Modélisation et analyse prédictive des risques et des conséquences post accident vasculaire cérébral

Youssef Hbid

THÈSE DE DOCTORAT

Spécialité : Mathématiques Appliquées

Écoles doctorales :

Sciences et Techniques, UCA

Et

Sciences Mathématiques de Paris-Centre (ED 386), Sorbonne Université

présentée par

YOUSSEF HBID

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ CADI AYYAD DE MARRAKECH

Et

DE SORBONNE UNIVERSITÉ, PARIS

Sujet de la thèse :

Modélisation et analyse prédictive des risques et des conséquences post accident vasculaire cérébral

Soutenue le 17/09/2021

Après avis des rapporteurs :

Stéphanie PORTET	Professeur, University of Manitoba, Winnipeg, Canada
Jean-Christophe POGGIALE	Professeur, Aix-Marseille Université, France
Mostafa ADIMY	Directeur de Recherches à l'INRIA de Lyon, France

Devant la commission d'examen :

Président :

Mohamed ELALAOUI TALIBI	PES, Faculté des Sciences Semlalia, Université Cadi Ayyad

Examinateurs :

Mostafa ADIMY	Directeur de Recherches à l'INRIA de Lyon
Bernard CAZELLES	Professeur, Sorbonne Université, Ecole Normale Supérieure, Paris
Pierre-Alexandre BLIMAN	Directeur de Recherches à l'INRIA, Sorbonne Université, Paris
Abdel DOUIRI	Professeur, King's College of London
Mohamed KHALADI	PES, Faculté des Sciences Semlalia, Université Cadi Ayyad
Jean- Christophe POGGIALE	Professeur, Aix-Marseille Université

Remerciements

La rédaction de ce manuscrit constitua l'une des étapes les plus importantes de mes quatre années de thèse. Et cette page fut sans aucun doute celle qui aura nécessité le plus d'attention de ma part, tant je tenais à remercier comme il se doit les personnes ayant contribué à la réalisation et l'achèvement de cette thèse.

Mes premiers remerciements vont à mes encadrants, Professeur Bliman, Professeur Douiri et Professeur Khaladi, qui ont dirigé et encadré mes travaux. Merci pour votre confiance, vos précieux conseils et votre disponibilité. Merci pour votre enthousiasme et votre passion, que vous avez su me communiquer. Je n'aurais pu imaginer meilleurs guides pour accompagner mes premiers pas dans le monde de la recherche scientifique.

Mes plus sincères remerciements vont également au Professeur Charles D.A Wolfe, directeur de la division de la santé publique au King's College de Londres, pour son appui et soutien, ses précieux conseils, et sa collaboration tout au long de cette thèse.

Je tiens également à remercier les professeurs Stéphanie Portet, Mostafa Adimy et Jean-Christophe Poggiale d'avoir rapporté ma thèse. Merci pour vos conseils et remarques très constructives qui ont amélioré ce manuscrit.

Je remercie très chaleureusement le Professeur Bernard Cazelles d'avoir accepté de faire partie des membres du jury de ma thèse, de même que le Professeur Mohamed El Alaoui Talibi qui m'a fait l'honneur de présider ce jury.

Mes plus sincères remerciements vont également à l'UMI-UMMISCO et la Délégation régionale Île-de-France de l'IRD qui ont financé ma thèse. En particulier, je remercie le Pr.Jean-Daniel Zucker, le Pr.Christophe Cambier, Madame Kathy Baumont, Madame Elisabeth Pereira et Madame Rolande Altemaire pour leur amabilité et assistance administrative.

Je tiens à remercier tous les membres du Laboratoire de Mathématiques et Dynamique des Populations pour leur soutien et encouragements. Je tiens également à remercier l'ensemble des membres du Laboratoire Jacques-Louis Lions pour leur convivialité et l'accueil qu'ils me réservaient lors de mes séjours à Jussieu.

Je remercie particulièrement mes amis qui furent présents à mes côtés lorsque j'en avais besoin. Merci donc à Hamza Benfdil, Anas Hachad, Ahmed Karafi, Youssef Bourfia, Mohamed Belakziz, Oussama Bouali, Zakaria Bouchlih, Zouhair el kaoukabi, Nabil Lamhaddar, Julio Cardenas.

