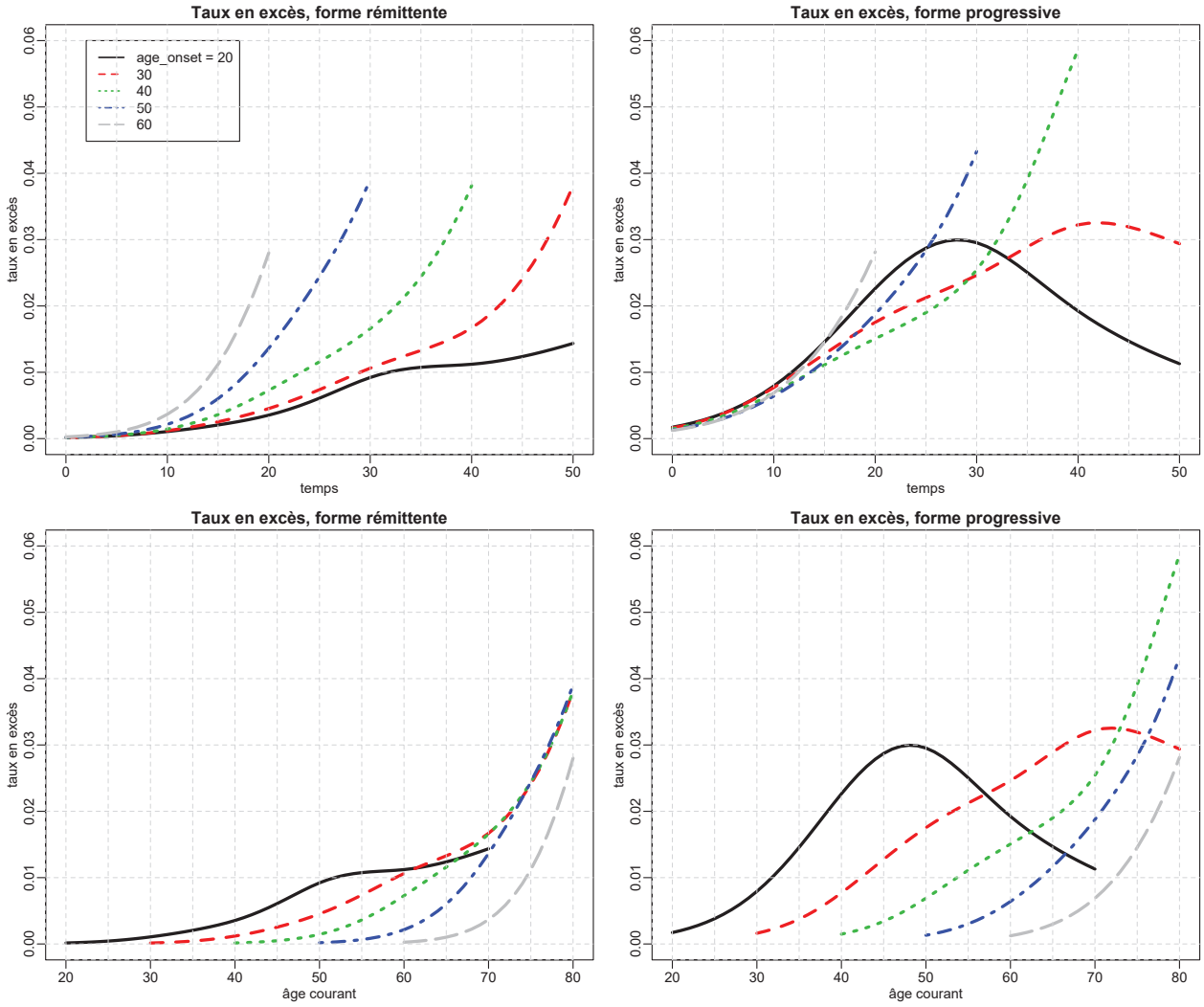

Figure 14.1 – Résultats principaux de l'étude sur la SEP

Chez les rémittents, l'excès de mortalité est très faible sur les 10 premières années de la maladie. Jusqu'à l'âge courant de 70 ans, les patients les plus jeunes au début de leur SEP présentent un taux en excès plus important que les autres. Cependant, à partir de 70 ans d'âge courant, les taux en excès sont comparables quel que soit l'âge de début de la maladie, les courbes étant parfaitement superposées (excepté pour l'âge 60). En d'autres termes, pour les rémittents, les taux de mortalité en excès ne dépendent pas de la durée de la SEP à partir de 70 ans d'âge courant.

Chez les progressifs, le taux en excès augmente rapidement dès le début de la SEP et dépend de la durée de la maladie quel que soit l'âge courant des patients. Chez les patients les plus jeunes (20 ans à l'entrée), on remarque une diminution du taux en excès après environ 30 ans de maladie (ou autour de 50 ans d'âge courant), observation qui reste à interpréter épidémiologiquement.

Pour juger de l'adéquation des modèles, les figures 14.2 et 14.3 comparent les prédictions des deux modèles au sein de différentes classes d'âges avec les prédictions de modèles de taux constants par intervalles. Les prédictions des deux modèles sont réalisées à l'âge médian (le taux populationnel donne

ici les mêmes résultats). Les barres verticales représentent l'étendue des intervalles de confiance pour les taux constants.

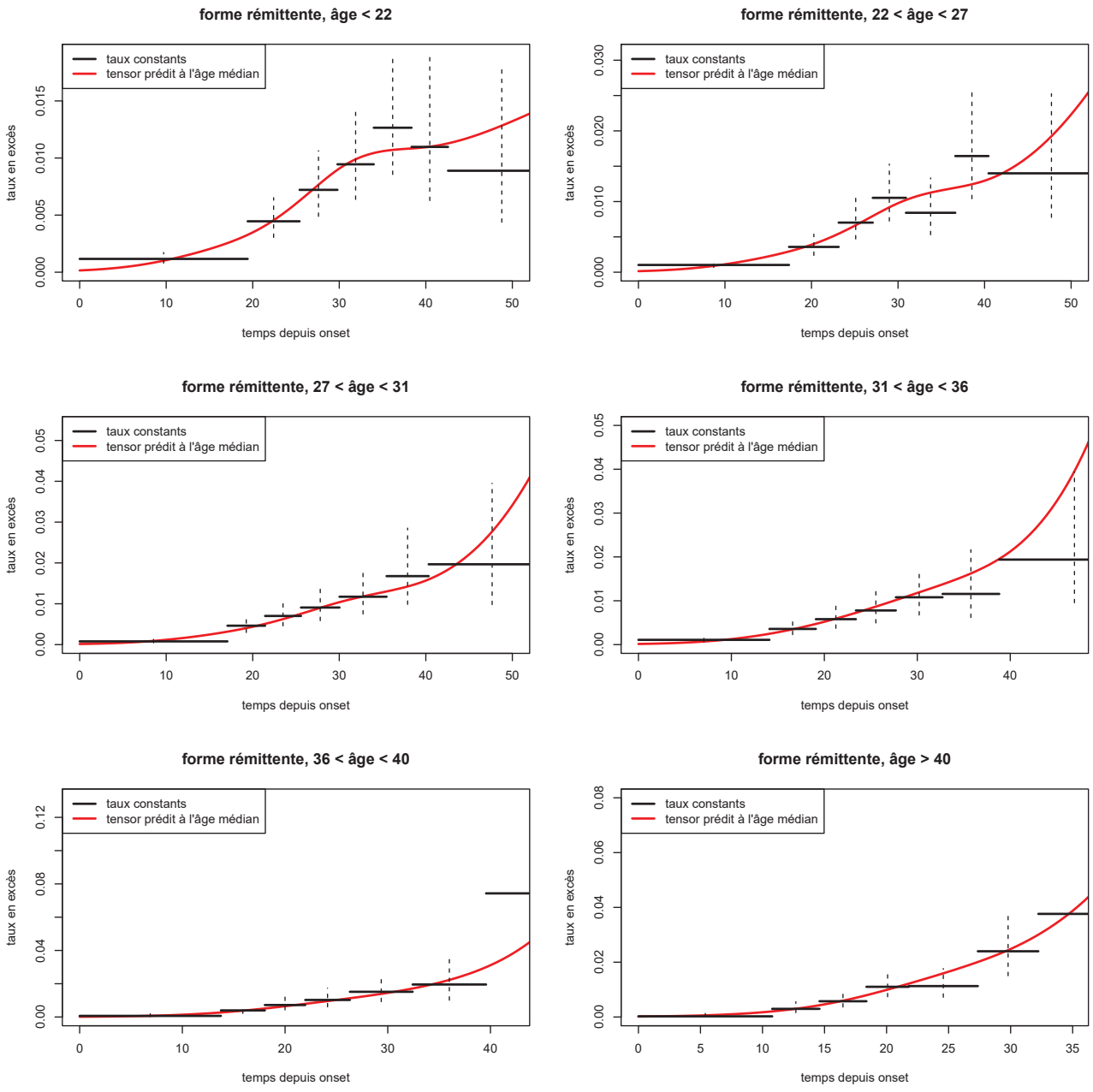

Figure 14.2 – Adéquation du modèle chez les rémittents

Les graphiques 14.2 et 14.3 montrent que la dynamique du taux en excès est restituée de manière satisfaisante au sein des différentes classes d'âges, que ce soit chez les rémittents ou chez les progressifs.

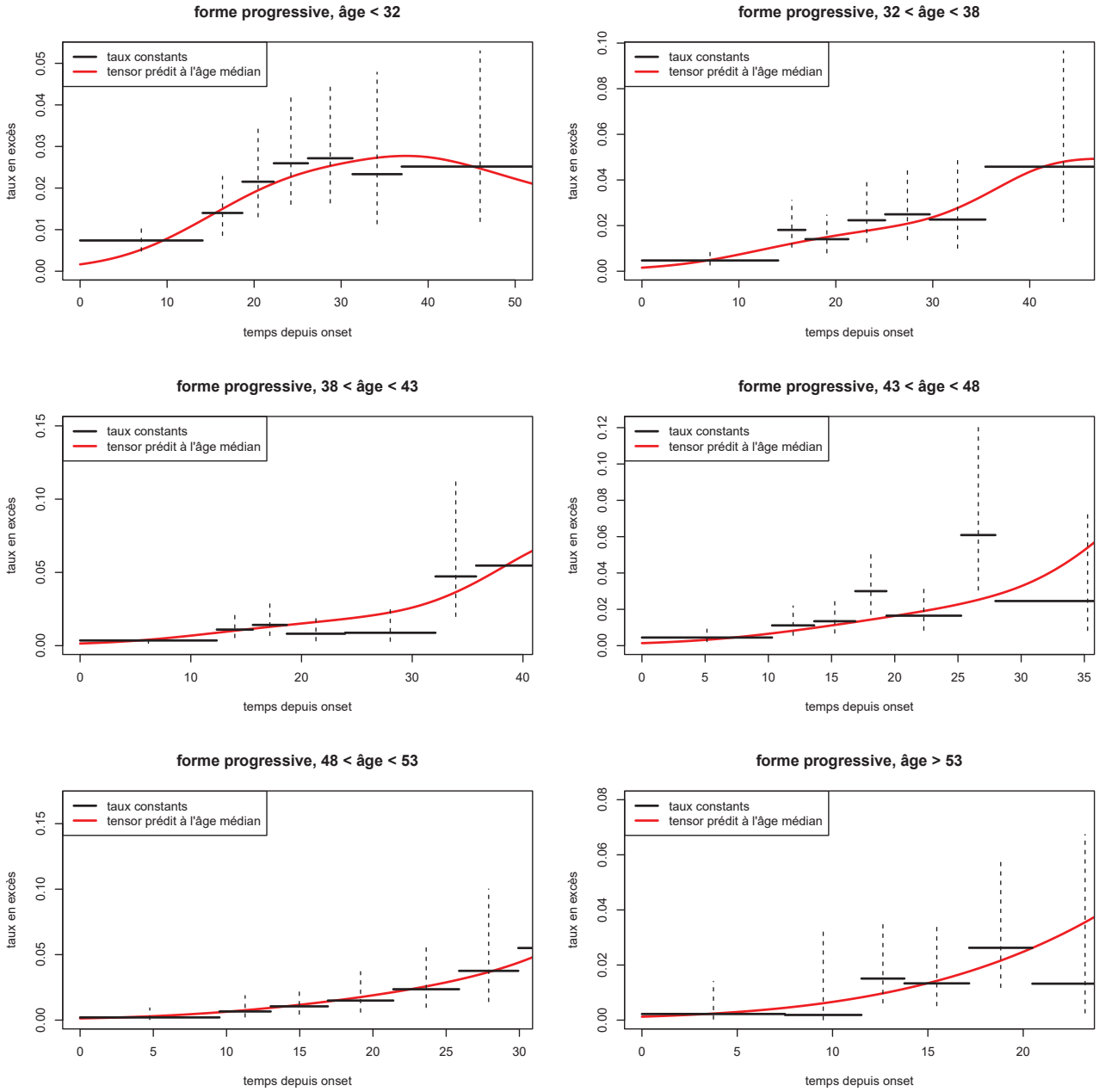

Figure 14.3 – Adéquation du modèle chez les progressifs

14.6 Conclusion

À partir d'un certain âge (autour de 70 ans), l'âge de début de la maladie n'impacte pas le taux de mortalité en excès chez les rémittents. Ce résultat, remarquable et épidémiologiquement important, est en cours d'interprétation par les épidémiologistes associés au projet, qui évoquent un processus amnésique dans la survie des patients atteints de SEP, similaire à celui précédemment observé dans l'évolution du handicap. Notons que ce résultat a pu être clairement mis en lumière grâce aux outils développés dans cette thèse et grâce à la cohorte OFSEP (cohorte unique en France représentant environ la moitié des cas de sclérose en plaques).

La publication associée à ce travail est en cours de rédaction : elle reprend les résultats exposés ici ainsi que l'analyse de l'effet du sexe et de l'année de début de la maladie. Elle fera aussi l'objet d'une présentation orale au 35^e congrès ECTRIMS (*European Committee for Treatment and Research in Multiple Sceloris*) en septembre 2019 à Stockholm.

Quatrième partie

Conclusion

Résumé de la contribution originale de la thèse

Le travail de thèse présenté dans ce manuscrit a consisté en différentes contributions originales :

- l'adaptation aux modèles de taux du cadre théorique et des algorithmes développés par Wood et al. (2016b)
- l'extension du cadre théorique aux modèles de taux en excès
- le calcul de la vraisemblance pénalisée et de ses dérivées première et seconde dans le cadre d'un modèle de taux et de taux en excès (avec notamment le recours à la quadrature de Gauss-Legendre)
- le calcul et l'implémentation du critère LCV et de ses dérivées première et seconde
- l'implémentation de la méthode au sein du package *survPen*
- l'évaluation des performances statistiques de la méthode via des simulations inspirées de données réelles
- l'application de la méthode à des analyses de données réelles dont l'étude de l'effet de la défavorisation sociale sur la mortalité en excès des personnes atteintes d'un cancer du col de l'utérus et l'étude de la mortalité en excès chez les patients atteints de sclérose en plaques.

Discussion

Dans cette thèse, nous avons développé un modèle de taux de survenue d'un événement capable de répondre aux problématiques épidémiologiques auxquelles nous étions confrontés. Ce travail, qui s'appuie sur des splines multidimensionnelles pénalisées dont le cadre théorique a été développé par Wood et al. (2016b), a fait l'objet d'une publication scientifique (Fauvernier et al., 2019b). La version publiée est disponible en section C de ce rapport. L'utilisation des splines assure la flexibilité du modèle tandis que la pénalisation a pour objectif de lisser les prédictions obtenues et ainsi de diminuer le risque de sur-ajustement.
La méthode développée a ensuite été implémentée dans un package R nommé *survPen* et appliquée sur des jeux de données réelles afin d'étudier les effets de facteurs pronostiques sur la mortalité en excès chez les patients atteints d'un cancer du col de l'utérus et chez ceux atteints d'une sclérose en plaques. Le package *survPen* a également fait l'objet d'une publication scientifique (Fauvernier et al., 2019a). La version publiée est disponible en section D de ce rapport.

Avantages

Le modèle de taux développé dans cette thèse permet de modéliser conjointement le taux de base ainsi que les effets non-proportionnels et non-linéaires des covariables. Le modèle gère aussi les interactions entre variables continues via un produit tensoriel pénalisé. Ce produit tensoriel peut être décomposé afin d'offrir un lissage différent selon les effets propres et les interactions. Ce genre de décomposition constitue une approche séduisante en analyse de survie car les effets propres sont souvent plus complexes que les interactions.

Dans la section 8.5 et le chapitre 10 de cette thèse, nous avons comparé les performances des critères LCV et LAML en termes d'estimation des paramètres de lissage. Au vu des résultats, le critère LAML semble un peu plus satisfaisant puisqu'il conduit à des prédictions moins variables que le critère LCV et apparaît donc moins sujet au sur-ajustement. En outre, le critère LAML demeure la seule manière de quantifier l'incertitude sur les paramètres de lissage estimés et de définir un critère AIC corrigé en ce sens. Notons toutefois que le critère LCV peut être utilisé avec n'importe quel type de pénalité (comme le LASSO) alors que le critère LAML est défini uniquement quand la pénalité est une forme quadratique des paramètres de régression.
Dans l'étude de simulation, la matrice de covariance Bayésienne offrait des probabilités de couverture proches de la valeur nominale de 95% sans tenir compte de l'incertitude sur les paramètres de lissage. Ce résultat avait déjà été mis en exergue par Wahba (1983) et Nychka (1988). Marra and Wood (2012) ont montré que ce résultat est attendu dès lors que les paramètres de lissage ne sont pas estimés à des valeurs trop importantes par manque d'information.
Bien que l'étude de simulation présente dans cette thèse ne s'intéresse pas aux petits effectifs, Remontet et al. (2019) ont démontré les bonnes propriétés des tensors pénalisés pour des effectifs allant de 250 à 1000 individus. Les tensors pénalisés présentent notamment un excellent compromis biais-variance,

et ce même pour des petits effectifs. En revanche, en prenant l'exemple de l'analyse de tendance pour le cancer de l'ovaire, Remontet et al. (2019) ont montré qu'un certain niveau d'information (au moins 1000 individus) était nécessaire afin d'obtenir des estimations d'une précision satisfaisante.

Le chapitre 12 démontre la supériorité de *survPen* par rapport à différentes approches en termes de temps d'exécution.
Ensuite, le chapitre 12 compare l'implémentation de *survPen* à celle de *rstpm2* dans le cadre des produits tensoriels pénalisés dans les modèles de taux en excès (cette comparaison correspond au *Supplementary Material* C.2 de Fauvernier et al. 2019b). Dans ce cadre, la supériorité de *survPen* par rapport à *rstpm2* en termes de performances statistiques et de stabilité a été démontrée.
Deux caractéristiques majeures peuvent expliquer la supériorité de *survPen* par rapport à *rstpm2* : i) *survPen* utilise les dérivées explicites de la vraisemblance tandis que *rstpm2* a recours à des dérivées numériques via les fonctions R *optim* et *nlm* ; et, ii) dans *survPen*, les paramètres sont définis sur l'échelle du log du taux (échelle considérée comme naturelle) alors que, dans *rstpm2*, cette échelle n'est pas disponible ce qui implique l'utilisation d'un système de contraintes qui ajoute de la complexité aux algorithmes d'optimisation.

Enfin, le chapitre 12 compare l'implémentation de *survPen* à celle de *mexhaz* dans le cadre des modèles de taux en excès pénalisés à effets mixtes. Dans ce cadre, *survPen* est supérieur en termes de performances statistiques mais perd l'avantage en termes de temps de calcul lorsque le nombre de clusters augmente.
En plus des comparaisons réalisées dans cette thèse, le *Supplementary Material* C.1 de Fauvernier et al. (2019b) présente des comparaisons entre les packages R *frailtypack*, *gss*, *R2BayesX*, *rstpm2*, *mgcv*, *bamlss*, et *survPen* en termes de temps d'exécution. Ces comparaisons démontrent l'efficience de *survPen*.

La troisième partie de cette thèse s'est intéressée à la résolution de problématiques épidémiologiques à l'aide du modèle développé. La première étude concernait l'impact de la défavorisation sociale (via l'EDI) sur la mortalité en excès de patientes atteintes de cancer du col de l'utérus. Le modèle consistait en un produit tensoriel pénalisé du temps, de l'âge au diagnostic et de l'EDI. Cette modélisation flexible a permis de montrer que, chez les femmes jeunes et d'âge moyen, le taux de mortalité en excès des patientes les plus défavorisées était significativement supérieur à celui des patientes les moins défavorisées. Cet écart, principalement observé entre un et deux ans après le diagnostic, impliquait des différences non négligeables de survie nette à cinq ans.
La seconde application de cette thèse concernait la mortalité en excès chez les patients atteints de SEP. Nous avons alors notamment montré que, chez les rémittents, l'âge courant a un impact plus important sur la mortalité due à la SEP que la durée de la maladie.

Même si la réalisation de cette thèse a été motivée par des considérations épidémiologiques précises, les méthodes statistiques développées ont une portée bien plus importante. En effet, de nombreuses études de temps d'événement peuvent bénéficier des splines multidimensionnelles pénalisées car, en général, il est difficile de spécifier un modèle paramétrique.

Limites

Puisque les bases marginales utilisées sont des splines cubiques restreintes, le nombre et la position des nœuds doivent être spécifiés en amont. Toutefois, cette restriction est bien moins forte que dans un contexte non pénalisé (Wood, 2017, Section 4.2.2). En pratique, il est raisonnable de spécifier un nombre de noeuds sensiblement plus important que ce que l'on pense nécessaire, la pénalisation se chargeant ensuite de lisser les prédictions et d'éviter le sur-ajustement.
En règle générale, le produit tensoriel atteint ses limites lorsque le nombre de covariables est supérieur ou égal à quatre. Dans ce cas, des stratégies de sélection de variables sont nécessaires. Dans le cadre des GAM, de telles stratégies sont proposées par Marra and Wood (2011). L'AIC corrigé, initialement proposé par Wood et al. (2016b), est une autre approche intéressante pour la sélection de modèles de taux pénalisés.

Le calcul du log déterminant de la matrice de pénalisation et de ses dérivées est un point clé de la stabilité de l'estimation par LAML. En effet, lorsque certains paramètres de lissage tendent vers l'infini tandis que d'autres sont très faibles, les calculs peuvent devenir instables. La résolution proposée ici diffère de celle initialement proposée par Wood et al. (2016b) et est théoriquement plus instable car elle repose sur le calcul explicite de l'inverse de la matrice de pénalisation. Toutefois, en pratique, sa stabilité est amplement satisfaisante.

La contrepartie de la méthode proposée ici est la difficile généralisation du calcul des dérivées à d'autres contextes. Par exemple, la prise en compte d'effets aléatoires telle que décrite dans la section 8.7 devient inefficiente lorsque le nombre de clusters augmente (Wood, 2011). Ainsi, il serait intéressant d'étendre les splines pénalisées à un véritable modèle mixte qui maximise la vraisemblance marginale pénalisée. Cependant, gérer l'intégration sur la distribution des effets aléatoires n'est pas chose aisée et représente un vrai défi.