Je remercie infiniment les familles HBID, LAKANAL et FORNERIS pour leur soutien constant et inconditionnel.

Je ne pourrais pas finir sans exprimer mon amour et mes sincères remerciements à ceux qui m'ont soutenu et encouragé, sans relâche, durant toute ma vie : mes très chers parents et mon petit frère Haitam. Rien n'aurait été possible sans vous !

Résumé

La modélisation prédictive enveloppe une variété de techniques issues des statistiques qui analyse des faits actuels et historiques afin de prédire des événements futurs ou inconnus. Ses applications concernent toutes les activités humaines et permettent par exemple l'amélioration de la qualité des soins et l'optimisation du diagnostic et des traitements.
L'objectif principal de cette thèse est d'identifier et proposer les meilleures stratégies pour établir des modèles capables de prédire avec précision les événements ou tendances futures, en particulier les risques et les conséquences des accidents vasculaires cérébraux (AVC).
Pour ce faire, nous avons étudié deux approches : une approche statistique basée sur le modèle mixte et une approche stochastique basée sur le modèle espace-état.
Concernant l'approche statistique, nous avons proposé en premier lieu une nouvelle méthode de régularisation pour le modèle de régression statistique en associant le cadre théorique des problèmes inverses avec celui des statistiques robustes. Nous avons évalué et validé les performances de la méthodologie proposée par rapport à d'autres méthodes (lasso, lasso adaptatif, ridge, elastic-net) à la fois par des simulations et des données réelles basées sur le registre d'AVC du Sud de Londres (SLSR). Par ailleurs, cette méthodologie peut être également étendue au modèle mixte. En second lieu, nous avons contribué à l'amélioration de la méthodologie des "courbes de récupération". En effet, des travaux antérieurs ont montré qu'il est possible de construire des "courbes de récupération" permettant de prédire l'évolution de l'incapacité physique d'un individu après un AVC à court ou moyen terme (52 semaines). Dans ce travail de thèse, nous avons étendu la méthodologie des courbes de récupération en utilisant des modèles mixtes régularisés qui ont notablement amélioré la prédiction. Nous avons également évalué la performance de la méthodologie des "courbes de récupération régularisée" en utilisant à la fois des métriques traditionnelles (discrimination et étalonnage) et non-traditionnelles (utilité clinique). D'un point de vue applicatif, nous avons utilisé l'approche des « courbes de récupération régularisées » pour prédire à long-terme (5 ans) le déclin cognitif chez un individu après un AVC. Cette méthodologie a ainsi été développée et validée, pour la première fois, pour des prévisions à long terme et d'autres conséquences d'AVC, en l'occurrence le déficit cognitif.
Concernant l'approche stochastique, les avantages tirés des modèles mixtes régularisés nous ont conduit à comparer cette classe de modèle avec d'autres approches de modélisation robustes, mais plus complexes, à savoir les modèles espace-état. À travers cette comparaison, nous avons proposé un cadre de modélisation optimal qui suggère l'utilisation de modèles mixtes dans des situations ne présentant pas de structure complexe, mais souvent abordées par des modèles espace-état. Motivés par des considérations théoriques et numériques, nous avons proposé une stratégie générale pour reformuler un modèle espace-état en un modèle mixte. Nous avons illustré et validé le cadre proposé à la fois par des études de simulation et des données réelles basées sur le registre d'AVC du Sud de Londres (SLSR).
Toutes les méthodologies proposées dans cette thèse, peuvent être appliquées non seulement à l'AVC, mais aussi à toutes les maladies chroniques où la prédiction du rétablissement des patients s'avère nécessaire.

Abstract

Predictive modeling encompasses a variety of statistical techniques, which analyze current and historical facts to forecast the risk of future events. They are important in various applications, including meteorology, finance, and medicine. In recent years, predictive models and tools become more pertinent in clinical fields, in particular, the application of personalized approaches to improve quality of care and to optimize testing, diagnosis and treatment.

The aim objective of this thesis is to identify and develop the best strategies to build models that predict accurately future events or trends, with application to stroke outcomes. To do this, we studied two approaches: a statistical approach based on mixed model and a stochastic approach based on space-state model.

Regarding the statistical approach, we first proposed a novel regularized regression modellings under the inverse problem framework, where the idea of regularization is based on Huber robust statistics. We showed improved performances of the proposed method over existing approaches, including (lasso, adaptive-lasso, ridge and elastic-net) in both simulation studies and real data using patients data from the South London Stroke Registry (SLSR). Previous work has shown that it is possible to build recovery curves over time which can predict the evolution of an individual's physical disability after a stroke in the short to mid- term up to 52 weeks. In this thesis work, we revisited these modeling and proposed a novel mathematical inverse problem approach to improve the methodology of predictive recovery curves . We extended recovery curves using regularized mixed models which notably improved the prediction. We compared the performance of these models using both metrics, discrimination and calibration, as well as clinical utility. From clinical perspective, we used regularized recovery curves methodology to develop and validate a patient-specific predictive model to estimate risk for cognitive decline up to 5 years (long-term) after ischemic stroke. This methodology was developed and validated, for the first time, for long-term predictions with different outcomes of stroke.

Regarding the stochastic approach, the benefits derived from penalized mixed models led us to comparisons with other robust but more complex modeling approaches, namely space-state models. Through this investigations, we offer an optimal modeling framework that suggest the use of mixed models in situations that not exhibit complex structure, but often approached by space-state models. Motivated by theoretical and numerical considerations, we proposed a general strategy to reformulate a space-state model as a mixed model. We illustrate and validate the proposed framework in both, simulation studies and real data based on patients data from the South London Stroke Registry (SLSR).

All the methodologies proposed in this thesis can be applied not only to stroke, but also to any ongoing, long term or recurring conditions that can have a significant impact on people's lives, as well as, planning innovative healthcare strategies.

Table des matières

Table des figures

Liste des abréviations :

AVC	Accidents Vasculaires Cérébraux.
PSCI	Déficience cognitive post-AVC.
SLSR	South London Stroke Register.
REML	Restricted Maximum Likelihood (REML).
ML	Maximum de vraisemblance.
BLUP	Best Linear Unbiased Prediction.
LM	Modèle linéaire.
GLM	Modèle linéaire généralisé.
MA	Modèle additif.
GAM	Modèle additif généralisé.
LMM	Modèle linéaire à effets mixtes (LMM).
GLMM	Modèle linéaire à effets mixtes généralisé.
GAMM	modèle additif mixte généralisé.
E-M	Espérance-maximisation (E-M).
NR	Newton-Raphson.
FS	Score-Fisher
SSR	Somme des carrés des résidus
Var ()	Variance.
E ()	Espérance.
COV ()	Covariance.
i.i.d	variables indépendantes et identiquement distribuées.
tr	Trace.
MSE	Erreur quadratique moyenne.
OLS	Moindres carrés ordinaire.
MAP	Maximum a posteriori.
AUC	Aire sous la courbe.
PPV	Valeur prédictive positive.
NPV	Valeur prédictive négative.
LR+	Likelihood ratio positif.
LR-	Likelihood ratio négatif.
DCA	Analyse de la courbe de décision.

Introduction générale

Objectif principal du sujet de thèse

L'objectif principal de cette thèse est d'identifier et proposer les meilleures stratégies pour établir des modèles capables de prédire avec précision les événements ou tendances futures, en particulier les risques et les conséquences des accidents vasculaires cérébraux.

Motivations et contextes

L'accident vasculaire cérébral (AVC) est la cause la plus fréquente d'incapacité physique chez les adultes et la troisième cause de décès à l'échelle mondiale. Malgré l'introduction de traitements efficaces pour l'AVC, la réadaptation précoce et la prévention secondaire, la plupart des survivants d'un AVC présentent des comorbidités médicales, des troubles physiques et / ou cognitifs qui nécessitent une évaluation et une gestion actives continues. Il a été estimé que le nombre annuel d'AVC incident dans le monde atteindrait 23 millions à l'horizon de 2030 [1], du fait du vieillissement de la population. La gestion et la prise en charge des patients survivant à un AVC représente alors, l'un des plus grands défis en santé publique. Les systèmes de santé peinent à trouver de meilleures solutions pour l'amélioration de la qualité, l'efficacité et la réduction des coûts des soins. L'une des solutions qui se présente est la construction et l'utilisation de modèles d'analyse prédictive.