Perspectives

Une des premières perspectives intéressantes de ce travail serait d'intégrer à la modélisation les variables dépendantes du temps, notamment afin de pouvoir mesurer l'impact sur le taux en excès d'une exposition cumulée à un facteur de risque. Pour l'heure, ce type de modélisation n'est disponible que dans le cadre des GAM et dans les modèles de survie s'appuyant sur les GAM (Gasparrini et al., 2017; Bender et al., 2018a,b).

L'information collectée dans les registres est toujours plus riche d'année en année et le nombre de facteurs pronostiques susceptibles d'intéresser les cliniciens et les épidémiologistes augmente en conséquence. Le développement de méthodes de sélection de variables et de choix de modèles dans le cadre des modèles de taux en excès pénalisés est ainsi de plus en plus prégnant. À ce titre, les techniques de *boosting*, issues de l'apprentissage machine ou *machine learning*, et utilisées notamment par Kneib et al. (2009) dans les *geoadditive models*, constituent une piste prometteuse. Assez récemment, Hofner et al. (2013) ont d'ailleurs étendu ce principe aux modèles de survie. D'autres pistes ont été développées dans le cadre des GAM et méritent également notre attention sur le sujet (Marra and Wood, 2011).

Puisque l'information à analyser est toujours plus volumineuse, les algorithmes utilisés doivent s'adapter. Afin de limiter le temps de calcul et l'utilisation de la mémoire vive, Wood et al. (2015) ont proposé

d'utiliser une décomposition QR de la matrice de design. Plus récemment encore, Wood et al. (2017) ont mis au point un système de discrétisation des valeurs de covariables : par exemple, si l'on collecte l'âge d'individus âgés de 21 à 80 ans, alors il existe 60 valeurs d'âges à l'âge entier près, 600 valeurs d'âges au dixième d'âge près etc. Cette approche est extrêmement intéressante car quelle que soit la précision nécessaire à l'étude statistique (l'âge au dixième près serait typiquement amplement suffisant dans la majorité des cas), le nombre de valeurs distinctes sera toujours très inférieur au nombre d'individus. Cette technique, perfectionnée par Li and Wood (2019), pourrait ainsi être intégrée dans l'implémentation de *survPen*.

Conclusion

La méthode proposée ainsi que son implémentation au sein du package *survPen* constituent un outil opérationnel, validé sur le plan théorique (par simulation) et sur le plan pratique grâce à son utilisation dans deux études épidémiologiques majeures. Il faut également noter qu'actuellement, l'étude sur la survie des patients atteints de cancer du réseau FRANCIM, analysant plus de 1,2 millions de tumeurs diagnostiquées entre 1989 et 2015, utilise la méthodologie développée dans cette thèse.

Nous espérons que, dans le futur, notre outil continuera à montrer son utilité en épidémiologie mais aussi en recherche clinique.

Annexes

A Intégration numérique

On s'intéresse au calcul de l'intégrale I d'une fonction f définie sur un intervalle borné $[a;b]$.

$$I = \int_a^b f(x)\mathrm{d}x$$

A.1 Cavalieri-Simpson

On définit les points $x_0 = a$, $x_1 = \frac{a+b}{2}$ et $x_2 = b$. On approxime la fonction f par le polynôme P de degré 2 qui passe par les points $f(x_0) = f_0$, $f(x_1) = f_1$ et $f(x_2) = f_2$.

$$P(x) = 2\frac{f_2 - 2f_1 + f_0}{(x_2 - x_0)^2}(x - x_1)^2 + \frac{f_2 - f_0}{x_2 - x_0}(x - x_1) + f_1$$

On approche alors I par :

$$\tilde{I} = \int_a^b P(x)\mathrm{d}x = (b-a)\frac{f_0 + 4f_1 + f_2}{6}$$

A.2 Gauss-Legendre

La méthode de Cavalieri-Simpson fait partie des méthodes plus générales dites de Newton-Cotes. Ces méthodes consistent à approximer I par $\tilde{I} = (b-a)\sum_{k=0}^{n}\omega_k f_k$. Les $n+1$ points sont **espacés de manière régulière** et on approxime f par le polynôme de degré n qui coïncide avec f sur ces $n+1$ points.
L'idée de la quadrature de Gauss-Legendre est de généraliser les méthodes de Newton-Cotes en évaluant la fonction f en des points **espacés de manière irrégulière** sur l'intervalle d'intégration. Si on ne fixe pas a priori les positions des $n+1$ points, cela laisse $n+1$ degrés de libertés supplémentaires et on peut choisir les positions de ces points de manière à obtenir une méthode optimale.
Afin de trouver l'expression des méthodes de Gauss pour différents nombres de points, on cherche à déterminer les positions des x_k et les coefficients ω_k associés de manière à ce que cette méthode soit exacte pour les polynômes 1, x, x^2, ... et x^{2n+1}. Le plus souvent, la méthode de Gauss-Legendre est présentée sur l'intervalle $[-1;1]$. Une intégration sur l'intervalle $[a;b]$ s'obtient par le changement de variable : $x \to \frac{a+b}{2} + \frac{b-a}{2}x$.

Exemple pour 1 point ($n = 0$) :

Nous voulons que $\tilde{I}$ soit égale à I pour les polynômes $P_0 = 1$ et $P_1 = x$. En règle générale, on intègre sur $[a;b] = [-1;1]$.

D'un côté nous avons :

$$\tilde{I}_0 = (b-a)\sum_{k=0}^{n}\omega_k f_k = 2\sum_{k=0}^{0}\omega_k P_0(x_k) = 2\sum_{k=0}^{0}\omega_k = 2\omega_0$$

Qui doit être égale à :

$$I_0 = \int_{-1}^{1} 1\mathrm{d}x = 2$$

Ainsi, on en déduit que $\omega_0 = 1$.

De même, on a :

$$\tilde{I}_1 = 2\sum_{k=0}^{0} \omega_k P_1(x_k) = 2\sum_{k=0}^{0} \omega_k x_k = 2\omega_0 x_0$$

Et :

$$I_1 = \int_{-1}^{1} x \mathrm{d}x = 0$$

Donc $x_0 = 0$.

Finalement, on trouve :

$$\tilde{I} = 2f(0)$$

On peut appliquer la même méthode pour trouver les coefficients ω_k et les points x_k de la méthode de Gauss-Legendre à 2 points, 3 points ...

Pour un même nombre de points évalués, la méthode de Gauss-legendre est bien plus optimale que les méthodes de Newton-Cotes.

A.3 Calibration du nombre de nœuds pour la quadrature de Gauss-Legendre

Dans les modèles de taux présentés dans cette thèse, nous utilisons la quadrature de Gauss-Legendre afin d'approximer le taux cumulé. L'utilisation de cette quadrature implique de spécifier en amont un nombre de points (nœuds) d'évaluation q. Si ce nombre est trop petit, on risque de manquer de précision, et s'il est trop grand, le temps de calcul sera trop important pour être acceptable.

Afin de choisir q, nous réalisons une étude de sensibilité à partir des données et des modèles suivants :

Jeu de données : 5977 patientes diagnostiquées d'un cancer du col de l'utérus en France entre 1989 et 2011 (données FRANCIM)

Modèles à ajuster :

$$\text{modèle univarié} \quad : \quad log\{h(t)\} = f(t)$$

$$\text{modèle trivarié} \quad : \quad log\{h(t,a,y)\} = f(t) \times g(a) \times h(y)$$

où t est le temps écoulé depuis le diagnostic, a est l'âge au diagnostic, y est l'année de diagnostic et f, g et h sont des splines cubiques naturelles avec respectivement 5, 4 et 3 nœuds. Le modèle trivarié est assimilable à un tensor non pénalisé.

On s'intéresse à la survie jusqu'à 10 ans après le diagnostic.

Idée : chaque modèle est ajusté quatre fois en retenant respectivement 10, 15, 20 et 50 nœuds pour la quadrature de Gauss-Legendre. Ensuite, nous comparons les survies et taux prédits par les quatre versions du modèle en considérant la version à 50 nœuds comme référence.

Les résultats sont donnés dans les figures A.1, A.2 et A.3.

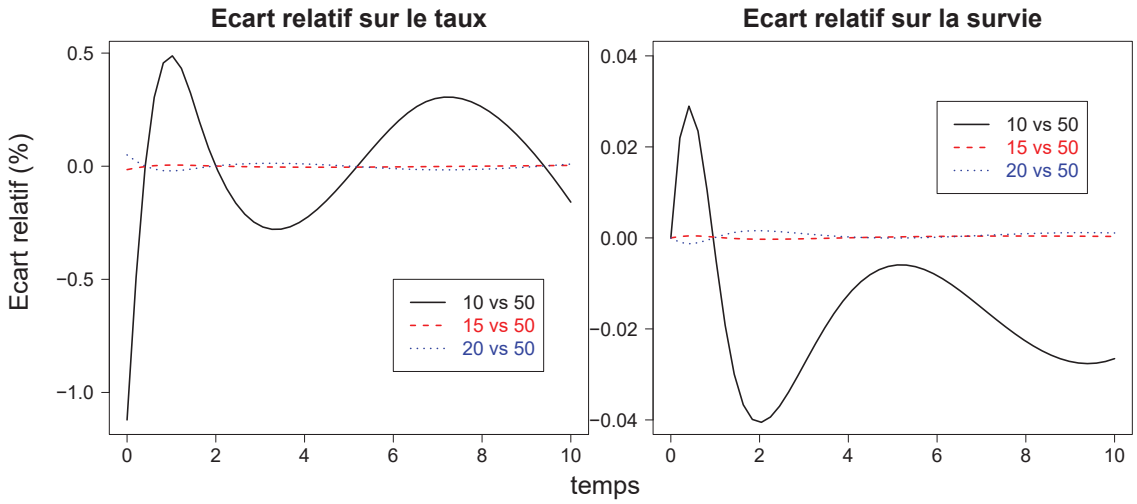

Figure A.1 – Calibration Gauss-Legendre : écarts sur les prédictions du modèle univarié

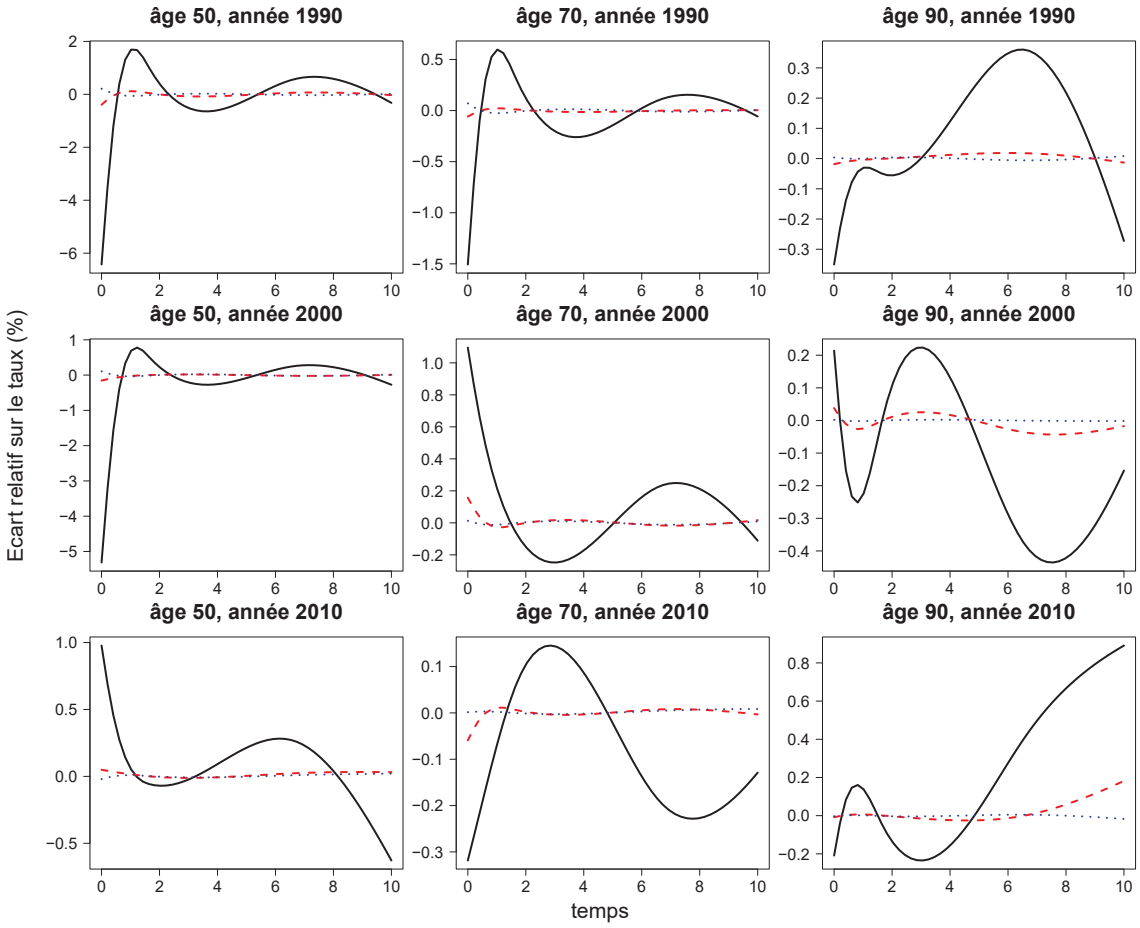

Figure A.2 – Calibration Gauss-Legendre : écarts sur les prédictions de taux du modèle trivarié

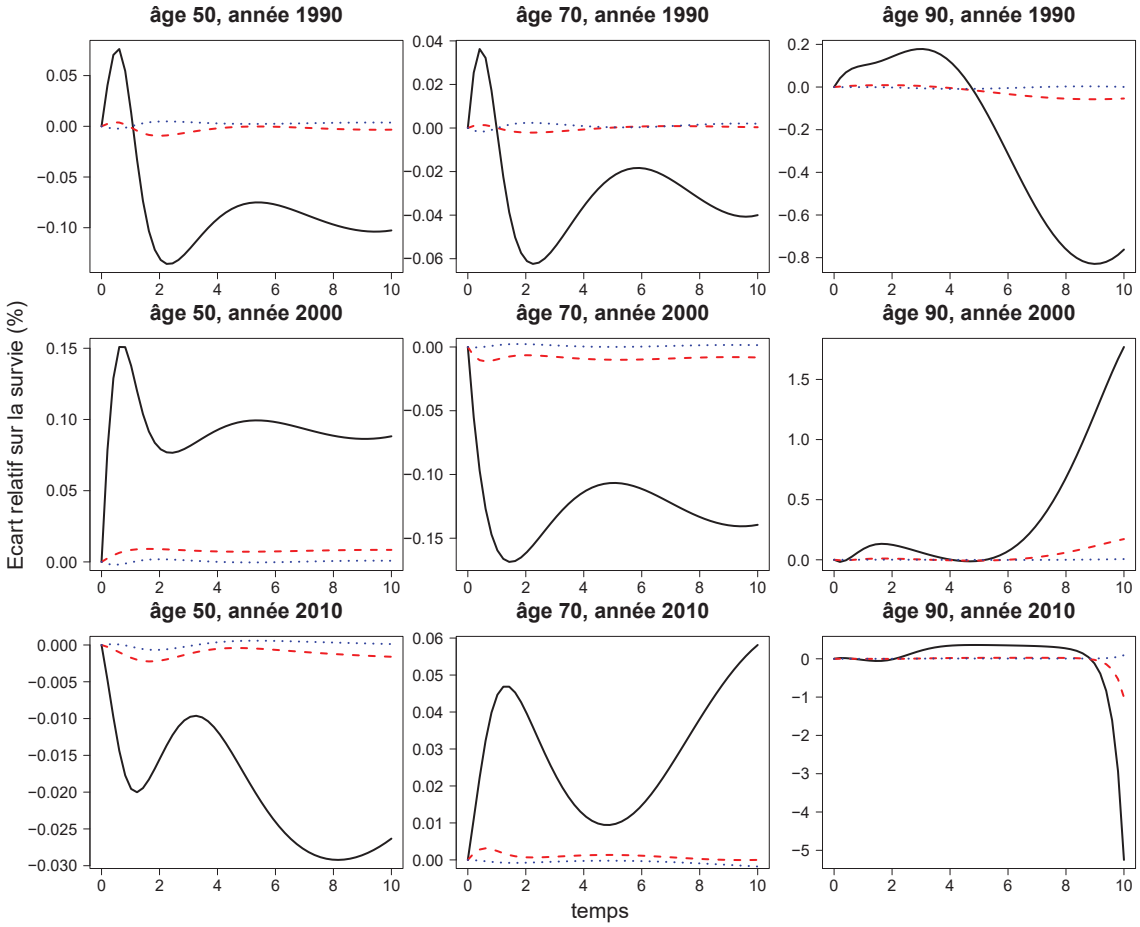

Figure A.3 – Calibration Gauss-Legendre : écarts sur les prédictions de survie du modèle trivarié

Conclusion : La valeur $q = 10$ présente une variabilité très importante par rapport à la référence $q = 50$. Cette variabilité est nettement diminuée lorsque l'on passe à $q = 15$ mais certains effets « de bords » persistent. Au final, la valeur $q = 20$ nous semble être un bon compromis en termes de précision et de temps de calcul raisonnable.

Valorisation scientifique

B Liste des articles et des communications scientifiques

Articles publiés

1. **Fauvernier**, M., Roche, L., Uhry, Z., Tron, L., Bossard, N. and Remontet, L. (2019). Multi-dimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival, *Journal of the Royal Statistical Society, series C*. doi: 10.1111/rssc.12368

2. **Fauvernier**, M., Remontet, L., Uhry, Z., Bossard, N. and Roche, L. (2019). survPen: an R package for hazard and excess hazard modelling with multidimensional penalized splines, *Journal of Open Source Software*, 4(40), 1434. doi:10.21105/joss.01434

3. Tron, L., Belot, A., **Fauvernier**, M., Remontet, L., Bossard, N., Launay, L., Bryere, J., Monnereau, A., Dejardin, O., Launoy, G. and the French Network of Cancer Registries (FRANCIM) (2019), Socioeconomic environment and disparities in cancer survival for 19 solid tumor sites: An analysis of the French Network of Cancer Registries (FRANCIM) data. *Int. J. Cancer*, 144: 1262-1274. doi:10.1002/ijc.31951

4. Delinière, A., Baranchuk, A., Giai, J., Bessiere, F., Maucort-Boulch, D., Defaye, P., Marijon, E., Le Vavasseur, O., Dobreanu, D., Scridon, A., Da Costa, A., Delacr´etaz, E., Kouakam, C., Eschalier, R., Extramiana, F., Leenhardt, A., Burri, H., Winum, P. F., Taieb, J., Bouet, J., **Fauvernier**, M., Rosianu, H., Carabelli, A., Duband, B., and Chevalier, P. (2019). Prediction of ventricular arrhythmias in patients with a spontaneous Brugada type 1 pattern: the key is in the electrocardiogram. EP Europace. https://doi.org/10.1093/europace/euz156

Communications orales

1. **Mathieu Fauvernier**, Laurent Roche, Zoé Uhry, Laure Tron, Nadine Bossard, Laurent Remontet and the CENSUR Working Survival Group. « Multidimensional penalized splines in survival models: applications to cancer epidemiology ». Group for Cancer Epidemiology and Registration in Latin Language Countries (GRELL) Meeting 2019, du 29 au 31 Mai 2019, Lisbonne (Portugal).

2. **Mathieu Fauvernier**, Laurent Roche, Zoé Uhry, Laure Tron, Nadine Bossard, Laurent Remontet and the CENSUR Working Survival Group. « Splines multidimensionnelles pénalisées dans les modèles de survie : applications en épidémiologie des cancers ». 13ème Conférence Francophone d'Epidémiologie Clinique (EPICLIN) et 26èmes Journées des statisticiens des Centre de Lutte Contre le Cancer (CLCC), du 15 au 17 Mai 2019, Toulouse (France).

3. **Mathieu Fauvernier**, Laurent Roche, Zoé Uhry, Laure Tron, Nadine Bossard, Laurent Remontet and the CENSUR Working Survival Group. « Multidimensional penalised splines for hazard and excess hazard regression ». International Conference of the Royal Statistical Society 2018 (RSS 2018), du 3 au 6 septembre 2018, Cardiff (Royaume-Uni).

4. **Mathieu Fauvernier**, Laurent Roche, Zoé Uhry, Laure Tron, Nadine Bossard, Laurent Remontet and the CENSUR Working Survival Group. « Multidimensional penalized splines in hazard regression for time-to-event data ». Joint International Society for Clinical Biostatistics and Australian Statistical Conference 2018 (ISCB ASC 2018), du 27 au 30 août 2018, Melbourne (Australie).

Poster

Mathieu Fauvernier, Laurent Remontet, Zoé Uhry, Laure Tron, Nadine Bossard, Laurent Roche and the CENSUR Working Survival Group. « survPen: multidimensional penalized splines for (excess) hazard regression in R ». 40th Annual Conference of the International Society for Clinical Biostatistics (ISCB 2019), du 14 au 18 Juillet 2019, Louvain (Belgique).

C Article méthodologique publié dans JRSSC

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2019)

Multi-dimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival

Mathieu Fauvernier and Laurent Roche,

Hospices Civils de Lyon and Université Lyon 1, France

Zoé Uhry,

Santé Publique France, Saint Maurice, Hospices Civils de Lyon and Université Lyon 1, France

Laure Tron,

Centre Hospitalier Universitaire de Caen, and Université de Caen Normandie, Caen, France

Nadine Bossard and Laurent Remontet

Hospices Civils de Lyon and Université Lyon 1, France

and the Challenges in the Estimation of Net Survival Working Survival Group

Marseille, France

[Received September 2018. Revised June 2019]

Summary. Describing the dynamics of patient mortality hazard is a major concern for cancer epidemiologists. In addition to time and age, other continuous covariates have often to be included in the model. For example, survival trend analyses and socio-economic studies deal respectively with the year of diagnosis and a deprivation index. Taking advantage of a recent theoretical framework for general smooth models, the paper proposes a penalized approach to hazard and excess hazard models in time-to-event analyses. The baseline hazard and the functional forms of the covariates were specified by using penalized natural cubic regression splines with associated quadratic penalties. Interactions between continuous covariates and time-dependent effects were dealt with by forming a tensor product smooth. The smoothing parameters were estimated by optimizing either the Laplace approximate marginal likelihood criterion or the likelihood cross-validation criterion. The regression parameters were estimated by direct maximization of the penalized likelihood of the survival model, which avoids data augmentation and the Poisson likelihood approach. The implementation proposed was evaluated on simulations and applied to real data. It was found to be numerically stable, efficient and useful for choosing the appropriate degree of complexity in overall survival and net survival contexts; moreover, it simplified the model building process.