Les questions de recherche et les objectifs de cette thèse

Pour garantir la construction d'un modèle de prédiction précis et interprétable, les méthodes de régularisation sont souvent des méthodes de prédilection. La régularisation est essentielle pour réussir la modélisation statistique des données modernes, en l'occurrence dans le domaine de la santé. En général, ces données sont de grande dimension, parfois bruitées et contiennent souvent des covariables non pertinentes. Le premier défi que nous relevons dans cette thèse est de mieux comprendre ces différents aspects, expliquer l'importance de la régularisation dans les modèles d'analyse prédictive et proposer une nouvelle méthode de régularisation. Pour ce faire, nous étudions le modèle de régression statistique avec une approche problème inverse. En associant le cadre théorique des problèmes inverses avec celui des statistiques robustes, nous caractérisons et proposons une nouvelle méthode de régularisation. Nous évaluons et validons les performances de la méthodologie proposée par rapport à d'autres méthodes (lasso [2], lasso adaptatif [3], ridge [4], elastic-net [5]) à l'aide de simulations et données réelles basées sur le registre d'AVC du Sud de Londres (SLSR)[6].
Le deuxième défi que nous relevons, est le développement et la conceptualisation d'une bonne stratégie de modélisation permettant de prédire à long-terme (5 ans), les conséquences post-AVC telles que le déficit cognitif, la dépression ou la mortalité. Tilling,

Toshke et Douiri [7, 8, 9] ont montré qu'il est possible de construire des «courbes de récupération» au fil du temps pouvant prédire l'évolution de l'incapacité physique d'un individu après un AVC à court ou moyen terme (52 semaines). Ce travail de thèse contribue au développement de la méthodologie des « courbes de récupération» [1] et ce d'un point de vue mathématique et applicatif. D'un point de vue mathématique, nous construisons des « courbes de récupération régularisées » à l'aide des modèles à effets mixtes au lieu des méthodes de régression classiques. Nous évaluons également la performance de ces modèles en utilisant à la fois des métriques traditionnelles (discrimination et étalonnage) et non traditionnelles (utilité clinique). D'un point de vue applicatif, nous utilisons l'approche des « courbes de récupération régularisées » pour prédire à long-terme (5 ans), le déclin cognitif chez un individu après un AVC. Cette méthodologie est ainsi développée et évaluée, pour la première fois, pour des prévisions à long terme et d'autres conséquences d'AVC, en l'occurrence le déficit cognitif.
Les avantages tirés des modèles à effets mixtes par rapport aux méthodes de régression traditionnelles, nous ont conduits à comparer cette classe de modèles avec des approches de modélisation plus complexes, à savoir les modèles espace-état.
Par ce travail, nous offrons un cadre de modélisation optimal permettant l'utilisation des modèles à effets mixtes dans des situations souvent approchées par des modèles espace-état, mais ne nécessitant pas un tel cadre de modélisation complexe.
C'est la première fois, à notre connaissance, que cette méthodologie est présentée d'une manière détaillée et explicite avec une application à l'AVC.
Motivés par des considérations à la fois théoriques, numériques et pratiques, nous proposons une stratégie générale pour transformer un modèle espace-état en modèle à effets mixtes. Cette méthodologie est ensuite appliquée aux séries chronologiques structurelles, généralement approchées par un cadre espace-état. Nous comparons également des procédures d'estimation, spécifiques aux modèles espace-état (filtre de Kalman) [11], à des procédures propres à un modèle à effets-mixtes (REML-BLUP)[12]. Nous illustrons et validons le cadre théorique proposé via des simulations et des données réelles basées sur le registre d'AVC du Sud de Londres (SLSR).

Plan de la thèse

Cette thèse est organisée en quatre chapitres.

Chapitre 1 : Revue bibliographique

Ce chapitre fait l'objet d'une revue bibliographique, et est scindé en trois grandes sections.

La première section, *modèle et modélisation*, donne un aperçu des modèles d'analyse prédictive, à savoir : le modèle linéaire (LM), le modèle linéaire généralisé (GLM), le modèle additif (MA), le modèle additif généralisé (GAM), le modèle linéaire à effets mixtes (LMM), le modèle linéaire à effets mixtes généralisé (GLMM), le modèle additif mixte généralisé (GAMM) et le modèle espace-état.

Dans la deuxième section, *techniques de régularisation et sélection de variables*, nous passons en revue les principales méthodes de régularisation et de sélection de variables existantes dans la littérature. Nous commençons par rappeler les définitions de ces mé-

1. Traduction française de "*recovery curve*", une méthodologie introduite par K. Tilling [10].

thodes et les conditions idoines pour leur utilisation. Ensuite, nous décrivons l'algorithme de descente par coordonnée (*coordinate descent*), connu pour son adaptabilité aux méthodes de régularisation, et nous l'appliquons à différentes méthodes de régularisation.

La troisième et dernière section de ce chapitre, *construction et conceptualisation d'un modèle de prédiction clinique avec application à l'accident vasculaire cérébral*, fait l'objet d'une réflexion et d'une synthèse globale sur les différentes méthodologies de construction de modèles de prédiction clinique.
Nous commençons par définir le concept et le processus derrière la construction de ces modèles et nous les classifions. Ensuite, nous discutons les conditions nécessaires à la conduite de cette méthodologie de recherche ainsi que ses limitations.
Comme application, nous avons choisi l'accident vasculaire cérébral (AVC), un sujet d'importance significative pour la santé publique. Nous définissons l'AVC et ses conséquences, en l'occurrence la déficience cognitive post-AVC. Finalement, nous passons en revue les travaux menés sur la modélisation prédictive du risque post-AVC, en particulier la méthodologie « des courbes de récupération ».

Chapitre 2 : Une approche problème inverse pour les modèles de régression régularisés avec application à la prédiction de la récupération fonctionnelle post-AVC.

Ce chapitre a été publié dans une revue [13] et est traduit en langue française dans ce manuscrit de thèse. Il a également fait l'objet de deux présentations orales :
La première lors de la ***"1st International Conference on Research in Applied Mathematics and Computer Science (ICRAMCS 2019), Université Hassan-2, Casablanca, Maroc (Mars 2019)."***
La deuxième lors de la ***"2nd IMA Conference On Inverse Problems From Theory To Application, University College London, Uk. (September 2019)".***

Dans ce chapitre, nous montrons que le modèle de régression peut être formulé comme un problème inverse qui mesure l'écart entre l'observation et les données produites par la représentation des prédicteurs. Cette approche pourrait effectuer simultanément la sélection de variables et l'estimation des coefficients. Nous nous sommes concentrés particulièrement sur un problème de régression linéaire, $Y \sim N(X\beta, \sigma I_n)$, où $\beta \in R^p$ est le paramètre d'intérêt et ses composantes sont les coefficients de régression.
Le problème inverse permet de trouver une estimation du paramètre β, qui est lié par l'opérateur linéaire $(L : \beta \to X\beta)$ aux données obsérvées $Y = X\beta + \epsilon$.
Ce problème pourrait être construit en trouvant une solution dans le sous-espace affine $L^{-1}(Y)$. Cependant, en présence de colinéarité, de données de grande dimension et d'un conditionnement élevé de la matrice de covariance relative aux données, la solution peut ne pas être unique. L'introduction d'une information préalable pour réduire le sous-ensemble $L^{-1}(Y)$ et régulariser le problème inverse est alors nécessaire.
Forts du solide cadre de la statistique robuste de Huber, nous proposons un régularisateur optimal pour le problème de régression. Nous évaluons et comparons la méthode de régularisation proposée par rapport à d'autres méthodes de régularisation : ridge, lasso, adaptive-lasso et l'elastic-net sous différents scénarios, à savoir, un conditionnement élevé de la matrice de covariance relative aux données et une grande amplitude d'erreur, et ce, sur des données simulées et réelles basées sur le registre des accidents vasculaires cérébraux du sud de Londres (SLSR). L'approche proposée peut être étendue au modèle linéaire à effets-mixtes. Le cadre des problèmes inverses, associé à la méthodologie statis-

tique robuste, offre ainsi de nouvelles perspectives de recherche en matière de régression statistique et d'apprentissage automatique.