Keywords: Excess hazard; Net survival; Non-linear effects; Penalized regression splines; Smooth models; Time-dependent effects

Address for correspondence: Mathieu Fauvernier, Service de Biostatistique—Bioinformatique, Centre Hospitalier Lyon Sud, 165 Chemin du Grand Revoyet, F-69130 Pierre-Benite, Lyon, France.
E-mail: mathieu.fauvernier@chu-lyon.fr

 0035–9254/19/68000

1. Introduction

Patient survival and its corresponding mortality hazard represent essential indicators in cancer epidemiology. Dedicated population-based studies aim at producing these indicators by cancer site and sex; they rely on data that are collected by cancer registries in well-defined geographical areas (for example, see Allemani *et al.* (2018) on the international study CONCORD-3). Two fundamental continuous variables impact cancer mortality hazard and must be accounted for in the analysis (Remontet *et al.*, 2018): the time elapsed since diagnosis and the age at cancer diagnosis. In addition to these two variables, the effect of other continuous covariates may be of interest. For example, trend analyses (that study the effect of the year of diagnosis on the mortality hazard) are essential to describe the way that medical practices impact patient survival. There is also an increasing interest nowadays in socio-economic inequality studies that aim at assessing the effect of a continuous social deprivation index on mortality. The problem here consists in specifying a methodology that enables a full and accurate description of the dynamics of mortality hazard and the corresponding survival according to age and other factors of interest. Besides, this methodology should be well suited for the routine production of epidemiological indicators, i.e. applicable to a wide variety of types of cancer. As a motivating example, survival data from 537291 cancer patients who were diagnosed between 1989 and 2010 collected by the French Network of Cancer Registries were available for studying net survival trends by sex for 53 cancers (Cowppli-Bony *et al.*, 2017; Monnereau *et al.*, 2016). The continuous European deprivation index (EDI) (Guillaume *et al.*, 2016) was also collected by the French Network of Cancer Registries, for 189144 patients with cancer (Bryere *et al.*, 2018) and, similarly, the effect of the EDI on mortality hazard by cancer site and sex was studied.

Consequently, in survival analysis, in addition to modelling the effect of time (via the baseline hazard), one has often to deal with several continuous covariates and to model their functional forms, their time-dependent effects and their interactions. As studying the joint effect of several continuous covariates is complex, most published works have been based on the categorization of one or several continuous covariates and stratified analyses. For example, to compare cancer survival trends between countries, the CONCORD-3 study (Allemani *et al.*, 2018) categorized the year of diagnosis and calculated the net survival by using a non-parametric estimator (Perme *et al.*, 2012) within each period of diagnosis whereas Antunes *et al.* (2016) used a categorized version of the EDI to study its effect on survival.

Nevertheless, it is well known that categorization leads to a loss of information and that multiplying stratified analyses, especially when using a non-parametric method, may decrease dramatically the precision of the estimates. Thus, accurate estimates of complex effects as well as interactions can only come from a modelling approach where the covariates are kept in their continuous forms; this can be done in the framework of fully parametric flexible hazard models (Remontet *et al.*, 2007; Charvat *et al.*, 2016). These models are attractive because they allow efficient inference via full likelihood and provide a full description of the dynamics of the hazard.

Specifying a flexible parametric model is complex because strong choices must be made regarding non-linearity, time dependence and interactions. Model selection procedures represent a possible solution to this model specification problem: a list of candidate models (which differ in terms of non-linearity, non-proportionality and interactions) is established and the 'best' model is retained by using some selection criterion, such as the Akaike information criterion (AIC). In recent work (Uhry *et al.*, 2017), the best model included a triple interaction between time since diagnosis, age and year of diagnosis in about half of 90 analyses. This emphasizes the interest of modelling higher order interactions. Nevertheless, specifying the pool of candidate models is very challenging and there is no certainty that the 'right' model is part of this pool.

Moreover, model selection uncertainty is then difficult to take into account and therefore often disregarded.

Penalized regression splines (Ruppert *et al.*, 2003; Wood, 2017) are an appealing alternative approach to model selection. In comparison with the unpenalized setting, model specification is made easier because of the reduced need for a model selection procedure. Indeed, in a penalized framework, choosing the appropriate degree of complexity relies largely on the estimates of the smoothing parameters. Important concerns are then the choice of the criterion to use in estimating the smoothing parameters and the way of inserting it into the optimization scheme.

Different penalized models for time-to-event data have already been proposed. Rondeau *et al.* (2003) proposed penalized estimation in hazard regression models in which the baseline hazard was modelled as a penalized spline. Liu *et al.* (2018) have recently published a penalized framework that enables modelling on different scales defined as functions of survival (typically, the log-cumulative hazard) but not the appealing log-hazard scale because of the integration problem that is raised in this case. This problem is avoided and becomes a derivation issue but, then, constraints are needed to ensure a positive associated hazard; this adds complexity to an already difficult optimization problem. To benefit from the generalize additive model (GAM) framework, Kauermann (2005), Becher *et al.* (2009) and Rodríguez-Girondo *et al.* (2013) presented a penalized model for hazard regression where the estimation procedure is based on data augmentation and a Poisson approach and relies on the classical relationship between survival likelihood and Poisson likelihood (Friedman, 1982). In a mixed-model-based approach, Kneib and Fahrmeir (2007) developed a penalized framework for Cox-type hazard models that allow for time varying coefficients and non-linear effects. However, this framework does not handle interactions between more than two continuous variables. Moreover, interactions between time and another covariate are not currently available in the implementation of the *BayesX* software (version 3.0.2 (Brezger *et al.*, 2003), that includes the approach of Kneib and Fahrmeir) or in the *R2BayesX* package (version 1.1-1 (Umlauf *et al.*, 2015)).

Fully Bayesian frameworks have also been proposed as in Hennerfeind *et al.* (2006), who introduced geoadditive survival models in which the regression and smoothing parameters are estimated simultaneously by using Markov chain Monte Carlo (MCMC) simulation. Martino *et al.* (2011) proposed a Bayesian survival model in which the computationally intensive MCMC inference is replaced by an integrated nested Laplace approximation. However, the latter model considers only either a proportional hazard model through Weibull regression or a piecewise log-constant baseline hazard. Recently, Umlauf *et al.* (2018) proposed the *bamlss* R package to fit complex additive structures on the log-hazard scale, including tensor product splines of time and covariates. Despite its attractiveness, this implementation relies on MCMC inference that requires a long computation time.

As cancer patients may die from cancer or from other causes, it is relevant to study the mortality due to cancer; also called the 'excess mortality'. This excess mortality is useful to make comparisons between different countries and time periods (Uhry *et al.*, 2017; Allemani *et al.*, 2018) and is directly linked to the concept of net survival which is another important indicator in cancer epidemiology (Perme *et al.*, 2012). Hennerfeind *et al.* (2008) proposed a Bayesian penalized geoadditive regression model for excess hazard analysis: non-linear effects of continuous covariates and time varying effects modelled by P-splines were especially considered. However, the inference that is performed via MCMC sampling can be computationally intensive. Recently, Remontet *et al.* (2018) developed a penalized excess hazard model that relies on a data augmentation and Poisson approach. With this model, a simulation showed that the method is suitable to study the trends in the dynamics of the (excess) hazard. Nevertheless, because of the data augmentation, Poisson approaches are quite difficult to set up and computationally demanding.

For example, the model that was proposed by Remontet *et al.* (2018) is not practicable for large data sets, such as prostate or breast cancer data sets (respectively 72558 and 96726 cases; see Cowppli-Bony *et al.* (2017)), with 16 Gbytes of random-access memory using 64-bit R software (R Core Team, 2018).

Recently, Wood *et al.* (2016a) provided a general framework for smooth regression modelling that extends beyond the exponential family. They suggested estimating the smoothing parameters by maximizing a Laplace approximate marginal likelihood (LAML) criterion and demonstrated that this way maintained statistical consistency. Their optimization scheme is based on a double-Newton–Raphson procedure that requires computing the third and fourth derivatives of the likelihood. The smoothing parameters are estimated by maximizing the LAML criterion with outer Newton iterations in which each new set of smoothing parameters leads to estimating the regression parameters by maximizing the penalized likelihood through inner Newton iterations. Moreover, they used a corrected AIC that accounts for smoothing parameter uncertainty to ease model selection.

The objective of the present paper is the development of a penalized hazard and excess hazard model that does not require the Poisson approach and in which the baseline hazard, the functional forms, the time-dependent effects and the interactions between continuous covariates are modelled simultaneously by using penalized splines. The model proposed belongs to the framework that was described by Wood *et al.* (2016a) but deals with challenges that are specific to the hazard model. The superiority of the LAML criterion (also called restricted maximum likelihood) over the cross-validation criteria, such as generalized cross-validation (GCV) (Wahba, 1985), has already been demonstrated in GAMs (Reiss and Ogden, 2009; Wood, 2011) but not yet in more general frameworks. Thus, in addition to LAML, the present study includes the implementation of the likelihood cross-validation (LCV) criterion (O'Sullivan, 1988) in the way that Wood *et al.* (2016a) used the LAML criterion, i.e. as an objective criterion for the outer Newton–Raphson algorithm. For that, for the first time to our knowledge, it is proposed to calculate the derivatives of the LCV criterion with respect to the smoothing parameters. The performances of LCV and LAML are compared in a simulation study.

The present paper is structured as follows. Sections 2.1 and 2.2 present the specification of the penalized hazard regression model and its associated penalized log-likelihood. Sections 2.3 and 2.4 describe the estimation of the regression parameters and the smoothing parameters, Section 2.5 displays various possibilities for variance estimation and Section 2.6 details the estimation of independent and identically distributed Gaussian random effects when treated as penalized splines. Section 3 describes the extension of the proposed approach to the excess hazard model. Section 4 details the simulation study that evaluates the method performance within the context of net survival trend analysis. Section 5 shows an application on real cancer data that considers the effects on the excess hazard of time, age and social deprivation.

The programs that were used to analyse the data can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets`

2. The penalized hazard model

2.1. Model specification

For better readability, lower-case bold characters (e.g. $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$) stand for vectors and upper-case bold characters (e.g. $\mathbf{S}$ and $\mathbf{X}$) for matrices. The ith row of a given matrix $\mathbf{M}$ is denoted $\mathbf{M}_i$ but element i, j is denoted M_{ij}. The ith element of a given vector $\mathbf{v}$ is denoted v_i.

2.1.1. Formulation with continuous variables

We assume first that time and covariate effects may be represented via 'marginal' bases of modest rank, e.g. natural cubic splines. Using the formulation and notation of Wood *et al.* (2016a), the general form of the model is

$$\log\{h(t;\mathbf{x})\} = \sum_{j=1}^{J} g_j(t;\mathbf{x}).$$

In this formula, h is the hazard function, t is the follow-up time, $\mathbf{x}$ is a vector of continuous covariates and each function g_j can be the marginal basis of time, the marginal basis of a covariate or the tensor product of the marginal bases of any number of elements of $(t, \mathbf{x})$.

More explicitly, in the case that g_j is just a marginal basis, it may be written as a linear combination of k_j known functions b_{ji} (k_j being the rank and dimension of the basis) such that

$$g_j(x) = \sum_{i=1}^{k_j} \beta_{ji} b_{ji}(x)$$

where β_{ji} are unknown coefficients to estimate.

In the case that g_j is a tensor product which deals with interactions between continuous covariates (e.g. with time t, covariate x_1 and covariate x_2) g_j becomes

$$g_j(t, x_1, x_2) = \sum_{i_t=1}^{k_t} \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \beta_{i_t,i_1,i_2} b_{ti_t}(t) b_{x_1 i_1}(x_1) b_{x_2 i_2}(x_2)$$

where k_t, k_1 and k_2 are respectively the dimensions of the marginal bases of time t, covariate x_1 and covariate x_2.

An important feature of the model proposed is that it includes a flexible modelling of the baseline hazard, time-dependent and non-linear effects, and interactions between continuous covariates via tensor products.

Let n be the number of cases that are analysed and $\mathbf{t}$ the n-vector of follow-up durations. If $\mathbf{X}(\mathbf{t})$ is the design matrix of the model (formed from functions b_{ji}), the model may be written in matrix notation as

$$[\log\{h(t_i;\mathbf{x}_i)\}]_{1 \leqslant i \leqslant n} = \mathbf{X}(\mathbf{t})\boldsymbol{\beta}.$$

$\boldsymbol{\beta}$ is the p-vector of unknown model parameters (presented above).

Each function g_j that is contained in the design matrix may be penalized or not and, when penalized, associated with one or several penalty term(s) in the case of a simple marginal basis or a tensor product respectively. Given the linearity of g_j with respect to β, each of the M penalty terms may be written as a quadratic form in the parameter vector: $\lambda_m \boldsymbol{\beta}^{\mathrm{T}} \mathbf{S}^m \boldsymbol{\beta}$ (Wood and Augustin, 2002).

$\mathbf{S}^m$ are known positive symmetric matrices that depend on the choice of the type of penalty: a penalty based on the second derivatives of g_j (e.g. $\int g_j''(x)^2 \mathrm{d}x$) is widely used (see Wood (2006a) for an extension to tensor product smooths) but other types of penalty, such as a ridge penalty, are also possible.

The λ_m are unknown positive smoothing parameters that control the trade-off between model fit and model smoothness.

Once the type of penalty has been chosen and the terms to be penalized have been specified, the full penalty matrix $\mathbf{S}^\lambda$ can be written as $\mathbf{S}^\lambda = \Sigma_{m=1}^{M} \lambda_m \mathbf{S}^m$ (Wood *et al.*, 2016a): the objective becomes then estimating the p-vector of unknown model parameters $\boldsymbol{\beta}$ and the M-vector of unknown smoothing parameters $\boldsymbol{\lambda}$.

Suppose that we want to model the effects of time t and of four covariates $x_1, \ldots, x_4$. A penalized or smooth interaction between the effects of t, x_1 and x_2 is expected and specified, as well as a penalized non-linear but proportional effect of x_3. Finally, we want to include a linear and proportional effect of x_4 without penalization. The corresponding model (model 1) is

$$\log\{h(t, x_1, x_2, x_3, x_4)\} = g_1(t, x_1, x_2) + g_2(x_3) + \beta_4 x_4$$

where g_1 is a tridimensional tensor product spline associated with three smoothing parameters and g_2 is a natural cubic spline associated with a single smoothing parameter so that the total number of smoothing parameters to estimate is 4 ($M = 4$). Suppose that each marginal basis has a dimension of 5 (intercept included); then g_1 and g_2 are associated respectively with 125 and five regression parameters. However, an identifiability constraint must be applied to g_1 and g_2 so that their intercepts cannot be confused with each other (Wood (2017), page 250). The total number of regression parameters is then $p = 125 + 4 + 1 = 130$.

2.1.2. Dealing with categorical variables

Categorical covariates can also be inserted in the model. Suppose that, in addition to time t and continuous covariates $x_1, \ldots, x_4$, we are interested in the effect of sex (1 for men and 0 for women). Keeping the same specifications as made for model 1, the simplest model that includes the effect of a categorical variable is the model that has an intercept for each modality (model 2):

$$\log\{h(t, x_1, x_2, x_3, x_4, \text{sex})\} = g_1(t, x_1, x_2) + g_2(x_3) + \beta_4 x_4 + \beta_{\text{sex}}\text{sex}.$$

In some cases, model 2 may be too restrictive. More complex models can be considered when incorporating a categorical variable, e.g. by repeating each penalized spline and each parametric term as many times as there are modalities in the categorical variable. Each function is then repeated twice (using the indicator functions $I_{(\text{sex}=1)}$ and $I_{(\text{sex}=0)}$) (model 3):

$$\begin{aligned}\log\{h(t, x_1, x_2, x_3, x_4, \text{sex})\} = &\{g_{1,\text{men}}(t, x_1, x_2) + g_{2,\text{men}}(x_3) + \beta_{4,\text{men}} x_4\} I_{(\text{sex}=1)} \\ &+ \{g_{1,\text{women}}(t, x_1, x_2) + g_{2,\text{women}}(x_3) + \beta_{4,\text{women}} x_4\} I_{(\text{sex}=0)}.\end{aligned}$$

In model 3, each of the number of regression parameters and the number of smoothing parameters is then twice the corresponding number in model 1. In practice, model 2 is too simple to be useful and, in contrast, model 3 is very complicated with numerous categorical variables. Finding a compromise between the two is a difficult task that must be handled during the model building process; here, the corrected AIC may be helpful (Wood *et al.*, 2016a).

2.2. The penalized log-likelihood and its calculation

We denote $\boldsymbol{\delta}$ the n-vector of right censoring indicators taking value 1 when the event had occurred, and 0 otherwise, and $\mathbf{t}_0$ the n-vector of follow-up starting times in the case of left-truncated data. Under the assumption of non-informative censoring, the contribution to the log-likelihood of an individual i corresponds to the contribution to a usual survival model with right-censored and left-truncated data:

$$l_i(\boldsymbol{\beta}) = \delta_i \log\{h(t_i; \mathbf{x}_i)\} - \int_{t_{0,i}}^{t_i} h(u; \mathbf{x}_i) \mathrm{d}u.$$

For ease of reading, starting from this point $t_{0,i}$ is considered equal to 0 for all i.

The log-likelihood and the penalized log-likelihood of the model are respectively

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta})$$

and

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = l(\boldsymbol{\beta}) - \tfrac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}^{\lambda}\boldsymbol{\beta}.$$

Computing the integral of the hazard in the log-likelihood (i.e. the cumulative hazard at t_i) is a specific problem of such survival models that was not considered by Wood *et al.* (2016a). The integral may be approximated by Gauss–Legendre quadrature, which requires

(a) q, the number of nodes that are associated with the quadrature,
(b) $(d_k^*)_{1\leqslant k\leqslant q}$, the q-vector of nodes for integration over $[-1; 1]$, and
(c) $(w_k^*)_{1\leqslant k\leqslant q}$, the q-vector of weights for integration over $[-1; 1]$.

For a given individual i, the $(d_k^i)_{1\leqslant k\leqslant q}$ and $(w_k^i)_{1\leqslant k\leqslant q}$ correspond to an integration over $[0; t_i]$ and are given by

$$d_k^i = \frac{t_i}{2} d_k^* + \frac{t_i}{2}$$

and

$$w_k^i = \frac{t_i}{2} w_k^*.$$

The contribution to the log-likelihood then becomes

$$l_i(\boldsymbol{\beta}) \approx \delta_i \log\{h(t_i; \mathbf{x}_i)\} - \sum_{k=1}^{q} w_k^i h(d_k^i; \mathbf{x}_i).$$

From a computational viewpoint, in addition to the design matrix of the model, the q design matrices of size (n, p) denoted by

$$\mathbf{GL}^k = \mathbf{X}\left(\frac{\mathbf{t}}{2} d_k^* + \frac{\mathbf{t}}{2}\right)$$

may be constructed by replacing $\mathbf{t}$, the follow-up duration vector, by n-vector $(\mathbf{t}/2)d_k^* + \mathbf{t}/2$. Similarly, the q weight vectors $\mathbf{w}^k$ of length n may be constructed such that $\mathbf{w}^k = (\mathbf{t}/2)w_k^*$. Finally

$$l_i(\boldsymbol{\beta}) \approx \delta_i \mathbf{X}_i \boldsymbol{\beta} - \sum_{k=1}^{q} w_i^k \exp(\mathbf{GL}_i^k \boldsymbol{\beta})$$

where $\mathbf{X}_i$ and $\mathbf{GL}_i^k$ are vectors of length p that correspond respectively to the ith row of $\mathbf{X}(\mathbf{t})$ and $\mathbf{GL}^k$. The choice of q is critical because it represents a compromise between precision and computing time. According to Charvat *et al.* (2016) and to previous simulations (which are not shown here), $q = 20$ seems to be a reasonable choice.

As described by Wood *et al.* (2016a), the estimation procedure is based on two nested Newton–Raphson algorithms. Indeed, the estimation of the smoothing parameters $\boldsymbol{\lambda}$ is based on outer Newton–Raphson iterations where the estimation of $\boldsymbol{\beta}$ is based on inner Newton–Raphson iterations for given $\boldsymbol{\lambda}$ (see Sections 2.3 and 2.4).

2.3. Estimation of regression parameters at fixed smoothing parameters (inner algorithm)

When the smoothing parameters $\boldsymbol{\lambda}$ are known, $\hat{\boldsymbol{\beta}}^{\lambda}$ should be

$$\hat{\boldsymbol{\beta}}^{\lambda} = \arg\max_{\beta}\{\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda})\}.$$

The solution is given by the Newton–Raphson method that requires the first and second derivatives of the penalized log-likelihood according to $\boldsymbol{\beta}$. The derivatives of the contribution to the unpenalized log-likelihood are

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_l} = \delta_i X_{il} - \sum_{k=1}^{q} w_i^k \mathrm{GL}_{il}^k \exp(\mathbf{GL}_i^k \boldsymbol{\beta})$$

and

$$\frac{\partial^2 l_i(\boldsymbol{\beta})}{\partial \beta_l \partial \beta_m} = -\sum_{k=1}^{q} w_i^k \mathrm{GL}_{il}^k \mathrm{GL}_{im}^k \exp(\mathbf{GL}_i^k \boldsymbol{\beta}).$$

The gradient of the penalized likelihood is

$$\boldsymbol{\mathcal{G}}(\boldsymbol{\beta}) = \left(\sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}} - \mathbf{S}^{\lambda}\boldsymbol{\beta}.$$

For legibility, the Hessian of the negative log-likelihood $\mathbf{H}$ in matrix notation is

$$\mathbf{H}(\boldsymbol{\beta}) = \sum_{k=1}^{q} (\mathbf{GL}^k)^{\mathrm{T}}\{\mathbf{w}^k \circ \mathbf{GL}^k \circ \exp(\mathbf{GL}^k\boldsymbol{\beta})\}.$$

In this equation '$\circ$' is the Hadamard product. The Hadamard product between an n-vector $\mathbf{v}$ and an $n \times p$ matrix $\mathbf{M}$ is defined as $(\mathbf{v} \circ \mathbf{M})_{i,j} = (\mathbf{M} \circ \mathbf{v})_{i,j} = v_i M_{i,j}$. The Hessian of the negative penalized log-likelihood is then $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}) = \mathbf{H}(\boldsymbol{\beta}) + \mathbf{S}^{\lambda}$. The Newton–Raphson step will therefore be $\boldsymbol{\Lambda}(\boldsymbol{\beta}) = \boldsymbol{\mathcal{H}}(\boldsymbol{\beta})^{-1}\boldsymbol{\mathcal{G}}(\boldsymbol{\beta})$. To ensure convergence at each iteration, a diagonal perturbation of $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta})$ is made, whenever necessary, to render it positive definite and $\boldsymbol{\Lambda}$ is halved until the log-likelihood is maximized (for more details, see Wood *et al.* (2016a)).

2.4. Laplace approximate marginal likelihood and likelihood cross-validation criteria for smoothing parameter estimation (outer algorithm)

As already elaborated, a Newton–Raphson algorithm is implemented to estimate the smoothing parameters $\boldsymbol{\lambda}$. This requires the first and second derivatives of the LCV and the LAML criteria with respect to the smoothing parameters. This section recalls the definition of each criterion and presents the derivative calculations in the survival context. For clarity, $\hat{\boldsymbol{\beta}}^{\lambda}$, $\mathbf{H}(\hat{\boldsymbol{\beta}}^{\lambda})$ and $\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\beta}}^{\lambda})$ will be respectively denoted $\hat{\boldsymbol{\beta}}$, $\mathbf{H}$ and $\boldsymbol{\mathcal{H}}$. All references about matrix calculation of this section can be found in Petersen and Pedersen (2008).