Chapitre 3 : Prédiction du risque de déclin cognitif à long-terme après un AVC ischémique.

Ce chapitre est accepté pour publication dans la revue *"Journal of stroke and cerebrovascular diseases, ELSEVIER"*. Il est repris dans ce manuscrit avec plus de détails au niveau de la stratégie de modélisation mathématique.

Dans ce chapitre, nous développons et validons un modèle prédictif centré sur le patient pour estimer le risque de déclin cognitif à long terme (5 ans) après un AVC ischémique. Nous évaluons les écarts entre la récupération observée et prévue ainsi que les différences dans les tendances de récupération. D'un point de vue mathématique, ce travail développe la stratégie de modélisation basée sur les *"courbes de récupération"*. Il évalue la robustesse de cette méthodologie et sa capacité à prédire, à long terme, une des conséquences importantes d'AVC, qu'est la déficience cognitive.
Une autre contribution inédite de ce travail, est le développement des *"courbes de récupération régularisées"* par le biais des modèles à effets mixtes pour des prédictions à long-terme. Nous évaluons également la performance de ces modèles en utilisant à la fois des métriques traditionnelles (discrimination et étalonnage) et non-traditionnelles (utilité clinique). Les résultats de cette recherche, d'un point de vue clinique, seront utilisés pour fournir des informations pronostiques aux survivants d'un AVC et à leurs familles, dans le but de faciliter leur suivi à long terme et aider à l'élaboration de plans de soins et de gestion adaptés.

Chapitre 4 : Séries temporelles structurelles : de la formulation espace-état à la représentation en modèle à effets mixtes.

Ce chapitre est un travail en cours et sera soumis pour publication après quelques améliorations.

Dans ce chapitre, nous proposons une stratégie générale pour transformer un modèle espace-état en modèle à effets mixtes. Cette stratégie est ensuite appliquée à la classe des séries chronologiques structurelles, souvent traitée par une approche espace-état, et dont le principal objectif demeure l'analyse et la prédiction.
Par ce travail, nous proposons une méthodologie optimale permettant l'utilisation des modèles à effets mixtes pour des cas ne nécessitant pas un cadre de modélisation complexe, à savoir les modèles espace-état. Comme limitation de la méthodologie proposée, nous étudions et discutons le cas de modèles espace-état générant une matrice de transition avec plus de deux paramètres.
Via des simulations et des données réelles, utilisant le registre d'AVC du Sud de Londres (SLSR) [6], nous montrons que les estimations basées sur le filtre de Kalman [11], spécifique aux modèles espace-état, sont quasi-équivalentes à la meilleure prédiction linéaire sans biais (BLUP), basée sur le RMLE [12], dans un modèle à effets mixtes.
La méthodologie proposée est appliquée pour la première fois à des données d'AVC, dans le but de décrire la variation du score de Barthel moyen des patients victimes d'AVC enregistrés dans le SLSR, et ce, au cours des 23 dernières années. Elle est également appliquée pour évaluer l'impact de la mise à jour des directives et conseils cliniques sur le diagnostic et la prise en charge des patients victime d'AVC, dans les 48 heures suivant l'apparition

des symptômes.
Dans cette étude, nous nous sommes restreints au "*local level model*" pour illustrer et valider la méthodologie proposée.
Comme perspectives, nous souhaitons élargir cette étude en appliquant les différents modèles étudiés théoriquement, sur d'autres problématiques de santé publique, en l'occurence les AVC.