2.4.1. Laplace approximate marginal likelihood criterion

As described by Wood *et al.* (2016a), in a Bayesian perspective where smoothing parameters are treated as variance hyperparameters, these parameters may be estimated by maximizing the posterior marginal likelihood over $\boldsymbol{\beta}$ (LAML criterion), which is expressed as

$$\mathrm{LAML}(\boldsymbol{\lambda}) = \mathcal{L}(\hat{\boldsymbol{\beta}}) + \frac{1}{2}\log|\mathbf{S}^{\lambda}|_{+} - \frac{1}{2}\log|\boldsymbol{\mathcal{H}}| + \frac{M_0}{2}\log(2\pi).$$

$|\mathbf{S}^{\lambda}|_{+}$ is the product of the positive eigenvalues of $\mathbf{S}^{\lambda}$ and M_0 is the number of zero eigenvalues of $\mathbf{S}^{\lambda}$ when all λ_j are strictly positive. Maximizing this criterion with respect to $\boldsymbol{\rho} = \log(\boldsymbol{\lambda})$ such that the estimated smoothing parameters remain positive during the optimization process requires the derivatives of LAML with respect to $\boldsymbol{\rho}$ (for the calculation of these derivatives, see the on-line supplementary material section A).

2.4.2. Likelihood cross-validation criterion

Prediction error criteria and cross-validation techniques (such as GCV) have been frequently used in GAMs to select the smoothing parameters. In a survival model, likelihood cross-validation techniques have been already used by O'Sullivan (1988), Verweij and Van Houwelingen (1993), Liu *et al.* (2018) and others. One important asset of likelihood cross-validation is that it can be seen as an estimator of the expected log-likelihood (Liquet and Commenges (2004); further details can be found in Commenges *et al.* (2007)). The LCV criterion is defined as follows:

$$\mathrm{LCV}(\boldsymbol{\lambda}) = -l(\hat{\boldsymbol{\beta}}) + \mathrm{tr}(\boldsymbol{\mathcal{H}}^{-1}\mathbf{H}).$$

To our knowledge, the first and second derivatives of the LCV criterion with respect to the smoothing parameters have been calculated for the first time. This enables implementing this criterion in a Newton–Raphson algorithm. In addition to numerical stability, the main advantage of this implementation is that it is not specific to hazard modelling. Indeed, by specifying another log-likelihood together with its gradient and Hessian, the derivatives of LCV that are proposed here can be used in any framework covered by Wood *et al.* (2016a).

The first derivative of LCV is

$$\frac{\partial \mathrm{LCV}}{\partial \rho_l} = -\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \rho_l} + \frac{\partial \mathrm{tr}(\boldsymbol{\mathcal{H}}^{-1}\mathbf{H})}{\partial \rho_l}.$$

Using the chain rule for $\partial l(\hat{\boldsymbol{\beta}})/\partial \rho_l$, the first derivative becomes

$$\frac{\partial \mathrm{LCV}}{\partial \rho_l} = -\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} + \frac{\partial \mathrm{tr}(\boldsymbol{\mathcal{H}}^{-1}\mathbf{H})}{\partial \rho_l}.$$

Finally, by classical derivation of the product in the above trace operator,

$$\frac{\partial \mathrm{LCV}}{\partial \rho_l} = -\frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} + \mathrm{tr}\left(\frac{\partial \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_l}\mathbf{H} + \boldsymbol{\mathcal{H}}^{-1}\frac{\partial \mathbf{H}}{\partial \rho_l}\right).$$

The second derivative of LCV is

$$\frac{\partial^2 \mathrm{LCV}}{\partial \rho_l \partial \rho_m} = -\frac{\partial^2 l(\hat{\boldsymbol{\beta}})}{\partial \rho_l \partial \rho_m} + \frac{\partial^2 \mathrm{tr}(\boldsymbol{\mathcal{H}}^{-1}\mathbf{H})}{\partial \rho_l \partial \rho_m}.$$

From the first derivative and using again the chain rule and product derivation, the second derivative becomes

$$\begin{aligned}\frac{\partial^2 \mathrm{LCV}}{\partial \rho_l \partial \rho_m} = &-\frac{\partial \hat{\boldsymbol{\beta}}^{\mathrm{T}}}{\partial \rho_m}\mathbf{H}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_l} - \frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\frac{\partial^2 \hat{\boldsymbol{\beta}}}{\partial \rho_l \partial \rho_m} \\ &+ \mathrm{tr}\left(\frac{\partial^2 \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_l \partial \rho_m}\mathbf{H} + \frac{\partial \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_l}\frac{\partial \mathbf{H}}{\partial \rho_m} + \frac{\partial \boldsymbol{\mathcal{H}}^{-1}}{\partial \rho_m}\frac{\partial \mathbf{H}}{\partial \rho_l} + \boldsymbol{\mathcal{H}}^{-1}\frac{\partial^2 \mathbf{H}}{\partial \rho_l \partial \rho_m}\right)\end{aligned}$$

with

$$\frac{\partial^2 \mathcal{H}^{-1}}{\partial \rho_l \partial \rho_m} = -\frac{\partial \mathcal{H}^{-1}}{\partial \rho_m}\frac{\partial \mathcal{H}}{\partial \rho_l}\mathcal{H}^{-1} - \mathcal{H}^{-1}\frac{\partial^2 \mathcal{H}}{\partial \rho_l \partial \rho_m}\mathcal{H}^{-1} - \mathcal{H}^{-1}\frac{\partial \mathcal{H}}{\partial \rho_l}\frac{\partial \mathcal{H}^{-1}}{\partial \rho_m}.$$

2.4.3. The outer algorithm

For each new proposal $\boldsymbol{\rho}$ in the optimization process of LAML or LCV, an inner Newton algorithm is needed to find $\hat{\boldsymbol{\beta}}^{\lambda}$ (which is referred to below as $\hat{\boldsymbol{\beta}}^{\rho}$ to relate to Section 2.5). The optimization of LAML or LCV is thus called the 'outer algorithm' and leads to smoothing parameters estimates $\hat{\boldsymbol{\rho}}$ with associated regression parameters $\hat{\boldsymbol{\beta}}^{\rho}$. In the outer algorithm, whenever necessary, Hessian perturbation is also performed and the step length is controlled so that LAML or LCV is optimized at each iteration (the step is divided by 10 in the outer algorithm instead of 2 in the inner algorithm). Moreover, to keep $\boldsymbol{\rho}$ inside the parameter space, the maximum absolute values of the components of the step vector are controlled at each iteration by an upper threshold (by default, this threshold is set to 5).

2.5. Estimation of the regression parameter variance

There are three possibilities for variance estimation proposed by Wood *et al.* (2016a). First, we may use a Bayesian large sample approximation to derive a Bayesian posterior covariance matrix:

$$\boldsymbol{\beta}|\boldsymbol{\rho} \sim N(\hat{\boldsymbol{\beta}}^{\rho}, \mathbf{V}_{\beta}) \qquad \mathbf{V}_{\beta} = \mathcal{H}^{-1}.$$

This Bayesian covariance matrix was first described by Wahba (1983).

Alternatively, we may consider a frequentist estimate (for further details, see Gray (1992)):

$$\mathbf{V}_{\hat{\beta}} = \mathcal{H}^{-1}\mathbf{H}\mathcal{H}^{-1}.$$

Finally, in addition to the uncertainty of $\hat{\boldsymbol{\beta}}$, Wood *et al.* (2016a) proposed to take into account the uncertainty arising from the estimation of the smoothing parameters. This leads to a third option for the covariance matrix for $\hat{\boldsymbol{\beta}}$:

$$\mathbf{V}'_{\beta} = \mathbf{V}_{\beta} + \mathbf{V}' + \mathbf{V}''.$$

In this equation, $\mathbf{V}' = \mathbf{J}\mathbf{V}_{\rho}\mathbf{J}^{\mathrm{T}}$,

$$V''_{jm} = \sum_{i}^{p}\sum_{l}^{M}\sum_{k}^{M}\frac{\partial R_{ij}}{\partial \rho_k}V_{\rho,kl}\frac{\partial R_{im}}{\partial \rho_l},$$

$\mathbf{J} = \partial\hat{\boldsymbol{\beta}}/\partial\boldsymbol{\rho}|_{\hat{\boldsymbol{\rho}}}$, $\mathbf{V}_{\rho}$ is the inverse of the Hessian of the negative log-marginal-likelihood with respect to $\boldsymbol{\rho}$ (calculated at $\hat{\boldsymbol{\rho}}$), $\mathbf{R}$ is such that $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{V}_{\beta}$, p is the number of regression parameters and M the number of smoothing parameters. This new covariance matrix is thus a Bayesian covariance matrix to which are added two corrected terms (for more details on the definitions of $\mathbf{V}'$ and $\mathbf{V}''$, see Wood *et al.* (2016a), section 4).

2.6. Hazard regression and random effects

The LAML criterion enables estimation of the smoothing parameters as if they were the variance hyperparameters of a Gaussian prior on the regression parameters. An interesting consequence is that this process can work in reverse: it is possible to estimate random effects by treating them

as penalized splines (Wood *et al.*, 2016a). In the hazard regression context, consider the model

$$[\log\{h(t_i;\mathbf{x}_i)\}]_{1\leqslant i\leqslant n} = \mathbf{X}(\mathbf{t})\boldsymbol{\beta} + \mathbf{Z}(\mathbf{t})\boldsymbol{\gamma}$$

where $\mathbf{Z}$ is the design matrix for the random effect and $\boldsymbol{\gamma}$ the associated regression parameters.

Denoting by $\mathbf{S}$ the corresponding penalty matrix and λ the smoothing parameter, the Gaussian prior on $\boldsymbol{\gamma}$ is of the form

$$f(\boldsymbol{\gamma}) \propto \exp\left(-\frac{\lambda}{2}\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{S}\boldsymbol{\gamma}\right).$$

Therefore, if we want to specify an independent and identically distributed Gaussian random effect, i.e. $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and

$$f(\boldsymbol{\gamma}) \propto \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{I}\boldsymbol{\gamma}\right),$$

then $\lambda = 1/\sigma^2$ and $\mathbf{S} = \mathbf{I}$. Thus, both $\boldsymbol{\gamma}$ and λ (and therefore σ^2) can be estimated by considering a ridge penalty on $\boldsymbol{\gamma}$ with LAML estimation of λ. In addition, as $\rho = \log(\lambda) = -2\log(\sigma)$, $\mathrm{var}\{\log(\hat{\sigma})\} = \frac{1}{4}\mathbf{V}_\rho$.

3. Extension to the penalized excess hazard model

In a survival study, examining the excess mortality of the patients in comparison with the mortality in a reference population can be more informative than examining the overall mortality. For example, oncologists would like to know the effect of tumours on mortality. However, the causes of death are not always reliable or available (in particular in population-based cancer registries) and death from cancer is difficult to ascertain in long-term follow-ups (cancer treatments may have long-term toxicities resulting in patients' deaths). The concept of 'excess mortality' is thus a relevant alternative to the concept of 'cause of death' in an attempt to isolate extra deaths that are attributable (directly or indirectly) to cancer. This excess mortality can be estimated by supposing that, in cancer patients, the mortality that is attributable to all other causes but the cancer under study may be approximated by the all-cause mortality of the general population (the 'expected mortality'). According to Estève *et al.* (1990), it is assumed that

$$h_{\mathrm{O}}(t,\mathbf{x}) = h_{\mathrm{E}}(t,\mathbf{x}) + h_{\mathrm{P}}(a+t,\mathbf{z}).$$

In this equation, h_{O} is the mortality that is observed in cancer patients, h_{E} is the excess mortality due to cancer, t is the time elapsed since cancer diagnosis, a is the age at cancer diagnosis, $\mathbf{x}$ is the vector of prognostic variables whose effects on h_{E} are being studied and $\mathbf{z}$ is the vector of demographic characteristics that provides for each individual its expected mortality h_{P} (h_{P} is the all-cause mortality rate of the general population at age $a+t$, given characteristics $\mathbf{z}$ of that individual).

The net survival is the survival that would be observed if cancer were the only cause of death; it can be directly obtained from h_{E} by using the classical relationship between hazard and survival.

In the excess hazard framework, h_{P} is considered known (available from national statistics) and h_{E} may be modelled as

$$[\log\{h_{\mathrm{E}}(t_i;\mathbf{x}_i)\}]_{1\leqslant i\leqslant n} = \mathbf{X}(\mathbf{t})\boldsymbol{\beta}.$$

By construction, the model ensures a positive excess hazard.

Let **a** be the n-vector of ages at diagnosis; the contribution of a given individual i to the log-likelihood is

$$l_i(\boldsymbol{\beta}) = \delta_i \log\{h_E(t_i, \mathbf{x}_i) + h_P(a_i + t_i, \mathbf{z}_i)\} - \int_0^{t_i} \{h_E(u, \mathbf{x}_i) + h_P(a_i + u, \mathbf{z}_i)\} \mathrm{d}u.$$

As in the overall survival context, left-truncated data can obviously be handled but, for ease of reading, the integration starts from 0. As h_P does not depend on $\boldsymbol{\beta}$, maximizing $l_i(\boldsymbol{\beta})$ is equivalent to maximizing

$$l_i(\boldsymbol{\beta}) = \delta_i \log\{h_E(t_i, \mathbf{x}_i) + h_P(a_i + t_i, \mathbf{z}_i)\} - \int_0^{t_i} h_E(u, \mathbf{x}_i) \mathrm{d}u.$$

Multi-dimensional penalized splines may be implemented in the excess hazard model following the same principle as that used for overall mortality. However, computing the derivatives is more difficult because h_P is present in the expression of $l_i(\boldsymbol{\beta})$ (for more details, see the on-line supplementary material section B).

4. Simulation study

The trends in net survival according to the year of cancer diagnosis are important in cancer epidemiology studies that use population-based registry data. The analysis of these trends takes into account the age at cancer diagnosis because age is a strong predictor of net survival. To assess the performance of the proposed approach within this context, survival data were simulated where the excess hazard varied according to the time since diagnosis, age and the year of diagnosis.

4.1. Design

The simulated data that are described here are similar to those used and detailed in Remontet *et al.* (2018). The simulation study considered two scenarios: one arising from oesophagus cancer (no trend) and one from *cervix uteri* cancer (complex trend). First, two parametric (unpenalized) models that take into account the effects of age and year of diagnosis were fitted to real French survival data (Cowppli-Bony *et al.*, 2017) and the model parameters were used to generate simulation data. Fractional polynomials (Royston and Sauerbrei, 2008) were used instead of splines to avoid overestimating the performance of the novel approach. The model building strategy of Sauerbrei *et al.* (2007) was adopted to choose the powers of the fractional polynomials.

Each scenario considered three sample sizes ($N = 2000$, 5000, 10000 patients). Scenarios with 2000 or 5000 patients used 1000 data sets ($D = 1000$), whereas scenarios with 10000 patients used only 200 data sets ($D = 200$) to limit the computing time. Each scenario considered a cohort whose age distribution is identical to that of the French patients with the same cancer. The year of diagnosis was randomly sampled from a uniform distribution between 1990 and 2010. The patients were censored at 5 years or at the end of follow-up (set to 2013).

The time to death of each patient, T, was the lowest between the time to death due to cancer, T_E, and the time to death from another cause, T_P. T_P was generated by using a piecewise exponential distribution as described by Danieli *et al.* (2012). Once the parameters of the two excess hazard models had been estimated, the cumulative distribution of T_E was derived and T_E was generated by using inverse transform sampling.

4.2. Description of the theoretical trends

In both scenarios, the effect of age was assumed non-linear and time dependent. The *cervix uteri* cancer scenario included a linear and time-dependent effect of the year of diagnosis, with a strong age–year interaction, and a triple interaction between time since diagnosis, age and year of diagnosis.

In contrast, the oesophagus cancer scenario assumed that the year of diagnosis has no effect on the excess mortality. This assumption was made to study the performance of the approach whenever a strong smoothing is needed in one dimension (year) whereas an important curvature is present in another dimension (time). Indeed, one interesting feature of this scenario is the non-monotony of the excess hazard function according to the time elapsed since diagnosis: the excess mortality hazard increased up to 1 year after diagnosis and then decreased steeply (Fig. 1).

4.3. Fitted models

The joint effects of time since diagnosis, age at diagnosis and year of diagnosis on the log-excess hazard of death were modelled as a tensor product smooth of the three covariates by using natural cubic splines as basis functions. The number of knots that were associated with time, age and year were 6, 5 and 4 respectively. Their locations were the following percentiles of observed deaths: 0, 0.20, 0.40, 0.60, 0.80 and 1 for time, 0, 0.25, 0.50, 0.75 and 1 for age, and 0, 0.33, 0.66 and 1 for year. The total number of regression parameters was then 120. The extent of penalization was controlled by three smoothing parameters. Two models were fitted to the simulated data sets: one with LAML and another with LCV estimation of the three smoothing parameters.

4.4. Assessment of the approach performance

The performance of the approach was assessed by estimating the excess hazard and the net survival at given combinations of times after cancer diagnosis (0.5, 1, 2, 3, 4 and 5 years), ages at diagnosis (in 5-year increments from age 30 to age 85 years for *cervix uteri* and from 45 to 85 years for oesophagus cancer), and years of diagnosis (in 5-year increments from 1990 to 2010). The boxplots and the median coverage probabilities that are presented below were calculated over all these combinations.

The study considered 12 settings (2 scenarios $\times$ 3 sample sizes $\times$ 2 smoothing parameter criteria). At a given time t and for a vector of covariates $\mathbf{x}$, let $h_{\mathrm{E}}(t,\mathbf{x})$ and $\mathrm{NS}(t,\mathbf{x})$ be the theoretical values of the excess hazard and the net survival, and $\hat{h}_{\mathrm{E}}^{d}(t,\mathbf{x})$ and $\widehat{\mathrm{NS}}^{d}(t,\mathbf{x})$ be their estimated counterparts at the dth simulated data set. In each of the 12 settings, the following parameters were estimated:

(a) the root mean integrated squared error (RMISE) (see for example Rondeau *et al.* (2007)) of the excess hazard,

$$\mathrm{RMISE}=\sqrt{\left(\frac{1}{D}\sum_{d=1}^{D}\left[\int\{\hat{h}_{\mathrm{E}}^{d}(t,\mathbf{x})-h_{\mathrm{E}}(t,\mathbf{x})\}^{2}\mathrm{d}t\right]\right)},$$

(b) the bias in estimating the net survival, $(1/D)\Sigma_{d=1}^{D}\{\widehat{\mathrm{NS}}^{d}(t,\mathbf{x})-\mathrm{NS}(t,\mathbf{x})\}$,

(c) the root-mean-squared error (RMSE) in estimating the net survival,

$$\sqrt{[(1/D)\sum_{d=1}^{D}\{\widehat{\mathrm{NS}}^{d}(t,\mathbf{x})-\mathrm{NS}(t,\mathbf{x})\}^{2}]},$$

and

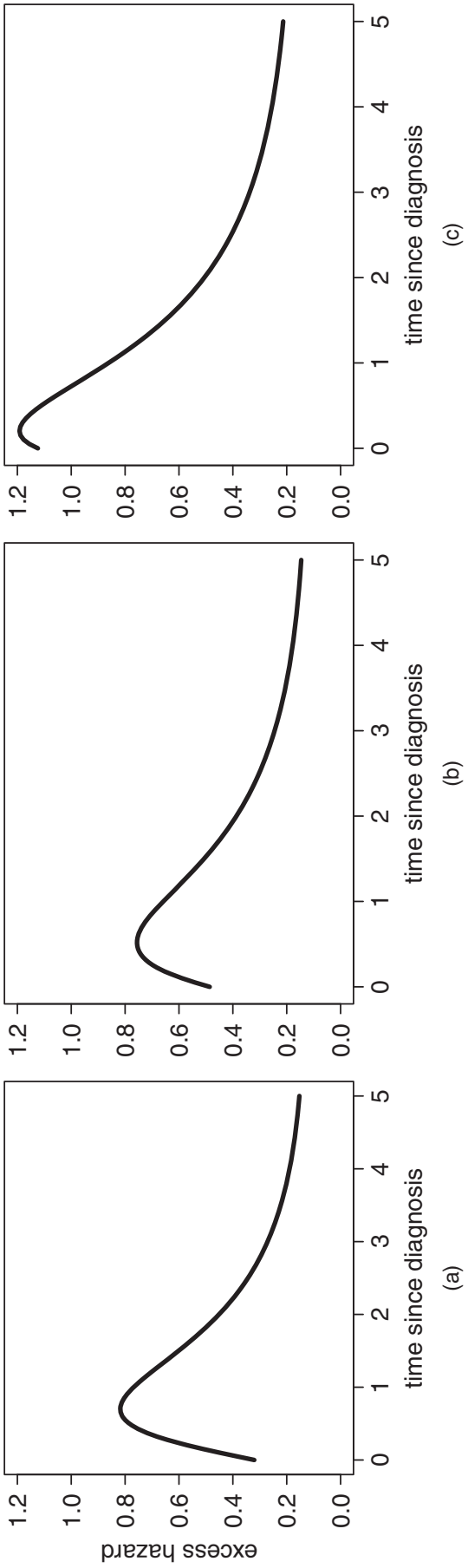

Fig. 1. Theoretical excess hazards in (a)–(c) oesophagus cancer and (d)–(f) *cervix uteri* cancer at three ages (a), (d) 40, (b), (e) 60 and (c), (f) 80 years, and three years of diagnosis 1990 (——), 2000 (– – –) and 2010 (· · · ·)

(d) the empirical coverage probability (CP) defined as the proportion of 95% confidence intervals (CIs) that include the theoretical value.

The RMISE was approximated by Gauss–Legendre quadrature using 20 values of time. CPs were estimated for CIs computed with the frequentist variance $\mathbf{V}_{\hat{\beta}}$, the Bayesian variance $\mathbf{V}_{\beta}$ and the corrected variance $\mathbf{V}'_{\beta}$, which was introduced to propose a version of the AIC that accounts for smoothing parameter uncertainty (for LAML estimation only). Nevertheless, because the use of $\mathbf{V}'_{\beta}$ for computing the CIs was discussed (Greven and Scheipl, 2016; Wood *et al.*, 2016b), the CPs of the CIs were also estimated by using $\mathbf{V}'_{\beta}$. The effective degrees of freedom edf were used to measure model complexity. They were defined from the Bayesian variance $\mathbf{V}_{\beta}$: $\text{edf} = \text{tr}(\mathbf{V}_{\beta}\mathbf{H})$. To assess the relative performance of the approach further and to validate the data generation algorithm, the true model (i.e. the model that generated the data) was also used and considered as the gold standard.

4.5. Results

Among the 8800 models fitted (4400 with LAML and 4400 with LCV), only one LCV model failed to converge. The overall convergence properties of the novel approach were then highly satisfactory. The median computing times of the LCV and the LAML algorithms were comparable with sample sizes 2000 and 5000 (around 60 s and 160 s per data set respectively, on a desktop computer with Intel i5-6600 3.30 GHz and 16 Gbytes of random-access memory). With sample size 10000, LAML optimization was slightly faster than LCV optimization (430 *versus* 460 s). The variance of the computing time was higher with LCV optimization. Regarding the edf, those derived from LCV tended to be slightly higher than those derived from LAML whatever the sample size and scenario. Table 1 shows the median values of edf obtained across all fitted models.

The results from Table 1 indicate that the estimates that were derived from LAML were smoother than those derived from LCV.

Fig. 2 shows the performance of the approach in terms of excess hazard RMISE and Fig. 3 shows the bias and RMSE in the net survival.

As expected, in the oesophagus cancer scenario, the tensor RMISE was much larger than that of the true model. The tensor was too flexible for this simple scenario, especially because it takes into account the effect of the year of diagnosis that is not present in the true model. Whatever the scenario or the sample size, LAML seemed to perform slightly better than LCV in terms of median values of the RMISE or the RMSE (e.g. median RMSE $\times$ 100 in the oesophagus scenario, $N = 2000$: 2.31 with LCV *versus* 2.24 with LAML). Moreover, the LCV results seemed to be slightly more dispersed than those of the LAML. These results agree with those of Ruppert

Table 1. Medians of the effective degrees of freedom calculated across all fitted models

Scenario	*Sample size*	*edf for LCV*	*edf for LAML*
Oesophagus	2000	22.8	21.5
	5000	27.5	26.7
	10000	31.5	31.0
Cervix uteri	2000	19.5	17.8
	5000	26.2	24.8
	10000	32.6	30.3

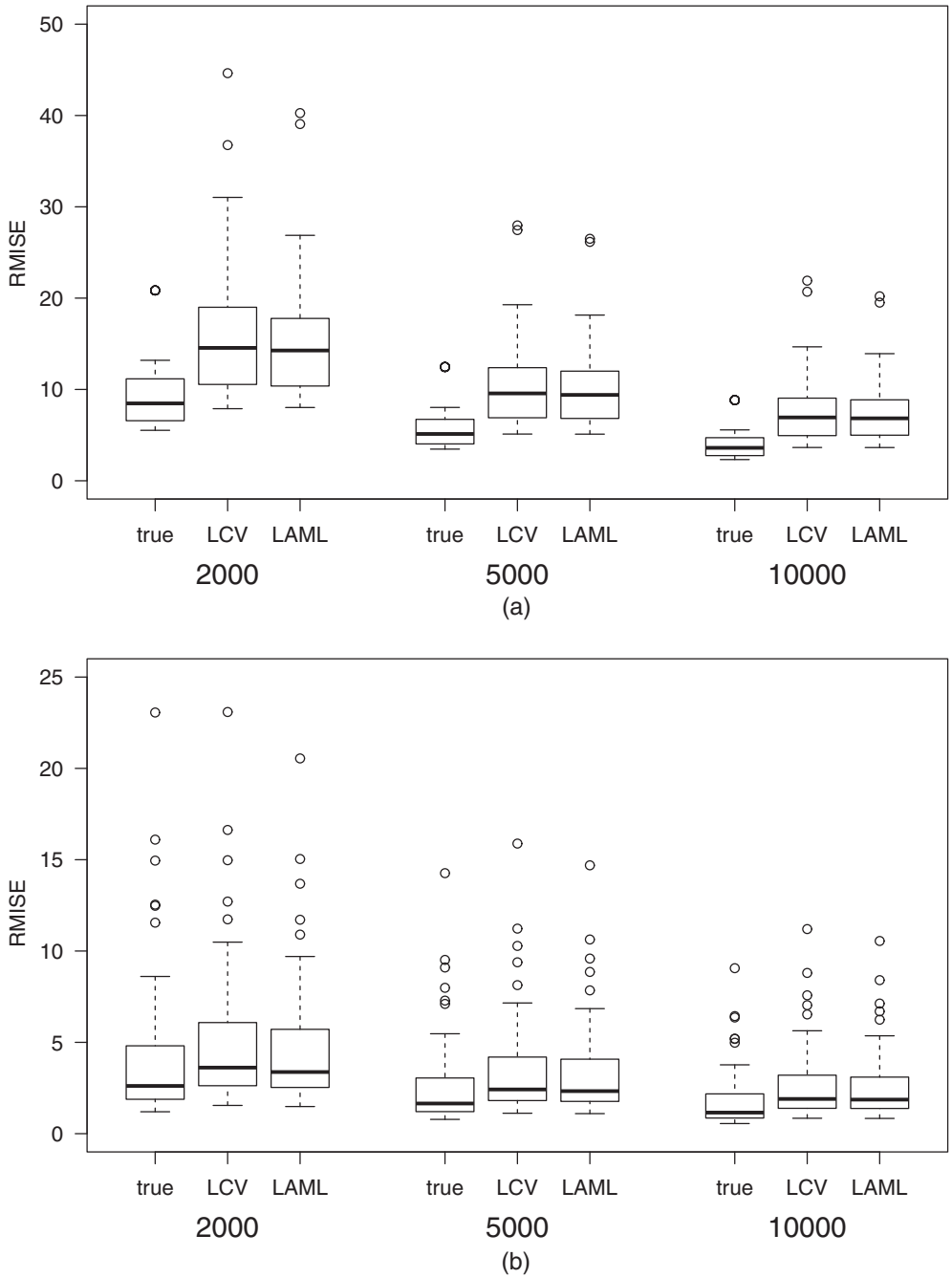

Fig. 2. Boxplots of excess hazard RMISE multiplied by 100 for all scenarios and sample sizes (for each scenario and sample size, the performances of the true model, the tensor model with LCV estimation of the smoothing parameters and the tensor model with LAML estimation are compared; the boxplots are calculated over all combinations of age (in 5-year increments from 45 to 85 years for oesophagus and from 30 to 85 years for cervix) and year of diagnosis (in 5-year increments from 1990 to 2010)): (a) oesophagus; (b) cervix

et al. (2003) who showed that the restricted maximum likelihood is likely to have less variance than the criterion stemming from GCV.

Tables 2 and 3 show the medians of the CPs that were obtained with each covariance matrix in estimating the excess hazard of death and the net survival.

The CPs that were obtained with the frequentist covariance matrix (either LCV or LAML) were quite low, especially in the *cervix uteri* cancer scenario. This poor result was expected because, generally, frequentist covariance matrices do not take into account the penalty-induced

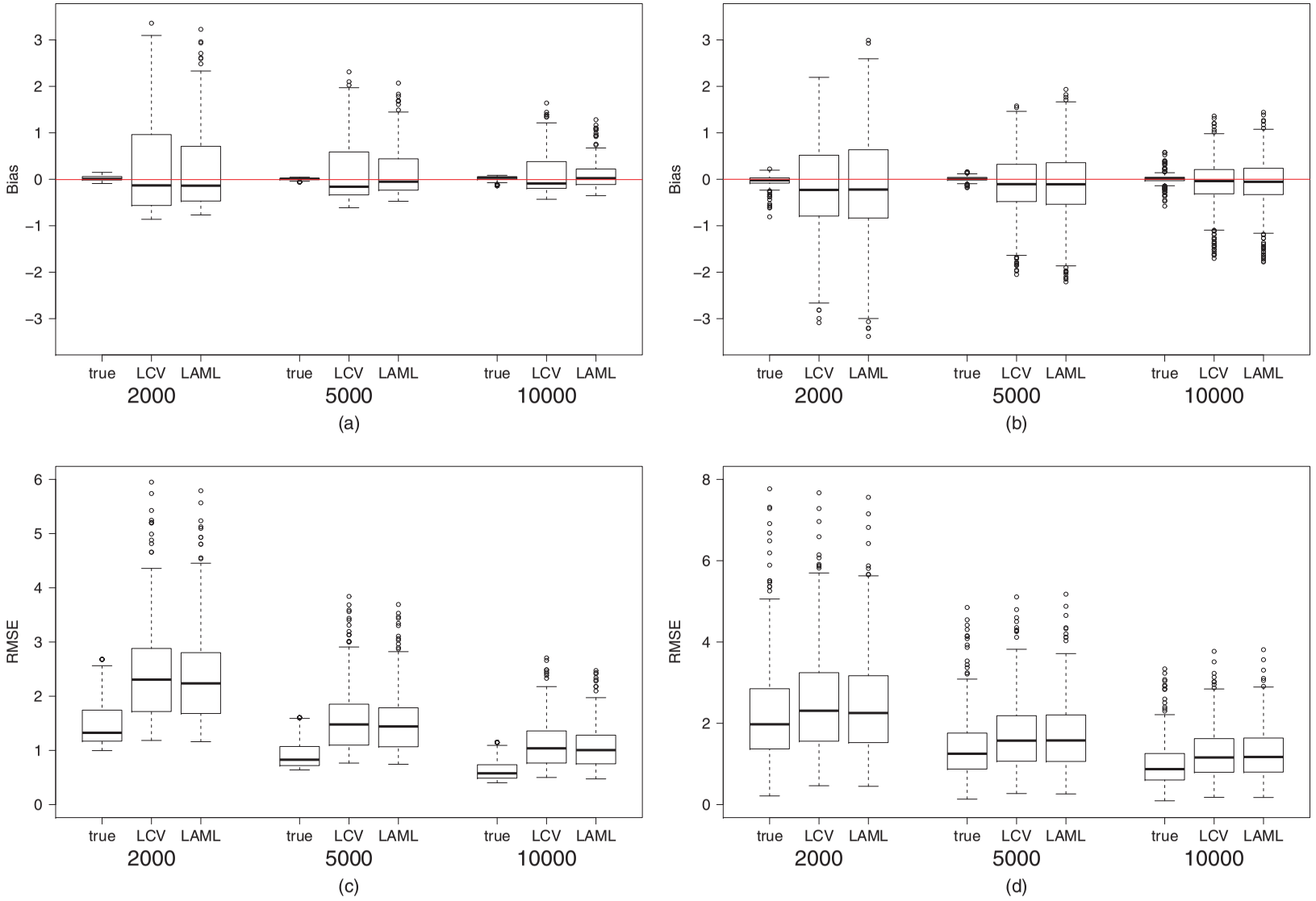

Fig. 3. Boxplots of (a), (b) net survival bias and (c), (d) RMSE multiplied by 100 for all scenarios and sample sizes (for each scenario and sample size, the performances of the true model, the tensor model with LCV estimation of the smoothing parameters and the tensor model with LAML estimation are compared; the boxplots are calculated over all combinations of time (0.5, 1, 2, 3, 4 and 5), age (in 5-year increments from 45 to 85 years for oesophagus and from 30 to 85 years for cervix) and year of diagnosis (in 5-year increments from 1990 to 2010)): (a), (c) oesophagus; (b), (d) cervix

Table 2. Medians of the CPs obtained when estimating the excess hazard of death

Scenario	*Sample size*	*Median of the CPs† with the following methods:*					
		True model	*Frequentist LCV*	*Frequentist LAML*	*Bayesian LCV*	*Bayesian LAML*	*Corrected LAML*
Oesophagus	2000	94.5	92.3	91.9	95.4	95.7	97.5
	5000	94.8	93.3	93.5	96.0	96.2	97.4
	10000	96.0	94.0	94.5	96.5	96.5	97.5
Cervix uteri	2000	95.1	92.0	92.4	96.1	96.3	98.2
	5000	94.6	92.0	92.0	96.0	96.1	97.8
	10000	95.0	92.5	92.5	96.0	96.5	97.5

†Calculated over all ages, years of diagnosis and times elapsed since diagnosis.

Table 3. Medians of the CPs obtained when estimating the net survival

Scenario	*Sample size*	*Median of the CPs† with the following methods:*					
		True model	*Frequentist LCV*	*Frequentist LAML*	*Bayesian LCV*	*Bayesian LAML*	*Corrected LAML*
Oesophagus	2000	94.8	91.5	92.0	94.1	94.8	96.9
	5000	95.2	92.2	93.1	94.9	95.6	96.9
	10000	96.0	93.5	94.5	96.0	96.5	97.5
Cervix uteri	2000	95.1	90.8	90.6	93.4	93.3	96.3
	5000	95.0	91.1	90.4	93.9	93.5	96.1
	10000	95.0	92.0	91.5	94.5	94.0	96.0

†Calculated over all ages, years of diagnosis and times elapsed since diagnosis.

bias (Wood, 2006a). In contrast, the CPs that were obtained with the Bayesian (uncorrected) covariance matrices were close to 95% even in the oesophagus cancer scenario where the smoothing parameter that is associated with the year of diagnosis is infinite. Thus, the Bayesian covariance matrix offers a close-to-nominal coverage without accounting for smoothing parameter uncertainty. As expected, the corrected Bayesian covariance matrix that takes parameter uncertainty into account provided larger variance estimates. In estimating the hazard of death, the CPs already slightly higher than 95% tended to move away from the nominal value. Generally, LCV tended to give lower CPs than did LAML, but the performances of the two criteria with Bayesian covariance matrices became equivalent with increasing sample sizes.

Overall, the LAML optimization performed slightly better than the LCV optimization in terms of RMISE, RMSE and CPs. However, the differences were very slight and both criteria remained useful.

5. Effect of social deprivation on the excess mortality in cancer patients

The penalized hazard approach that is described in this paper was applied to real data to study the effect of social deprivation on the excess hazard in cancer patients. It was particularly interesting to study whether the effect of social deprivation varies with age.

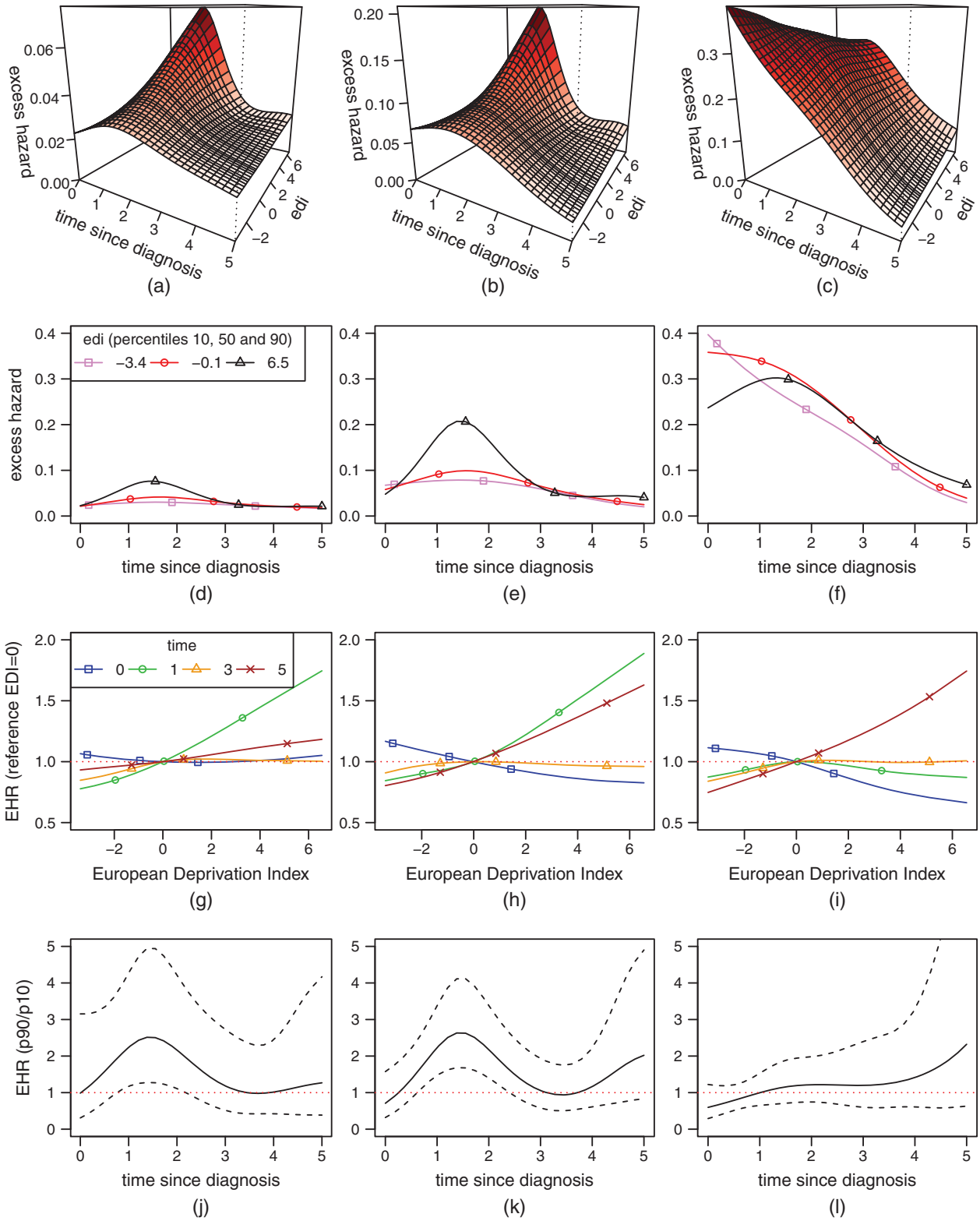

Fig. 4. Results of the model fit: (a)–(c) fitted surfaces of excess hazard *versus* time (in years) and EDI; (d)–(f) excess hazard *versus* time for three EDI values; (g)–(i) excess hazard ratio *versus* EDI (the reference is EDI = 0) at four time points after diagnosis (0, 1, 3 and 5 years); (j)–(l) excess hazard ratio between the 90th and the 10th percentile of EDI *versus* time along with the 95% CIs; (a), (d), (g), (j) age 35 years; (b), (e), (h), (k) age 51 years; (c), (f), (i), (l) age 80 years

This application used the French version of the EDI (Guillaume *et al.*, 2016), which is a continuous variable whose value increases with the degree of social deprivation. The study concerned the effects of time since diagnosis, t, age at diagnosis, a, and EDI on the excess hazard among 1865 *cervix uteri* cancer patients who were diagnosed between 2006 and 2009 (end of follow-up: June 30th, 2013) within an area covered by 14 French cancer registries. In this population, 689 patients (37%) died within 5 years after cancer diagnosis. The expected mortality rates h_P were derived from the observed mortality rates available by sex, age, year of death and *Département* of residence (the data were provided by the French Institut National de la Statistique et des Études Économiques) and were smoothed beforehand by using a Poisson regression model that included a bidimensional penalized spline of age and year of diagnosis (for each given sex and *Département*).

The model to fit was

$$\log\{h_E(t, a, \text{EDI})\} = \text{tensor}(t, a, \text{EDI}).$$

Each marginal basis was a natural cubic spline with five knots (placed at the percentiles 0, 0.25, 0.5, 0.75 and 1). Thus, the tensor product led to 125 parameters. The three smoothing parameters were estimated via LAML optimization. Fig. 4 shows different predictions from the model. In Fig. 4, the three ages and the three EDI values correspond to the 10th, 50th and 90th percentiles of age and EDI. The fitted surfaces from the model proved the ability of the tensor to capture complex interactions without too much wiggliness.

The excess mortality was more important among patients living in the most deprived environments (90th percentiles of EDI *versus* 10th or 50th), though this was not found at age 80 years during the first year of follow-up after diagnosis (Figs 4(d)–4(f)). Moreover, the ratio in excess hazard between the least and the most deprived environments was statistically significant at ages 35 and 51 years around 18 months after diagnosis (see the CIs that are shown in Figs 4(j)–4(l)). This ratio decreases then with the increase of the time elapsed since diagnosis. The method detected no significant effect of social deprivation at age 80 years.

Regarding net survival, the model predicted, at age 51 years, a 1-year net survival of 93% [90.7–94.7%] for the 10th EDI percentile and 90.2% [86.7–92.8%] for the 90th percentile. The corresponding predicted 5-year net survivals were 74.5% [69.9–78.4%] and 61% [53.9–67.2%] respectively. Thus, the differences that were observed in excess hazard dynamics between 1 and 2 years after diagnosis were reflected in terms of 5-year net survival. These results illustrate the need to interpret survival estimates jointly with the dynamics of the excess hazard, the latter being able to identify differences along the time elapsed since diagnosis.

6. Discussion

6.1. Summarizing remarks

This paper proposes a penalized hazard model that is suited for extensive production of epidemiological indicators in cancer survival. Indeed, the method proposed, which was applied to a real data set and whose performance was assessed by simulation, is applicable to a wide variety of types of cancer. Even the most challenging cancer sites, such as breast and prostate, can be readily analysed with this new methodology.

The approach proposed is based on the general smooth framework of Wood *et al.* (2016a). It offers an alternative smoothing parameter estimation procedure via LCV that can be extended beyond hazard regression models. As explained in Section 2.4.2, by specifying another log-likelihood with its gradient and Hessian, the derivatives of LCV may be used in any of the

frameworks that were considered by Wood *et al.* (2016a). The paper details also an extension of the model to excess hazard regression.

6.2. Technical points

The novel approach allows a flexible modelling of the baseline hazard as well as the time-dependent and non-linear effects of the covariates. Because natural cubic splines are used as marginal bases, the number of knots and their locations need to be specified. However, this is less important in a penalized than in an unpenalized context. Indeed, if k is the number of knots, then

> 'provided that k is large enough [...], neither the exact choice of k, nor the precise selection of knot locations has a great deal of influence on the model fit'

(Wood (2017), section 4.2.2). In practice, it is reasonable to specify a slightly higher number of knots than deemed necessary and to let the penalization avoid overfitting.

In the simulation study based on the trend analysis context, the performance of LCV was slightly less satisfactory than that of LAML (proposed by Wood *et al.* (2016a)). In practice, LAML tends to give smoother estimates (see Table 1) and therefore seems less prone to overfitting than is LCV. Besides, up to now, LAML optimization remains the only way of quantifying smoothing parameter uncertainty in a non-fully-Bayesian setting. However, an important feature of LCV is that it can be used with all kinds of penalties (i.e. not necessarily quadratic penalties such as the lasso penalty) whereas the Bayesian approach that is used to define LAML relies on quadratic penalties.

Calculating the log-determinant of the penalty matrix and its derivatives is the key to stable computation with LAML. Indeed, when some smoothing parameters tend to ∞ whereas others remain finite, these calculations may become unstable. Regarding this issue, the approach that is described in the on-line supplementary material section A differs from that of Wood *et al.* (2016a) and may appear more unstable when extreme cases occur (e.g. when one smoothing parameter is close to 10^6 whereas another is close to 10^{-6}). However, these extreme cases have never been observed in practice.

In the simulation study, the Bayesian covariance matrix offered a close-to-nominal CP without accounting for smoothing parameter uncertainty. Wahba (1983) and Nychka (1988) have previously shown that the CPs that are obtained from Bayesian CIs can be very close to the nominal values. Furthermore, as described by Marra and Wood (2012), as long as heavy over-smoothing is avoided, close-to-nominal CPs can be achieved by Bayesian CIs. Finally, recent results showed that bad coverage was not necessarily the consequence of overlooking smoothing parameter uncertainty but rather the consequence of a poor smoothing parameter estimation procedure (Wood (2017), page 297).

The method deals easily also with interactions between continuous covariates by considering a tensor product smooth. Interactions may be even specified as tensor product interactions (based on functional analysis-of-variance decomposition; see Wood (2006b)) to estimate separately the extent of penalization of the marginal bases and the interaction terms. This kind of decomposition is appealing in survival analysis because the main effects are often more complex than the interactions.

6.3. Epidemiological considerations

This paper investigated the effect of social deprivation (EDI) on the excess mortality hazard in *cervix uteri* cancer patients. A penalized tensor product smooth was used to model simultaneously the non-linear effects of time, age and EDI as well as interactions. This flexible approach enabled us to show that, for young and middle-aged women, the excess hazard in the most

deprived patients was significantly higher than the excess hazard in the least deprived. This superiority, mostly observed between 1 and 2 years after diagnosis, resulted in important differences in terms of 5-year net survival.

Although motivated by trend analyses and socio-economic inequality studies in cancer survival, the statistical methods that were developed in this paper have a wider applicability. Indeed, several time-to-event data analyses may benefit from this novel approach because it is generally difficult to specify a parametric model (even with only the survival time and two other covariates like in the above application on social deprivation).

6.4. Comparison with alternative approaches

Though the derivatives that are needed for the inner and outer Newton–Raphson algorithms in survival models are difficult to obtain, they provide a numerically stable and efficient way of estimation with highly satisfactory convergence properties. Moreover, Gauss–Legendre quadrature has proven to be an accurate technique to approximate the cumulative hazard while avoiding artificial data splitting. The on-line supplementary material section C.1 presents comparisons between R packages `frailtypack`, `gss`, `R2BayesX`, `rstpm2`, `mgcv`, `bamlss` and `survPen` in terms of computational times. These comparisons show the efficiency of `survPen` (the package that was derived from the present work).

The on-line supplementary material section C.2 presents a comparison between `survPen` and `rstpm2` within the frame of the above-detailed simulation study and shows the superiority of `survPen` over `rstpm2` in terms of statistical performance and reliability.

Two features explain this superiority:

(a) `survPen` uses explicit fourth-order derivatives of the likelihood whereas `rstpm2` resorts to numerical derivatives via R function `optim` and
(b) in `survPen`, the parameters are defined on the logarithm of the hazard (a scale that is deemed 'natural') whereas the scales of the parameters that are available in the `pstpm2` function require specifying a complex system of constraints over many parameters, which adds complexity to the optimization algorithm.

However, further research is needed to confirm these explanations.

6.5. Limits

The counterpart of determining high order derivatives is that extending the approach to other settings may be challenging. For example, incorporating random effects, as described in Section 2.6, may be inefficient when the number of clusters is very large (Wood, 2011). Thus, it would be interesting to extend the approach to a 'genuine' frailty model that maximizes a penalized marginal likelihood during the inner Newton–Raphson step. However, deriving such a likelihood is not easy because of the integration over the random-effect distribution.

In general, the tensor product approach reaches its limit when the number of covariates equals or exceeds 4. In this case, variable-selection strategies are needed (for such selection in the GAM setting, see Marra and Wood (2011)) and the corrected AIC (Wood *et al.*, 2016a) for penalized hazard model selection is an interesting possibility.

6.6. Conclusion

In comparison with other approaches for hazard regression, the method proposed offers four main advantages:

(a) its optimization procedure uses the exact derivatives of the objective functions (penalized likelihood, LCV and LAML criteria) for high numerical stability and speed;

(b) it belongs to the theoretical framework that was developed by Wood *et al.* (2016a) and offers therefore good statistical performances;
(c) it offers great flexibility on the log-hazard scale through penalized splines and tensor product splines without requiring time-consuming MCMC, smoothing splines or Poisson approximation techniques;
(d) it offers a common and straightforward setting for hazard and excess hazard regression.

The methods that are detailed in this paper (including random-effects models) were implemented in R package `survPen` that is available in the Comprehensive R Archive Network repository (`https://CRAN.R-project.org/package=survpen`).

Acknowledgements

This research was conducted as part of the first author's doctoral thesis supported by the French Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation. The authors thank the Agence Nationale de la Recherche for supporting this study of the Challenges in the Estimation of Net Survival group (grant ANR-12-BSV1-0028). This research was also carried out within the context of a four-institute cancer surveillance programme partnership involving the Institut National du Cancer, Santé Publique France, the French Network of Cancer Registries and Hospices Civils de Lyon through a grant from the Institut National du Cancer (attributive decision 2016-131). The authors are also grateful to Guy Launoy (Centre Hospitalier Universitaire de Caen, Caen, France) and his team for the access to the EDI data. They also thank Jean Iwaz for his thorough checking of the manuscript and Jacques Estève for his valuable advice, as well as the French Network of Cancer Registries for providing high quality cancer data.

References

Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Niksic, M., Bonaventure, A., Valkov, M., Johnson, C. J., Estève, J., Ogunbiyi, O. J., Azevedo, E. S. G., Chen, W. Q., Eser, S., Engholm, G., Stiller, C. A., Monnereau, A., Woods, R. R., Visser, O., Lim, G. H., Aitken, J., Weir, H. K., Coleman, M. P. and CONCORD Working Group (2018) Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37513025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet*, **391**, 1023–1075.

Antunes, L., Mendonca, D., Bento, M. J. and Rachet, B. (2016) No inequalities in survival from colorectal cancer by education and socioeconomic deprivation—a population-based study in the North Region of Portugal, 2000-2002. *BMC Cancer*, **16**, article 608.

Becher, H., Kauermann, G., Khomski, P. and Kouyate, B. (2009) Using penalized splines to model age- and season-of-birth-dependent effects of childhood mortality risk factors in rural Burkina Faso. *Biometr. J.*, **51**, 110–122.

Brezger, A., Kneib, T. and Lang, S. (2003) BayesX: analysing Bayesian structured additive regression models (No. 332). *Discussion Paper*. Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München, Munich. (Available from `https://www.econstor.eu/bitstream/10419/31014/1/483860263.PDF`.)

Bryere, J., Dejardin, O., Launay, L., Colonna, M., Grosclaude, P., Launoy, G. and French Network of Cancer Registries FRANCIM (2018) Socioeconomic status and site-specific cancer incidence, a Bayesian approach in a French Cancer Registries Network study. *Eur. J. Cancer. Prevn*, **27**, 391–398.

Charvat, H., Remontet, L., Bossard, N., Roche, L., Dejardin, O., Rachet, B., Launoy, G., Belot, A. and CENSUR Working Survival Group (2016) A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statist. Med.*, **35**, 3066–3084.

Commenges, D., Joly, P., Gégout-Petit, A. and Liquet, B. (2007) Choice between semi-parametric estimators of Markov and non-Markov multi-state models from coarsened observations. *Scand. J. Statist.*, **34**, 33–52.

Cowppli-Bony, A., Uhry, Z., Remontet, L., Voirin, N., Guizard, A. V., Tretarre, B., Bouvier, A. M., Colonna, M., Bossard, N., Woronoff, A. S., Grosclaude, P. and French Network of Cancer Registries (FRANCIM) (2017)

Survival of solid cancer patients in France, 1989-2013: a population-based study. *Eur. J. Cancer Prevn*, **26**, 461–468.
Danieli, C., Remontet, L., Bossard, N., Roche, L. and Belot, A. (2012) Estimating net survival: the importance of allowing for informative censoring. *Statist. Med.*, **31**, 775–786.
Estève, J., Benhamou, E., Croasdale, M. and Raymond, L. (1990) Relative survival and the estimation of net survival: elements for further discussion. *Statist. Med.*, **9**, 529–538.
Friedman, M. (1982) Piecewise exponential models for survival data with covariates. *Ann. Statist.*, **10**, 101–113.
Gray, R. J. (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Am. Statist. Ass.*, **87**, 942–951.
Greven, S. and Scheipl, F. (2016) Comment. *J. Am. Statist. Ass.*, **111**, 1568–1573.
Guillaume, E., Pornet, C., Dejardin, O., Launay, L., Lillini, R., Vercelli, M., Mari-Dell'olmo, M., Fernandez Fontelo, A., Borrell, C., Ribeiro, A. I., Pina, M. F., Mayer, A., Delpierre, C., Rachet, B. and Launoy, G. (2016) Development of a cross-cultural deprivation index in five European countries. *J Epidem. Commty Hlth*, **70**, 493–499.
Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2006) Geoadditive survival models. *J. Am. Statist. Ass.*, **101**, 1065–1075.
Hennerfeind, A., Held, L. and Sauleau, E. A. (2008) A Bayesian analysis of relative cancer survival with geoadditive models. *Statist. Modllng*, **8**, 117–139.
Kauermann, G. (2005) Penalized spline smoothing in multivariable survival models with varying coefficients. *Computnl Statist. Data Anal.*, **49**, 169–186.
Kneib, T. and Fahrmeir, L. (2007) A mixed model approach for geoadditive hazard regression. *Scand. J. Statist.*, **34**, 207–228.
Liquet, B. and Commenges, D. (2004) Estimating the expectation of the log-likelihood with censored data for estimator selection. *Liftim. Data Anal.*, **10**, 351–367.
Liu, X.-R., Pawitan, Y. and Clements, M. (2018) Parametric and penalized generalized survival models. *Statist. Meth. Med. Res.*, **27**, 1531–1546.
Marra, G. and Wood, S. N. (2011) Practical variable selection for generalized additive models. *Computnl Statist. Data Anal.*, **55**, 2372–2387.
Marra, G. and Wood, S. N. (2012) Coverage properties of confidence intervals for generalized additive model components. *Scand. J. Statist.*, **39**, 53–74.
Martino, S., Akerkar, R. and Rue, H. (2011) Approximate Bayesian inference for survival models. *Scand. J. Statist.*, **38**, 514–528.
Monnereau, A., Uhry, Z., Bossard, N., Cowppli-Bony, A., Voirin, N., Delafosse, P., Remontet, L., Troussard, X. and Maynadié, X. (2016) Survie des personnes atteintes de cancer en France, 1989-2013: étude à partir des registres des cancers du réseau Francim, Partie 2—Hémopathies malignes. Institut de Veille Sanitaire, Saint-Maurice. (Available from `http://invs.santepubliquefrance.fr/fr../layout/set/print/Publications-et-outils/Rapports-et-syntheses/Maladies-chroniques-et-traumatismes/2016/Survie-des-personnes-atteintes-de-cancer-en-France-metropolitaine-1989-2013-Partie-2-hemopathies-malignes`.)
Nychka, D. (1988) Bayesian confidence intervals for smoothing splines. *J. Am. Statist. Ass.*, **83**, 1134–1143.
O'Sullivan, F. (1988) Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Comput.*, **9**, 363–379.
Perme, M. P., Stare, J. and Estève, J. (2012) On estimation in relative survival. *Biometrics*, **68**, 113–120.
Petersen, K. B. and Pedersen, M. S. (2008) The matrix cookbook. Technical University of Denmark, Lyngby. (Available from `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`.)
R Core Team (2018) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Reiss, P. T. and Ogden, R. T. (2009) Smoothing parameter selection for a class of semiparametric linear models. *J. R. Statist. Soc.* B, **71**, 505–523.
Remontet, L., Bossard, N., Belot, A., Estève, J. and French Network of Cancer Registries FRANCIM (2007) An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statist. Med.*, **26**, 2214–2228.
Remontet, L., Uhry, Z., Bossard, N., Iwaz, J., Belot, A., Danieli, C., Charvat, H., Roche, L. and CENSUR Working Survival Group (2018) Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables: performance of this multidimensional penalized spline approach in net survival trend analysis. *Statist. Meth. Med. Res.*, to be published, doi 10.1177/0962280218779408.
Rodríguez-Girondo, M., Kneib, T., Cadarso-Suárez, C. and Abu-Assi, E. (2013) Model building in nonproportional hazard regression. *Statist. Med.*, **32**, 5301–5314.
Rondeau, V., Commenges, D. and Joly, P. (2003) Maximum penalized likelihood estimation in a gamma-frailty model. *Liftim. Data Anal.*, **9**, 139–153.
Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V. and Soubeyran, P. (2007) Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, **8**, 708–721.

Royston, P. and Sauerbrei, W. (2008) *Multivariable Model-building: a Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables*. Chichester: Wiley.
Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.
Sauerbrei, W., Royston, P. and Look, M. (2007) A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometr. J.*, **49**, 453–473.
Uhry, Z., Bossard, N., Remontet, L., Iwaz, J., Roche, L. and GRELL EUROCARE-5 Working Group and the CENSUR Working Survival Group (2017) New insights into survival trend analyses in cancer population-based studies: the SUDCAN methodology. *Eur. J. Cancer Prevn*, **26**, suppl., S9–S15.
Umlauf, N., Adler, D., Kneib, T., Lang, S. and Zeileis, A. (2015) Structured additive regression models: an R interface to BayesX. *J. Statist. Softwr.*, **63**, 1–46.
Umlauf, N., Klein, N. and Zeileis, A. (2018) BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *J. Computnl Graph. Statist.*, **27**, 612–627.
Verweij, P. J. and Van Houwelingen, H. C. (1993) Cross-validation in survival analysis. *Statist. Med.*, **12**, 2305–2314.
Wahba, G. (1983) Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.* B, **45**, 133–150.
Wahba, G. (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, **13**, 1378–1402.
Wood, S. N. (2006a) On confidence intervals for generalized additive models based on penalized regression splines. *Aust. New Zeal. J. Statist.*, **48**, 445–464.
Wood, S. N. (2006b) Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, **62**, 1025–1036.
Wood, S. N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Statist. Soc.* B, **73**, 3–36.
Wood, S. N. (2017) *Generalized Additive Models: an Introduction with R*. Boca Raton: Chapman and Hall–CRC.
Wood, S. N. and Augustin, N. H. (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecol. Modllng*, **157**, 157–177.
Wood, S. N., Pya, N. and Säfken, B. (2016a) Smoothing parameter and model selection for general smooth models. *J. Am. Statist. Ass.*, **111**, 1548–1563.
Wood, S. N., Pya, N. and Säfken, B. (2016b) Rejoinder. *J. Am. Statist. Ass.*, **111**, 1573–1575.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material: Multidimensional penalized hazard model with continuous covariates: application for studying trends and social inequalities in cancer survival'.

Journal of the Royal Statistical Society: Series C (Applied Statistics)

Supplementary Material for

Multidimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival

Mathieu Fauvernier, Laurent Roche, Zoé Uhry, Laure Tron, Nadine Bossard, Laurent Remontet and the CENSUR Working Survival Group.

Corresponding author:
Mathieu FAUVERNIER
mathieu.fauvernier@chu-lyon.fr

A. Derivatives of the LAML criterion

According to Wood et al. (2016a), the derivatives of LAML are:

$$\frac{\partial LAML}{\partial \rho_l} = -\frac{\lambda_l}{2}\widehat{\boldsymbol{\beta}}^T \boldsymbol{S}^l \widehat{\boldsymbol{\beta}} + \frac{1}{2}\frac{\partial log\left|\boldsymbol{S}^\lambda\right|_+}{\partial \rho_l} - \frac{1}{2}\frac{\partial log|\boldsymbol{\mathcal{H}}|}{\partial \rho_l}$$

and

$$\frac{\partial^2 LAML}{\partial \rho_l \partial \rho_m} = -\kappa_l^m \frac{\lambda_l}{2}\widehat{\boldsymbol{\beta}}^T \boldsymbol{S}^l \widehat{\boldsymbol{\beta}} - \frac{\partial \widehat{\boldsymbol{\beta}}^T}{\partial \rho_m}\boldsymbol{\mathcal{H}}\frac{\partial \widehat{\boldsymbol{\beta}}}{\partial \rho_l} + \frac{1}{2}\frac{\partial^2 log\left|\boldsymbol{S}^\lambda\right|_+}{\partial \rho_l \partial \rho_m} - \frac{1}{2}\frac{\partial^2 log|\boldsymbol{\mathcal{H}}|}{\partial \rho_l \partial \rho_m}$$

where $\kappa_l^m = 1$ when $l = m$, 0 otherwise. The derivatives of the log determinant of $\boldsymbol{\mathcal{H}}$ (Wood et al. 2016a) are the following:

$$\frac{\partial log|\boldsymbol{\mathcal{H}}|}{\partial \rho_l} = tr\left\{\boldsymbol{\mathcal{H}}^{-1}\frac{\partial \boldsymbol{\mathcal{H}}}{\partial \rho_l}\right\}$$

and

$$\frac{\partial^2 log|\boldsymbol{\mathcal{H}}|}{\partial \rho_l \partial \rho_m} = -tr\left\{\boldsymbol{\mathcal{H}}^{-1}\frac{\partial \boldsymbol{\mathcal{H}}}{\partial \rho_l}\boldsymbol{\mathcal{H}}^{-1}\frac{\partial \boldsymbol{\mathcal{H}}}{\partial \rho_m}\right\} + tr\left\{\boldsymbol{\mathcal{H}}^{-1}\frac{\partial^2 \boldsymbol{\mathcal{H}}}{\partial \rho_l \partial \rho_m}\right\}$$

Concerning the derivatives of $log\left|\boldsymbol{S}^\lambda\right|_+$, one may partially follow Wood et al. (2016a) and form the symmetric eigen-decomposition $\sum_{j=1}^M \frac{\boldsymbol{S}^j}{\|\boldsymbol{S}^j\|_F} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{U}}^T$. Noting $\boldsymbol{U}_+$ the columns of $\widetilde{\boldsymbol{U}}$ that correspond to positive eigenvalues leads to $\tilde{\boldsymbol{S}}^j = \boldsymbol{U}_+{}^{\boldsymbol{T}}\boldsymbol{S}^j\boldsymbol{U}_+$. Then $\left|\boldsymbol{S}^\lambda\right|_+ = \left|\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right|$ with $\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j$ a full-rank matrix. Actually, matrix $\widetilde{\boldsymbol{U}}$ is used for an initial reparameterization performed before the optimization process. The reparameterization is reversed at convergence. Concerning the stable evaluation of $log\left|\boldsymbol{S}^\lambda\right|_+$ and its derivatives, one may not follow the complex procedure proposed by Wood (2011, Appendix B). Indeed, once the reparameterization is done, one may use the following QR-decomposition: $\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j = \widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{R}}$. Then $log\left|\boldsymbol{S}^\lambda\right|_+$ is just the sum of the absolute values of the diagonal elements of $\widetilde{\boldsymbol{R}}$ and the derivatives require only $\left(\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right)^{-1}$ which is computed through LU decomposition. The derivatives of $log\left|\boldsymbol{S}^\lambda\right|_+$ are given by:

$$\frac{\partial log\left|\boldsymbol{S}^\lambda\right|_+}{\partial \rho_l} = \frac{\partial log\left|\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right|}{\partial \rho_l} = tr\left\{\left(\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right)^{-1}\frac{\partial\left(\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right)}{\partial \rho_l}\right\}$$

$$= tr\left\{\left(\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right)^{-1}\lambda_l \tilde{\boldsymbol{S}}^l\right\}$$

and

$$\frac{\partial^2 log\left|\boldsymbol{S}^\lambda\right|_+}{\partial \rho_l \partial \rho_m} = -tr\left\{\left(\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right)^{-1}\frac{\partial\left(\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right)}{\partial \rho_l}\left(\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right)^{-1}\frac{\partial\left(\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right)}{\partial \rho_m}\right\}$$

$$+ tr\left\{\left(\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right)^{-1}\frac{\partial^2\left(\sum_{j=1}^M \lambda_j \tilde{\boldsymbol{S}}^j\right)}{\partial \rho_l \partial \rho_m}\right\}$$

$$= -tr\left\{\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\lambda_l\tilde{\boldsymbol{S}}^l\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\lambda_m\tilde{\boldsymbol{S}}^m\right\} + tr\left\{\left(\sum_{j=1}^{M}\lambda_j\tilde{\boldsymbol{S}}^j\right)^{-1}\kappa_l^m\lambda_l\tilde{\boldsymbol{S}}^l\right\}$$

In the survival context, the derivatives of $\boldsymbol{H}$ and $\boldsymbol{\mathcal{H}}$ are**:**

$$\frac{\partial \boldsymbol{H}}{\partial \rho_l} = \sum_{k=1}^{q}(\boldsymbol{GL}^k)^T\left[\boldsymbol{w}^k \circ \boldsymbol{GL}^k \circ \left\{\left(\boldsymbol{GL}^k \circ exp(\boldsymbol{GL}^k\widehat{\boldsymbol{\beta}})\right)\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\rho_l}\right\}\right]$$

$$\frac{\partial \boldsymbol{\mathcal{H}}}{\partial \rho_l} = \frac{\partial \boldsymbol{H}}{\partial \rho_l} + \lambda_l\boldsymbol{S}^l$$

and

$$\frac{\partial^2 \boldsymbol{H}}{\partial\rho_l\partial\rho_m} = \sum_{k=1}^{q}(\boldsymbol{GL}^k)^T\left[\boldsymbol{w}^k \circ \boldsymbol{GL}^k \circ\left\{\left(\boldsymbol{GL}^k \circ \left[\left(\boldsymbol{GL}^k \circ exp(\boldsymbol{GL}^k\widehat{\boldsymbol{\beta}})\right)\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\rho_m}\right]\right)\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\rho_l} + \left(\boldsymbol{GL}^k \circ exp(\boldsymbol{GL}^k\widehat{\boldsymbol{\beta}})\right)\frac{\partial^2\widehat{\boldsymbol{\beta}}}{\partial\rho_l\partial\rho_m}\right\}\right]$$

$$\frac{\partial^2 \boldsymbol{\mathcal{H}}}{\partial\rho_l\partial\rho_m} = \frac{\partial^2 \boldsymbol{H}}{\partial\rho_l\partial\rho_m} + \kappa_l^m\lambda_l\boldsymbol{S}^l$$

The derivatives of $\widehat{\boldsymbol{\beta}}$ are obtained via implicit differentiation (Wood et al. 2016a). For convenience, their matrix notations are:

$$\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\rho_l} = -\boldsymbol{\mathcal{H}}^{-1}(\lambda_l\boldsymbol{S}^l)\widehat{\boldsymbol{\beta}}$$

$$\frac{\partial^2\widehat{\boldsymbol{\beta}}}{\partial\rho_l\partial\rho_m} = -\left[\frac{\partial\boldsymbol{\mathcal{H}}^{-1}}{\partial\rho_m}(\lambda_l\boldsymbol{S}^l) + \boldsymbol{\mathcal{H}}^{-1}(\lambda_l\boldsymbol{S}^l)\kappa_l^m\right]\widehat{\boldsymbol{\beta}} - [\boldsymbol{\mathcal{H}}^{-1}(\lambda_l\boldsymbol{S}^l)]\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\rho_m}$$

where the first derivative of $\boldsymbol{\mathcal{H}}^{-1}$ is obtained via classical matrix operations:

$$\frac{\partial\boldsymbol{\mathcal{H}}^{-1}}{\partial\rho_m} = -\boldsymbol{\mathcal{H}}^{-1}\frac{\partial\boldsymbol{\mathcal{H}}}{\partial\rho_m}\boldsymbol{\mathcal{H}}^{-1}$$

B. Derivatives for the penalized excess hazard model

For ease of reading, we write $r_i = h_P(a_i + t_i, \boldsymbol{z}_i)$. In an excess hazard context, the contribution to the log-likelihood of an individual i is:

$$\mathrm{l_i}(\boldsymbol{\beta}) = \delta_i\log[h_E(t_i, \boldsymbol{x}_i) + r_i] - \int_0^{t_i} h_E(u, \boldsymbol{x}_i)du$$

Approximating the integral via Gauss-Legendre quadrature we obtain:

$$\mathrm{l_i}(\boldsymbol{\beta}) \approx \delta_i\log[exp(\boldsymbol{X}_i\boldsymbol{\beta}) + r_i] - \sum_{k=1}^{q} w_i^k exp(\boldsymbol{GL}_i^k\boldsymbol{\beta})$$

The first derivative of $\mathrm{l_i}$ with respect to β_l is:

$$\frac{\partial \mathrm{l_i}(\boldsymbol{\beta})}{\partial \beta_l} = \frac{\delta_\mathrm{i} X_{il} exp(\boldsymbol{X}_i\boldsymbol{\beta})}{exp(\boldsymbol{X}_i\boldsymbol{\beta}) + r_i} - \sum_{k=1}^{q} w_i^k GL_{il}^k exp(\boldsymbol{GL}_i^k\boldsymbol{\beta})$$

and the second derivative is then:

$$\frac{\partial^2 \mathrm{l_i}(\boldsymbol{\beta})}{\partial \beta_l \partial \beta_m} = \frac{r_i \delta_\mathrm{i} X_{il} X_{im} exp(\boldsymbol{X}_i\boldsymbol{\beta})}{(exp(\boldsymbol{X}_i\boldsymbol{\beta}) + r_i)^2} - \sum_{k=1}^{q} w_i^k GL_{il}^k GL_{im}^k exp(\boldsymbol{GL}_i^k\boldsymbol{\beta})$$

N.B.: In this section, we need to divide matrices by vectors and use exponentiation of vectors. When $\boldsymbol{v}$ is a n-vector, $\boldsymbol{M}$ a n-by-p matrix, and y a real number: $\left(\frac{\boldsymbol{M}}{\boldsymbol{v}}\right)_{i,j} = \frac{M_{i,j}}{v_i}$ and $(\boldsymbol{v}^y)_i = (v_i)^y$.

As in the overall survival context, $\boldsymbol{H}$ may be expressed in matrix notation:

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\boldsymbol{X}^T\left[\frac{\boldsymbol{r}\circ\boldsymbol{\delta}\circ\boldsymbol{X}\circ\exp(\boldsymbol{X\beta})}{(\exp(\boldsymbol{X\beta})+\boldsymbol{r})^2}\right] + \sum_{k=1}^{q}(\boldsymbol{GL}^k)^T[\boldsymbol{w}^k\circ\boldsymbol{GL}^k\circ exp(\boldsymbol{GL}^k\boldsymbol{\beta})]$$

We note $\boldsymbol{H_1}(\boldsymbol{\beta}) = -\boldsymbol{X}^T\left[\frac{\boldsymbol{r}\circ\boldsymbol{\delta}\circ\boldsymbol{X}\circ\exp(\boldsymbol{X\beta})}{(\exp(\boldsymbol{X\beta})+\boldsymbol{r})^2}\right]$ and $\boldsymbol{H_2}(\boldsymbol{\beta}) = \sum_{k=1}^{q}(\boldsymbol{GL}^k)^T[\boldsymbol{w}^k\circ\boldsymbol{GL}^k\circ exp(\boldsymbol{GL}^k\boldsymbol{\beta})]$. To select properly the smoothing parameters in an excess hazard context, we finally need the derivatives of $\boldsymbol{H}(\widehat{\boldsymbol{\beta}})$ with respect to the smoothing parameters. Actually, the derivatives of $\boldsymbol{H_2}(\widehat{\boldsymbol{\beta}})$ are already known because they correspond to the ones obtained in the overall survival context. Thus, only the derivatives of $\boldsymbol{H_1}(\widehat{\boldsymbol{\beta}})$, referred to as $\boldsymbol{H_1}$, are needed:

$$\frac{\partial \boldsymbol{H_1}}{\partial \rho_l} = -\boldsymbol{X}^T\left[\boldsymbol{r}\circ\boldsymbol{\delta}\circ\boldsymbol{X}\circ\left\{\left(\frac{\boldsymbol{X}\circ\exp(\boldsymbol{X\widehat{\beta}})\circ(\boldsymbol{r}-\exp(\boldsymbol{X\widehat{\beta}}))}{(\exp(\boldsymbol{X\widehat{\beta}})+\boldsymbol{r})^3}\right)\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\rho_l}\right\}\right]$$

and

$$\begin{aligned}\frac{\partial^2 \boldsymbol{H_1}}{\partial\rho_l\partial\rho_m} = -\boldsymbol{X}^T\Bigg[&\boldsymbol{r}\circ\boldsymbol{\delta}\circ\boldsymbol{X}\\ &\circ\left\{\left(\frac{\boldsymbol{X}\circ\boldsymbol{X}\circ\exp(\boldsymbol{X\widehat{\beta}})\circ\left[\exp(\boldsymbol{X\widehat{\beta}})\circ\exp(\boldsymbol{X\widehat{\beta}}) - \boldsymbol{4r}\circ\exp(\boldsymbol{X\widehat{\beta}}) + \boldsymbol{r}^2\right]}{(\exp(\boldsymbol{X\widehat{\beta}})+\boldsymbol{r})^4}\right)\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\rho_l}\right.\\ &\left.+\left(\frac{\boldsymbol{X}\circ\exp(\boldsymbol{X\widehat{\beta}})\circ(\boldsymbol{r}-\exp(\boldsymbol{X\widehat{\beta}}))}{(\exp(\boldsymbol{X\widehat{\beta}})+\boldsymbol{r})^3}\right)\frac{\partial^2\widehat{\boldsymbol{\beta}}}{\partial\rho_l\partial\rho_m}\right\}\Bigg]\end{aligned}$$

C. Comparisons between survPen package and other methods

C.1 Comparison in terms of computational time

This section aims to compare the computational times of R packages that implement penalized survival models: *frailtypack, R2BayesX, bamlss, gss, rstpm2*, and *survPen*. To this list, we have also added under denomination "Poisson + *gam*" the Poisson approach for survival model described by Remontet et al. (2018). In this approach a Poisson regression model is fitted on split data using the *gam* function from the *mgcv* package (the link function being modified to include expected rates).
Table 1 presents a brief description of model functionalities in terms of penalized splines. The columns indicate whether the function or method allows: 1) fitting penalized splines for the effect of a continuous covariate (column "Penalized spline"); 2) fitting penalized tensor product splines to model the effects and interactions of several continuous covariates ("Tensor"); 3) fitting excess hazard models ("Excess hazard"); 4) fitting regression splines as opposed to smoothing splines ("Regression splines"); and 5) modelling directly the hazard or log-hazard ("Hazard").

Table 1. Functionalities of various packages in terms of penalized splines for survival model

			Functionality [1]				
Method	R package	Function	Penalized spline	Tensor	Excess hazard	Regression splines	Hazard
frailtypack	yes	*frailtyPenal*	only for time	no	no	yes	yes
R2BayesX	yes	*bayesx*	yes	no	no	yes	yes
bamlss	yes	*bamlss*	yes	yes	no	yes	yes
gss	yes	*sshzd*	yes	yes	no	no [2]	yes
rstpm2	yes	*pstpm2*	yes	yes	yes	yes	no [3]
Poisson + *gam*	no [4]	-	yes	yes	yes	yes	yes
survPen	yes	*survPen*	yes	yes	yes	yes	yes

[1] This table is restricted to a comparison of functionalities versus those offered by package *survPen* (multidimensional penalized regression splines for hazard and excess hazard models); it does not provide an exhaustive description of each package.

[2] The *gss* package uses smoothing splines.

[3] The *rstpm2* package allows modelling on several scales, including the log-cumulative hazard, but not directly on the hazard or log-hazard.

[4] The "Poisson + *gam*" method is not available as a package but the R code is available on github (https://github.com/RocheLHCL/SMMR_Remontet2018/).

Briefly, *frailtypack* is useful to fit shared, joint, and nested frailty models; *R2BayesX* and *bamlss* are useful to fit complex additive models, for example by modelling jointly the location and shape of a distribution (not restricted to survival models); *gss* offers fitting general smoothing splines not confined to survival models; *rstpm2* proposes different modelling scales for survival models and frailty models; and, finally, *mgcv* is the reference in terms of penalized regression splines with generalized linear models and likelihoods beyond the exponential family (Cox model, for example).

To compare computational times between these packages, three survival models and one net survival model were considered (the first model is actually declined in two versions):

$$\text{model "spline\_time"} : \log\{f(t)\} = spline(t, df = 10)$$
$$\text{model "spline\_time", } frailtypack \text{ version: } f(t) = spline(t, df = 10)$$

$$\text{model "tensor2"} : \log\{f(t,a)\} = tensor(t, a, df = c(5,5))$$

$$\text{model "tensor3"} : \log\{f(t,a,y)\} = tensor(t, a, y, df = c(5,5,5))$$

$$\text{model "tensor3\_net"} : \log\{f_E(t,a,y)\} = tensor(t, a, y, df = c(5,5,5))$$

In these models, t is the follow-up time, a is the age at diagnosis, and y is the year of diagnosis. Function f represents the hazard function (the excess hazard for f_E) in all packages but *rstpm2* when f represents the cumulative hazard function (excess cumulative hazard for f_E). The first model is associated with 10 regression parameters and 1 smoothing parameter. The second model is associated with 25 regression parameters and 2 smoothing parameters. Each of the third and the fourth model is associated with 125 regression parameters and 3 smoothing parameters. The number of smoothing parameters differs for *gss* because it relies on functional ANOVA decomposition (Wood 2017, section 5.6.3).

All models were fitted on two simulated datasets (sizes: 2,000 and 20,000) taken from the simulations of Cervix scenario (N=2,000) described in Remontet et al. (2018). The dataset of size 2,000 is randomly chosen (it corresponds to the *survPen* package dataset called "datCancer") and the dataset of size 20,000 is the concatenation of the 10 first simulated datasets. Given the functionalities described in Table 1, the four models cannot be fitted with each of the seven methods. For example, the fourth tensor3_net can be fitted only by *rstpm2, survPen* and the "Poisson + *gam*" method.

The initial data splitting procedure needed before using *"*Poisson + *gam"* transforms the 2,000 and 20,000 individuals into 93,963 and 934,379 pseudo-observations, respectively.

Concerning *R2BayesX* package, we considered the REML version (based on Kneib and Fahrmeir 2007) and the MCMC version. As with *survPen*, we used the LCV criterion to allow comparisons with *rstpm2.*

We used R version 3.5.2 with the following versions of the packages: *frailtypack*_3.0.2.1, *R2BayesX*_1.1-1, *bamlss*_1.0-1, *gss*_2.1-9, *rstpm2*_1.4.5, *mgcv*_1.8-26 and *survPen*_1.0.1 Tables 2 and 3 show the computational times performed by an Intel Xeon CPU E3-1245 v5 3.50 GHz with 16 Go RAM.

Table 2. Computational times (in seconds) with N=2,000.

	Model			
Method	Spline_time	Tensor 2	Tensor 3	Tensor_net3
frailtypack	0.7			
R2BayesX REML	1.2			
R2BayesX MCMC	37.0			
bamlss	353.0	1,378.0	10,851.7	
gss	0.0	4.6	5,608.3	
rstpm2	1.2	19.1	562.7	456.2
Poisson + *gam*	1.3	6.8	232.5	463.3
survPen	0.5	3.6	55.4	75.6

Table 3. Computational times (in seconds) with N=20,000.

	Model			
Method	Spline_time	Tensor 2	Tensor 3	Tensor_net3
frailtypack	19.0			
R2BayesX REML	2.8			
R2BayesX MCMC	525.6			
bamlss	4,828.3	13,751.9	109,708.2	
gss	0.2	16.2	4,712.1	
rstpm2	38.1	37.2	4,270.6	4,197.7
Poisson + *gam*	12.4	91.3	1,755.4	4,270.7
survPen	5.2	23.4	576.3	680.1

With both sample sizes, *survPen* package was the fastest in the most complex settings (tensor3 and tensor3_net) but *R2BayesX* in its REML version and *gss* were faster with the first two models and sample size N=20,000. The absolute gain in computational time with *survPen* in the most complex settings is substantial. *R2BayesX* (MCMC version) and *bamlss* that rely on MCMC algorithms were unsurprisingly more time-consuming than the other approaches. *bamlss* was by far the slowest.

Despite the use of an efficient approximation for smoothing splines fitting (Du and Gu 2006, Gu 2014), we observed that the smoothing splines of the *gss* package are more time-consuming than the penalized regression splines used by *survPen* in the most complex settings. However, the models fitted here by *survPen* and *gss* are not fully comparable because *gss* relies on functional ANOVA decomposition. Finally, the *gss* models were also fitted with function *sshzd1* which is a faster version of *sshzd* but performance degradation was too important (not shown).

C.2 Comparison with *rstpm2* in terms of statistical performance

As the *rstpm2* package was the only other method that could, theoretically, fit the most complex model (tensor3_net), it was necessary to compare the statistical performance of *rstpm2* versus *survPen* in terms of excess hazard RMISE, net survival bias, and net survival RMSE (only with sample size 2,000 for computational time reasons). The results obtained with *survPen* are shown in Figure 2 of the article. With *rstmp2,* in both scenarios (Cervix uteri and Oesophagus), we fitted a log-cumulative excess hazard model on each of the 1,000 simulated datasets. The model was a penalized tensor product spline of log-time, age, and year and *pstpm2* was run with the same bases (natural cubic splines) and the same number of knots as in the simulation study (6 for time, 5 for age, and 4 for year). The locations of the knots were the default values of *pstpm2* (percentiles). The results with *pstpm2* are shown in Figures 1 and 2 below.

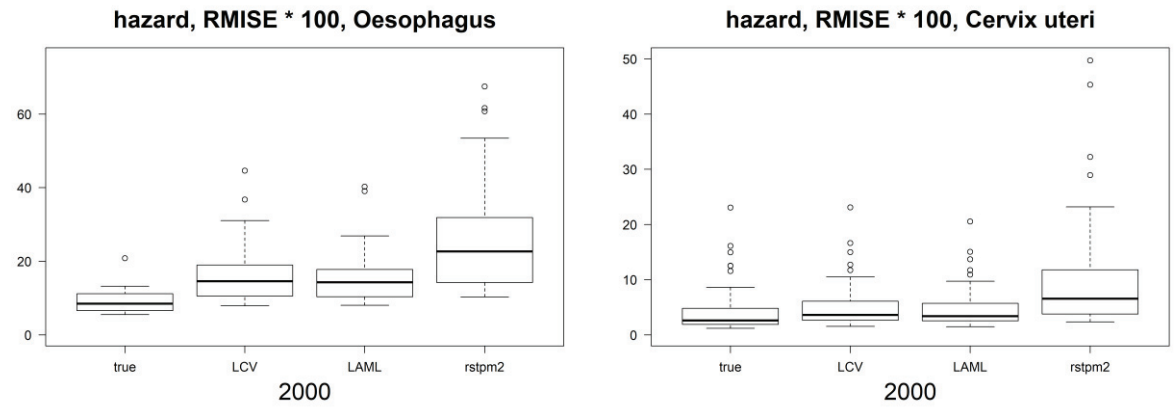

Figure 1: Boxplots of excess hazard RMISE multiplied by 100 in both scenarios with 2,000 sample size. "true" corresponds to the predictions of the fitted true model, "LCV" stands for survPen with LCV smoothing parameter estimation, "LAML" stands for survPen with LAML smoothing parameter estimation, and "rstpm2" stands for the pstpm2 function.

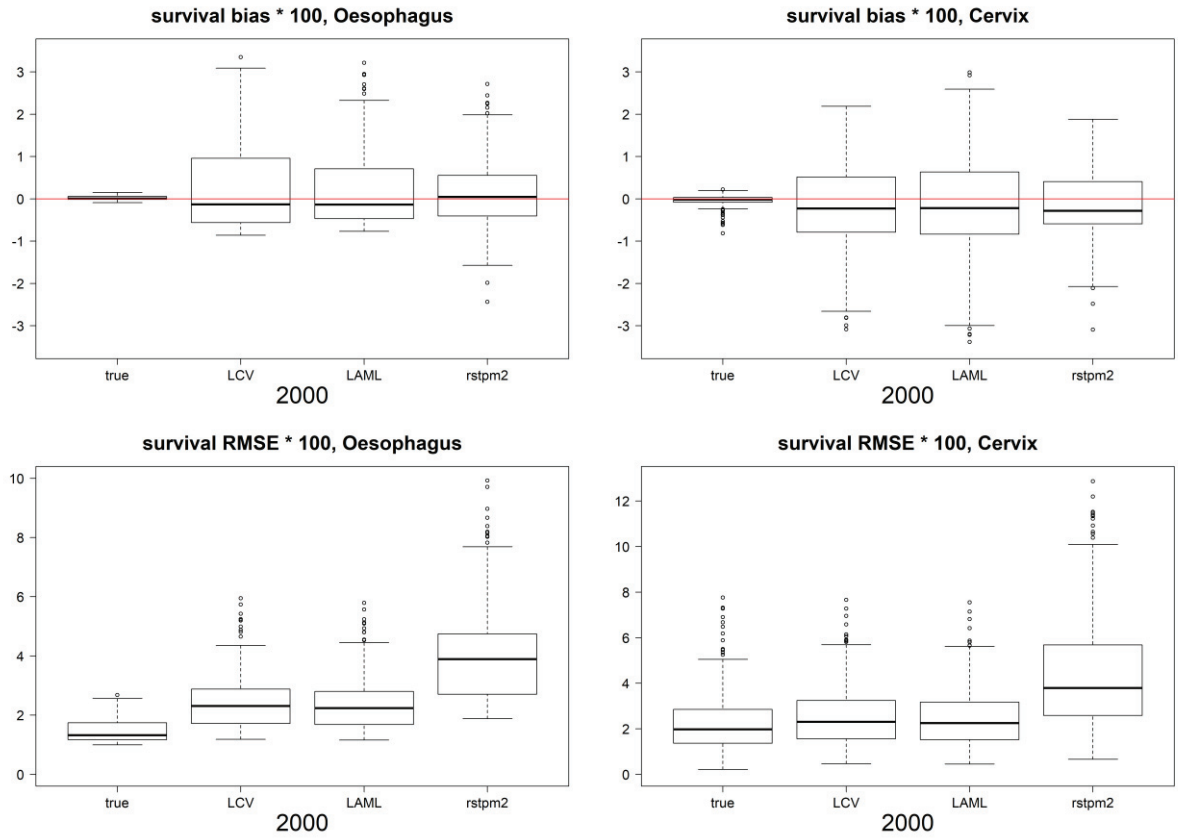

Figure 2: Boxplots of net survival bias and RMSE multiplied by 100 in both scenarios with 2,000 sample size. "true" corresponds to the predictions of the fitted true model, "LCV" stands for survPen with LCV smoothing parameter estimation, "LAML" stands for survPen with LAML smoothing parameter estimation and "rstpm2" stands for the pstpm2 function.

In the Cervix scenario, *rstpm2* showed slightly less bias than *survPen*. However, in both scenarios, *survPen* was superior to *rstpm2* in terms of excess hazard RMISE and net survival RMSE. In the Oesophagus scenario, the RMISE from *survPen* LCV was inferior to the

RMISE from *rstpm2* in 81% of age-year combinations. In the Cervix scenario, this proportion reached 84%. Besides, 44 Oesophagus models and 11 Cervix models failed to converge with *rstpm2* while only one *survPen* model failed to converge (Cervix scenario, LCV criterion). It appears then that our method outperforms *rstpm2* in terms of statistical performance and reliability in the two studied scenarios.

References

Du, P., and Gu, C. (2006). Penalized likelihood hazard estimation: efficient approximation and Bayesian confidence intervals. *Stat Probab Lett*, **76**, 244-254.

Gu, C. (2014). Smoothing spline ANOVA models: R package gss. *J Stat Softw*, **58**, 1-25.

Kneib, T. and Fahrmeir, L. (2007) A mixed model approach for geoadditive hazard regression. *Scand J Statist,* **34**, 207-228.

Remontet, L., Uhry, Z., Bossard, N., Iwaz, J., Belot, A., Danieli, C., Charvat, H., Roche, L. and CENSUR Working Survival Group (2018) Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables: Performance of this multidimensional penalized spline approach in net survival trend analysis. *Stat Methods Med Res*, 0962280218779408.

Wood, S. N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Statist Soc Series B Stat Methodol,* **73**, 3-36.

Wood, S. N. (2017) *Generalized additive models: an introduction with R*: Chapman and Hall/CRC.

Wood, S. N., Pya, N. and Säfken, B. (2016a) Smoothing parameter and model selection for general smooth models. *J Am Stat Assoc,* **111**, 1548-1563.

D Article *survPen* publié dans JOSS

The Journal of Open Source Software

survPen: an **R** package for hazard and excess hazard modelling with multidimensional penalized splines

Mathieu Fauvernier[1, 2], **Laurent Remontet**[1, 2], **Zoé Uhry**[1, 2, 3], **Nadine Bossard**[1, 2], **and Laurent Roche**[1, 2]

1 Hospices Civils de Lyon, Pôle Santé Publique, Service de Biostatistique - Bioinformatique, Lyon, France **2** Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique-Santé, Villeurbanne, France **3** Département des Maladies Non-Transmissibles et des Traumatismes, Santé Publique France, Saint-Maurice, France

DOI: 10.21105/joss.01434

Software

- Review
- Repository
- Archive

Submitted: 03 May 2019
Published: 23 August 2019

Background

Survival analysis deals with studying the elapsed time until an event occurs. When the event of interest is death, it aims at describing the survival probability and its corresponding mortality hazard. In epidemiology, as patients may die from their disease or from other causes, it is relevant to study the mortality due to their disease; also called "excess mortality". This excess mortality is useful to make comparisons between different countries and time periods (Allemani et al., 2018; Uhry et al., 2017) and is directly linked to the concept of net survival (Perme, Stare, & Estève, 2012), i.e. the survival that would be observed if patients could only die from their disease.

`survPen` is an R package that implements flexible regression models for (net) survival analysis. Model specification is carried out on the logarithm of the (excess) hazard scale. `survPen` provides an efficient procedure to estimate the model parameters, and tools for (excess) hazard and (net) survival predictions with associated confidence intervals.

In survival and net survival analysis, in addition to modelling the effect of time (via the baseline hazard), one has often to deal with several continuous covariates and model their functional forms, their time-dependent effects, and their interactions. Model specification becomes therefore a complex problem and penalized regression splines (Ruppert, Wand, & Carroll, 2003; Wood, 2017) represent an appealing solution to that problem as splines offer the required flexibility while penalization limits overfitting issues.

Current implementations of penalized survival models can be slow or unstable and sometimes lack some key features like taking into account expected mortality to provide net survival and excess hazard estimates. In contrast, `survPen` provides an automated, fast, and stable implementation (thanks to explicit calculation of the derivatives of the likelihood) and offers a unified framework for multidimensional penalized hazard and excess hazard models.

Summary

`survPen` is an implementation of multidimensional penalized hazard and excess hazard models for time-to-event data in **R** (R Core Team, 2018). It implements the method detailed in Fauvernier et al. (2019) which is itself included in the framework for general smooth models proposed by Wood, Pya, & Säfken (2016). Other **R** packages propose to fit flexible survival models via penalized regression splines (`rstpm2`, `bamlss`, `R2BayesX`, etc). However, the way they estimate the smoothing parameters is not optimal as they rely on either

derivative-free optimization (`rstpm2`) or MCMC (`bamlss`, `R2BayesX`), leading to possibly unstable or time-consuming analyses. The main objective of the `survPen` package is to offer a fully automatic, fast, stable and convergent procedure in order to model simultaneously non-proportional, non-linear effects of covariates and interactions between them. A second objective is to extend the approach to excess hazard modelling (J. Estève, Benhamou, Croasdale, & Raymond, 1990; L. Remontet, Bossard, Belot, Estève, & French Network of Cancer Registries, 2007). `survPen` is a free and open-source **R** package, available via GitHub at https://github.com/fauvernierma/survPen or via the CRAN repository at https://CRAN.R-project.org/package=survPen. The major features of `survPen` are documented in a walkthrough vignette that is included with the package (https://htmlpreview.github.io/?https://github.com/fauvernierma/survPen/blob/master/inst/doc/survival_analysis_with_survPen.html)

Those features include:

- Univariate penalized splines for the baseline hazard as well as any other continuous covariate.
- Penalized tensor product splines for time-dependent effects and interactions between several continuous covariates.
- Interactions between penalized splines and unpenalized continuous or categorical variables.
- Automatic smoothing parameter estimation by either optimizing the Laplace approximate marginal likelihood (LAML; Wood et al., 2016) or likelihood cross-validation criterion (LCV; O'Sullivan, 1988).
- Excess hazard modelling by specifying expected mortality rates.

`survPen` may be of interest to those who 1) analyse any kind of time-to-event data: mortality, disease relapse, machinery breakdown, unemployment, etc 2) wish to describe the associated hazard and to understand which predictors impact its dynamics.

Using the `survPen` package for time-to-event data analyses will help choose the appropriate degree of complexity in survival and net survival contexts while simplifying the model building process.

Multidimensional splines with `survPen` are currently being used in three major ongoing projects:

- Modelling the effects of time since diagnosis, age at diagnosis and year of diagnosis on the mortality due to cancer using French cancer registries data (FRANCIM network, around 1,200,000 tumours diagnosed between 1989 and 2015). This study will provide the new national estimates of cancer survival in France and its results will be used in the evaluation of the French "Plan Cancer" at the end of 2019.
- Modelling the effect of the European Deprivation Index (EDI) on the mortality due to cancer in France, using data from the FRANCIM network; this is the first time that EDI is available in all FRANCIM registries (around 210,000 tumours, diagnosed between 2006 and 2009 in 18 registries).
- For the first time modelling the effects of time since onset, age at onset, current age, year of onset and sex on the mortality due to multiple sclerosis in the biggest cohort of multiple sclerosis patients in France (37,524 patients diagnosed over the period 1960-2014 in 18 OFSEP centres).

Acknowledgements

This research was conducted as part of the first author's PhD thesis supported by the French Ministère de l Enseignement supérieur, de la Recherche et de l Innovation. The authors thank

the ANR (Agence Nationale de la Recherche) for supporting this study of the CENSUR group (ANR grant number ANR-12-BSV1-0028). This research was also carried out within the context of a four-institute cancer surveillance program partnership involving the Institut National du Cancer (INCa), Santé Publique France (SPF), the French network of cancer registries (FRANCIM), and Hospices Civils de Lyon (HCL) through a grant from INCa (attributive decision N° 2016-131). The authors are grateful to Jacques Estève for his valuable advice.

References

Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Niksic, M., Bonaventure, A., et al. (2018). Global surveillance of trends in cancer survival 2000-14 (concord-3): Analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet*, *391*(10125), 1023–1075. doi:10.1016/S0140-6736(17)33326-3

Estève, J., Benhamou, E., Croasdale, M., & Raymond, L. (1990). Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine*, *9*(5), 529–38. doi:10.1002/sim.4780090506

Fauvernier, M., Roche, L., Uhry, Z., Tron, L., Bossard, N., Remontet, L., & the Challenges in the Estimation of Net Survival Working Survival Group. (2019). Multidimensional penalized hazard model with continuous covariates: Applications for studying trends and social inequalities in cancer survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. doi:10.1111/rssc.12368

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, *9*(2), 363–379. doi:10.1137/0909024

Perme, M. P., Stare, J., & Estève, J. (2012). On estimation in relative survival. *Biometrics*, *68*(1), 113–120. doi:10.1111/j.1541-0420.2011.01640.x

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Remontet, L., Bossard, N., Belot, A., Estève, J., & French Network of Cancer Registries. (2007). An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine*, *26*(10), 2214–2228. doi:10.1002/sim.2656

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press. doi:10.1017/CBO9780511755453

Uhry, Z., Bossard, N., Remontet, L., Iwaz, J., Roche, L., Grell Eurocare-5 Working Group, & Censur Working Survival Group. (2017). New insights into survival trend analyses in cancer population-based studies: The SUDCAN methodology. *European Journal of Cancer Prevention*, *26 Trends in cancer net survival in six European Latin Countries: the SUDCAN study*, S9–S15. doi:10.1097/CEJ.0000000000000301

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Second Edition, Chapman; Hall/CRC.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. doi:10.1080/01621459.2016.1180986

Bibliographie

Abrahamowicz, M. and MacKenzie, T. A. (2007). Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine*, 26(2) :392–408. doi : https://doi.org/10.1002/sim.2519.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle,[w :] proceedings of the 2nd international symposium on information, bn petrow, f. *Czaki, Akademiai Kiado, Budapest*. doi : https://doi.org/10.1007/978-1-4612-1694-0_15.

Anderson, D. and Burnham, K. (2004). *Model selection and multi-model inference*. Second. NY : Springer-Verlag. doi : https://doi.org/10.1007/b97636.

Banner, K. M. and Higgs, M. D. (2017). Considerations for assessing model averaging of regression coefficients. *Ecological Applications*, 27(1) :78–93. doi : https://doi.org/10.1002/eap.1419.

Becher, H., Kauermann, G., Khomski, P., and Kouyate, B. (2009). Using penalized splines to model age- and season-of-birth-dependent effects of childhood mortality risk factors in rural burkina faso. *Biom J*, 51(1) :110–22. doi : https://doi.org/10.1002/bimj.200810496.

Bender, A., Groll, A., and Scheipl, F. (2018a). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, 18(3-4) :299–321. doi : https://doi.org/10.1177/1471082X17748083.

Bender, A., Scheipl, F., Hartl, W., Day, A. G., and Küchenhoff, H. (2018b). Penalized estimation of complex, non-linear exposure-lag-response associations. *Biostatistics*, 20(2) :315–331. doi : https://doi.org/10.1093/biostatistics/kxy003.

Brezger, A., Kneib, T., and Lang, S. (2005). Bayesx : Analyzing bayesian structural additive regression models. *Journal of Statistical Software, Articles*, 14(11) :1–22. doi : https://doi.org/10.18637/jss.v014.i11.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference : understanding aic and bic in model selection. *Sociological methods & research*, 33(2) :261–304. doi : https://doi.org/10.1177/0049124104268644.

Casella, G. and Berger, R. L. (2008). *Statistical inference (updated version)*, volume 2. Duxbury Pacific Grove, CA. url : https://www.isbns.net/isbn/9780495391876.

Charvat, H., Remontet, L., Bossard, N., Roche, L., Dejardin, O., Rachet, B., Launoy, G., Belot, A., and Group, C. W. S. (2016). A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Stat Med*, 35(18) :3066–84. doi : https://doi.org/10.1002/sim.6881.

Commenges, D., Bureau, J., and Putter, H. (2014). Inference with penalized likelihood. *arXiv preprint*. url : https://arxiv.org/abs/1401.7893.

Commenges, D., Joly, P., Gégout-Petit, A., and Liquet, B. (2007). Choice between semi-parametric estimators of markov and non-markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics*, 34(1) :33–52. doi : https://doi.org/10.1111/j.1467-9469.2006.00536.x.

Cornillon, P.-A. and Matzner-Løber, E. (2011). *Régression spline et régression à noyau*. Springer Paris, Paris. doi : https://doi.org/10.1007/978-2-8178-0184-1_10.

Cowppli-Bony, A., Uhry, Z., Remontet, L., Voirin, N., Guizard, A. V., Tretarre, B., Bouvier, A. M., Colonna, M., Bossard, N., Woronoff, A. S., Grosclaude, P., and French Network of Cancer, R. (2017). Survival of solid cancer patients in france, 1989-2013 : a population-based study. *Eur J Cancer Prev*, 26(6) :461–468. doi : https://doi.org/10.1097/CEJ.0000000000000372.

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *The Annals of Statistics*, pages 1676–1695. doi : https://doi.org/10.1214/aos/1176347872.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34(2) :187–202. doi : https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

Danieli, C., Bossard, N., Roche, L., Belot, A., Uhry, Z., Charvat, H., and Remontet, L. (2017). Performance of two formal tests based on martingales residuals to check the proportional hazard assumption and the functional form of the prognostic factors in flexible parametric excess hazard models. *Biostatistics*, 18(3) :505–520. url : https://www.ncbi.nlm.nih.gov/pubmed/28334368.

Danieli, C., Remontet, L., Bossard, N., Roche, L., and Belot, A. (2012). Estimating net survival : the importance of allowing for informative censoring. *Stat Med*, 31(8) :775–86. doi : https://doi.org/10.1002/sim.4464.

De Boor, C. (1972). On calculating with b-splines. *Journal of Approximation theory*, 6(1) :50–62. doi : https://doi.org/10.1016/0021-9045(72)90080-9.

De Boor, C. (1986). B(asic)-spline basics. Technical report, UW–Madison Mathematics Research Center. url : ftp://ftp.cs.wisc.edu/Approx/bsplbasic.pdf.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., and De Boor, C. (1978). *A practical guide to splines*, volume 27. springer-verlag New York. url : https://www.springer.com/gp/book/9780387953663.

Du, P. and Gu, C. (2006). Penalized likelihood hazard estimation : efficient approximation and bayesian confidence intervals. *Statistics & probability letters*, 76(3) :244–254. doi : https://doi.org/10.1016/j.spl.2005.08.008.

Estève, J., Benhamou, E., Croasdale, M., and Raymond, L. (1990). Relative survival and the estimation of net survival : elements for further discussion. *Stat Med*, 9(5) :529–38. doi : https://doi.org/10.1002/sim.4780090506.

Fahrmeir, L., Kneib, T., and Konrath, S. (2010). Bayesian regularisation in structured additive regression : a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20(2) :203–219. doi : https://doi.org/10.1007/s11222-009-9158-3.

Fauvernier, M., Remontet, L., Uhry, Z., Bossard, N., and Roche, L. (2019a). survPen : an R package for hazard and excess hazard modelling with multidimensional penalized splines. *Journal of Open Source Software*. doi : https://doi.org/10.21105/joss.01434.

Fauvernier, M., Roche, L., Uhry, Z., Tron, L., Bossard, N., and Remontet, L. (2019b). Multidimensional penalized hazard model with continuous covariates : applications for studying trends and social inequalities in cancer survival. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*. doi : https://doi.org/10.1111/rssc.12368.

Fleming, T. R. and Harrington, D. P. (1984). Nonparametric estimation of the survival distribution in censored data. *Communications in Statistics-Theory and Methods*, 13(20) :2469–2486. doi : https://doi.org/10.1080/03610928408828837.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10(1) :101–113. doi : https://doi.org/10.1214/aos/1176345693.

Gasparrini, A., Scheipl, F., Armstrong, B., and Kenward, M. G. (2017). A penalized framework for distributed lag non-linear models. *Biometrics*, 73(3) :938–948. doi : https://doi.org/10.1111/biom.12645.

Giorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Esteve, J., Gouvernet, J., and Faivre, J. (2003). A relative survival regression model using b-spline functions to model non-proportional hazards. *Statistics in medicine*, 22(17) :2767–2784. doi : https://doi.org/10.1002/sim.1484.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2) :255–277. doi : https://doi.org/10.1093/biomet/58.2.255.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420) :942–951. doi : https://doi.org/10.2307/2290630.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models : a roughness penalty approach*. CRC Press. url : https://www.crcpress.com/Nonparametric-Regression-and-Generalized-Linear-Models-A-roughness-penalty/Green-Silverman/p/book/9780412300400.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, 97(4) :773–789. doi : https://doi.org/10.1093/biomet/asq042.

Gu, C. (2014). Smoothing spline anova models : R package gss. *Journal of Statistical Software, Articles*, 58(5) :1–25. doi : https://doi.org/10.18637/jss.v058.i05.

Gu, C. and Kim, Y.-J. (2002). Penalized likelihood regression : general formulation and efficient approximation. *Canadian Journal of Statistics*, 30(4) :619–628. doi : https://doi.org/10.2307/3316100.

Gu, C. and Qiu, C. (1994). Penalized likelihood regression : a simple asymptotic analysis. *Statistica Sinica*, pages 297–304. url : https://www.jstor.org/stable/24305288.

Guillaume, E., Pornet, C., Dejardin, O., Launay, L., Lillini, R., Vercelli, M., Mari-Dell'Olmo, M., Fernandez Fontelo, A., Borrell, C., Ribeiro, A. I., Pina, M. F., Mayer, A., Delpierre, C., Rachet, B., and Launoy, G. (2016). Development of a cross-cultural deprivation index in five european countries. *J Epidemiol Community Health*, 70(5) :493–9. doi : https://doi.org/10.1136/jech-2015-205729.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. url : https://books.google.fr/books?id=qa29r1Ze1coC.

Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, 101(475) :1065–1075. doi : https://doi.org/10.1198/016214506000000348.

Hennerfeind, A., Held, L., and Sauleau, E. A. (2008). A bayesian analysis of relative cancer survival with geoadditive models. *Statistical Modelling*, 8(2) :117–139. doi : https://doi.org/10.1177/1471082x0800800201.

Herndon, J. E. and Harrell, F. E. (1990). The restricted cubic spline hazard model. *Communications in Statistics - Theory and Methods*, 19(2) :639–663. doi : https://doi.org/10.1080/03610929008830224.

Herndon, J. E. and Harrell, F. E. (1995). The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. *Statistics in Medicine*, 14(19) :2119–2129. doi : https://doi.org/10.1002/sim.4780141906.

Höge, M., Wöhling, T., and Nowak, W. (2018). A primer for model selection : The decisive role of model complexity. *Water Resources Research*, 54(3) :1688–1715. doi : https://doi.org/10.1002/2017WR021902.

Hofner, B., Hothorn, T., and Kneib, T. (2013). Variable selection and model choice in structured survival models. *Computational Statistics*, 28(3) :1079–1101. doi : https://doi.org/10.1007/s00180-012-0337-x.

Huber, P. J. (2011). *Robust statistics*. Springer. doi : https://doi.org/10.1007/978-3-642-04898-2_594.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282) :457–481. doi : https://doi.org/10.2307/2281868.

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational statistics & data analysis*, 49(1) :169–186. doi : https://doi.org/10.1016/j.csda.2004.05.006.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, 34(1) :207–228. doi : https://doi.org/10.1111/j.1467-9469.2006.00524.x.

Kneib, T., Hothorn, T., and Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics*, 65(2) :626–634. doi : https://doi.org/10.1111/j.1541-0420.2008.01112.x.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1) :79–86. doi : https://doi.org/10.1214/aoms/1177729694.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society : Series B (Methodological)*, 40(2) :113–132. doi : https://doi.org/10.1111/j.2517-6161.1978.tb01655.x.

Leray, E., Vukusic, S., Debouverie, M., Clanet, M., Brochet, B., De Sèze, J., Zéphir, H., Defer, G., Lebrun-Frenay, C., Moreau, T., et al. (2015). Excess mortality in patients with multiple sclerosis starts at 20 years from clinical onset : data from a large-scale french observational study. *PLoS One*, 10(7) :e0132033. doi : https://doi.org/10.1371/journal.pone.0132033.

Li, Z. and Wood, S. N. (2019). Faster model matrix crossproducts for large generalized linear models with discretized covariates. *Statistics and Computing*. doi : https://doi.org/10.1007/s11222-019-09864-2.

Lin, D. and Spiekerman, C. (1996). Model checking techniques for parametric regression with censored data. *Scandinavian journal of statistics*, pages 157–177. url : www.jstor.org/stable/4616394.

Liquet, B. and Commenges, D. (2004). Estimating the expectation of the log-likelihood with censored data for estimator selection. *Lifetime Data Analysis*, 10(4) :351–367. doi : https://doi.org/10.1007/s10985-004-4772-z.

Liu, X.-R., Pawitan, Y., and Clements, M. (2018). Parametric and penalized generalized survival models. *Statistical methods in medical research*, 27(5) :1531–1546. doi : https://doi.org/10.1177/0962280216664760.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7) :2372–2387. doi : https://doi.org/10.1016/j.csda.2011.02.004.

Marra, G. and Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1) :53–74. doi : https://doi.org/10.1111/j.1467-9469.2011.00760.x.

Martino, S., Akerkar, R., and Rue, H. (2011). Approximate bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38(3) :514–528. doi : https://doi.org/https://doi.org/10.1111/j.1467-9469.2010.00715.x.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2 edition. doi : http://dx.doi.org/10.1007/978-0-387-40065-5.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, 83(404) :1134–1143. doi : https://doi.org/10.2307/2290146.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on scientific and statistical computing*, 9(2) :363–379. doi : https://doi.org/10.1137/0909024.

Parker, R. and Rice, J. (1985). Discussion of "some aspects of the spline smoothing approach to nonparametric curve fitting" by BW Silverman. *Journal of the Royal Statistical Society, Series B*, 47 :40–42. url : https://www.ece.uvic.ca/~bctill/papers/mocap/Silverman_1985.pdf.

Perme, M. P., Stare, J., and Estève, J. (2012). On estimation in relative survival. *Biometrics*, 68(1) :113–120. doi : https://doi.org/10.1111/j.1541-0420.2011.01640.x.

Petersen, K. B. and Pedersen, M. S. (2008). The matrix cookbook. Journal Article 15, Technical University of Denmark. url : http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf.

R Core Team (2018). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. url : http://www.R-project.org/.

Ramsay, J. O. et al. (1988). Monotone regression splines in action. *Statistical science*, 3(4) :425–441. doi : https://doi.org/10.1214/ss/1177012761.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik*, 10(3) :177–183. doi : https://doi.org/10.1007/bf02162161.

Reiss, P. T., Huang, L., Chen, H., and Colcombe, S. (2014). Varying-smoother models for functional responses. *arXiv preprint*. url : https://arxiv.org/abs/1412.0778.

Remontet, L., Bossard, N., Belot, A., Estève, J., and the French Network of Cancer Registries (2007). An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in medicine*, 26(10) :2214–2228. doi : https://doi.org/10.1002/sim.2656.

Remontet, L., Uhry, Z., Bossard, N., Iwaz, J., Belot, A., Danieli, C., Charvat, H., Roche, L., and the CENSUR Working Survival Group (2019). Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables : Performance of this multidimensional penalized spline approach in net survival trend analysis. *Statistical Methods in Medical Research*, 28(8) :2368–2384. doi : https://doi.org/10.1177/0962280218779408.

Rodríguez-Girondo, M., Kneib, T., Cadarso-Suárez, C., and Abu-Assi, E. (2013). Model building in nonproportional hazard regression. *Statistics in medicine*, 32(30) :5301–5314. doi : https://doi.org/10.1002/sim.5961.

Rondeau, V., Commenges, D., and Joly, P. (2003). Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime data analysis*, 9(2) :139–153. url : https://www.hal.inserm.fr/inserm-00138554/document.

Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., and Soubeyran, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation : application on cancer events. *Biostatistics*, 8(4) :708–721. doi : https://doi.org/10.1093/biostatistics/kxl043.

Royston, P. and Sauerbrei, W. (2008). *Multivariable model-building : a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*, volume 777. John Wiley and Sons. doi : https://doi.org/10.1002/9780470770771.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press. doi : https://doi.org/10.1017/CBO9780511755453.

Sauerbrei, W., Royston, P., and Look, M. (2007). A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal*, 49(3) :453–473. doi : https://doi.org/10.1002/bimj.200610328.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1) :239–241. doi : https://doi.org/10.2307/2335876.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3) :379–423. doi : https://doi.org/10.1002/j.1538-7305.1948.tb00917.x.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810. doi : https://doi.org/10.1214/aos/1176345872.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–52. doi : https://doi.org/10.1111/j.2517-6161.1985.tb01327.x.

Stare, J., Pohar, M., and Henderson, R. (2005). Goodness of fit of relative survival models. *Statistics in Medicine*, 24(24) :3911–3925. doi : https://doi.org/10.1002/sim.2414.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 44–47. doi : https://doi.org/10.1111/j.2517-6161.1977.tb01603.x.

Tron, L., Belot, A., Fauvernier, M., Remontet, L., Bossard, N., Launay, L., Bryere, J., Monnereau, A., Dejardin, O., Launoy, G., et al. (2019). Socioeconomic environment and disparities in cancer survival for 19 solid tumor sites : An analysis of the french network of cancer registries (francim) data. *International journal of cancer*, 144(6) :1262–1274. doi : https://doi.org/10.1002/ijc.31951.

Ueberhuber, C. W. (1997). *Numerical computation 1 : methods, software, and analysis*. Springer Science & Business Media. doi : https://doi.org/10.1007/978-3-642-59118-1.

Uhry, Z., Bossard, N., Remontet, L., Iwaz, J., Roche, L., the Grell Eurocare-5 Working Group, and the Censur Working Survival Group (2017). New insights into survival trend analyses in cancer population-based studies : the sudcan methodology. *Eur J Cancer Prev*, 26 Trends in cancer net survival in six European Latin Countries : the SUDCAN study :S9–S15. doi : https://doi.org/10.1097/CEJ.0000000000000301.

Umlauf, N., Adler, D., Kneib, T., Lang, S., Zeileis, A., et al. (2015). Structured additive regression models : An r interface to bayesx. *Journal of Statistical Software*, 63(i21). doi : https://doi.org/10.18637/jss.v063.i21.

Umlauf, N., Klein, N., and Zeileis, A. (2018). Bamlss : Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27(3) :612–627. doi : https://doi.org/10.1080/10618600.2017.1407325.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. doi : https://doi.org/10.1017/CBO9780511802256.

Vaupel, J. W. and Yashin, A. I. (1985). Heterogeneity's ruses : some surprising effects of selection on population dynamics. *The American Statistician*, 39(3) :176–185. doi : https://doi.org/10.2307/2683925.

Verweij, P. J. and Van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in medicine*, 12(24) :2305–2314. doi : https://doi.org/10.1002/sim.4780122407.

Wahba, G. (1983). Bayesian" confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–150. url : https://www.jstor.org/stable/2345632.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 62(2) :413–428. doi : https://doi.org/10.1111/1467-9868.00240.

Wood, S. N. (2006a). *Generalized additive models : an introduction with R*. Chapman and Hall/CRC.

Wood, S. N. (2006b). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4) :1025–1036. doi : https://doi.org/10.1111/j.1541-0420.2006.00574.x.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73(1) :3–36. doi : https://doi.org/10.1111/j.1467-9868.2010.00749.x.

Wood, S. N. (2017). *Generalized additive models : an introduction with R*. Second Edition, Chapman and Hall/CRC. doi : https://doi.org/10.1201/9781315370279.

Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 64(1) :139–155. doi : https://doi.org/10.1111/rssc.12068.

Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. (2017). Generalized additive models for gigadata : Modeling the u.k. black smoke network daily data. *Journal of the American Statistical Association*, 112(519) :1199–1210. doi : https://doi.org/10.1080/01621459.2016.1195744.

Wood, S. N., Pya, N., and Säfken, B. (2016a). Rejoinder. *Journal of the American Statistical Association*, 111(516) :1573–1575. doi : https://doi.org/10.1080/01621459.2016.1250583.

Wood, S. N., Pya, N., and Säfken, B. (2016b). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516) :1548–1563. doi : https://doi.org/10.1080/01621459.2016.1180986.

Wynant, W. and Abrahamowicz, M. (2014). Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Statistics in medicine*, 33(19) :3318–3337. doi : https://doi.org/10.1002/sim.6178.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441) :120–131. doi : https://doi.org/10.1080/01621459.1998.10474094.