Chapitre 1

Revue bibliographique

1.1 Modèles et modélisation

Les modèles sont utilisés dans la plupart des disciplines scientifiques comme substituts de la réalité. Il se peut qu'il soit pratiquement impossible de mener des expériences sur un système physique, et donc faire recours à un modèle soit pour remplacer le dit système ou bien le généraliser à de nouvelles situations. Un modèle d'évolution décrit le comportement d'un système à l'aide d'un langage mathématique. Ce dernier pourrait être un ensemble d'équations différentielles comme il pourrait être une règle pour combiner des observations passées. Les modèles mathématiques présentent un intérêt et une utilisation particuliers pour l'ingénierie et la science. Étant donné qu'ils sont utilisés et revêtent une importance dans tant de domaines différents, il existe bien entendu une grande variété de types de modèles et de techniques de modélisation. Il existe également plusieurs domaines étudiant l'acte de modélisation, chacun avec sa propre nomenclature. L'art de la modélisation consiste à trouver un équilibre permettant de répondre aux questions posées ou de se poser de nouvelles questions. La complexité du modèle dépendra alors du problème et de la réponse requise, de sorte que différents modèles et analyses peuvent être appropriés pour un même ensemble de données.
Dans cette thèse, l'intérêt est porté particulièrement à la modélisation statistique avec ses différents approches et modèles.
La modélisation statistique peut être considérée comme un outil puissant pour développer et tester des théories à travers l'explication causale, la prédiction et la description.
Dans de nombreuses disciplines, il existe une utilisation quasi-exclusive de la modélisation statistique pour l'exploration causale, selon laquelle les modèles à fort pouvoir explicatif sont d'un pouvoir prédictif élevé. Dans différents domaines, tels que l'économie, la psychologie, l'éducation et les sciences de l'environnement, les modèles statistiques sont majoritairement utilisés pour l'explication causale, et les modèles possédant un pouvoir explicatif élevé jouissent d'un pouvoir prédictif inhérent. Dans des domaines tels que le traitement du langage naturel, la bio-informatique et l'épidémiologie, l'accent mis sur l'explication causale par rapport à la prédiction empirique est plus mitigé.
La caractéristique clé d'un modèle statistique est que la variabilité est représentée à l'aide de distributions de probabilités. Ces distributions forment les éléments de base à partir desquels le modèle est construit. En règle générale, le modèle doit tenir compte des variations aléatoires et systématiques. Le caractère aléatoire associé à la distribution de probabilité explique la dispersion aléatoire dans les données, tandis que le modèle systémique est supposé être généré par la structure du modèle.
Concernant le volet modèles et modélisation, nous nous intéressons majoritairement au modèle linéaire à effets mixtes et au modèle espace-état.

Ces différentes approches de modélisation ont largement servi à l'obtention des résultats clés de cette thèse. C'est pourquoi nous présentons et décrivons soigneusement les modèles, l'intuition, les hypothèses et les compromis derrière chacune des méthodes que nous considérons.
En premier lieu, nous examinons le modèle linéaire (LM), qui est le point de départ fondamental de tous les modèles d'analyse prédictive. Nous présentons également des extensions de ce modèle, à savoir, le modèle linéaire généralisé (GLM), le modèle additif (AM) et le modèle additif généralisé (GAM). La figure 1.1 donne un aperçu général des liens existants entre différents modèles d'analyse prédictive. Elle montre également que nous pouvons combiner des modèles mixtes et additifs pour obtenir une classe de modèle plus sophistiquée à savoir le modèle additif mixte généralisé (GAMM).
En second lieu, nous donnons une emphase spéciale au modèle linéaire à effets mixtes (LMM). Dans ce sens, nous définissons le (LMM) et les notions d'effets fixes et aléatoires. Nous estimons les paramètres de variance de ce modèle via la méthode du maximum de vraisemblance (ML) et celle du maximum de vraisemblance restreint (REML). Nous décrivons et comparons trois algorithmes permettant d'optimiser le calcul de ces estimations : Espérance-maximisation (E-M), Newton-Raphson (NR) et Score de Fisher (Fisher scoring). Nous généralisons le (LMM) au modèle linéaire mixte généralisé (GLMM).
Finalement, nous fournissons une description des modèles espace-état que nous jugeons nécessaire à la compréhension du chapitre 4. Des méthodes d'estimation de tels modèles sont ensuite expliquées en deux temps : l'estimation des variables cachées avec le filtre de Kalman, puis celle des paramètres avec l'algorithme E-M.

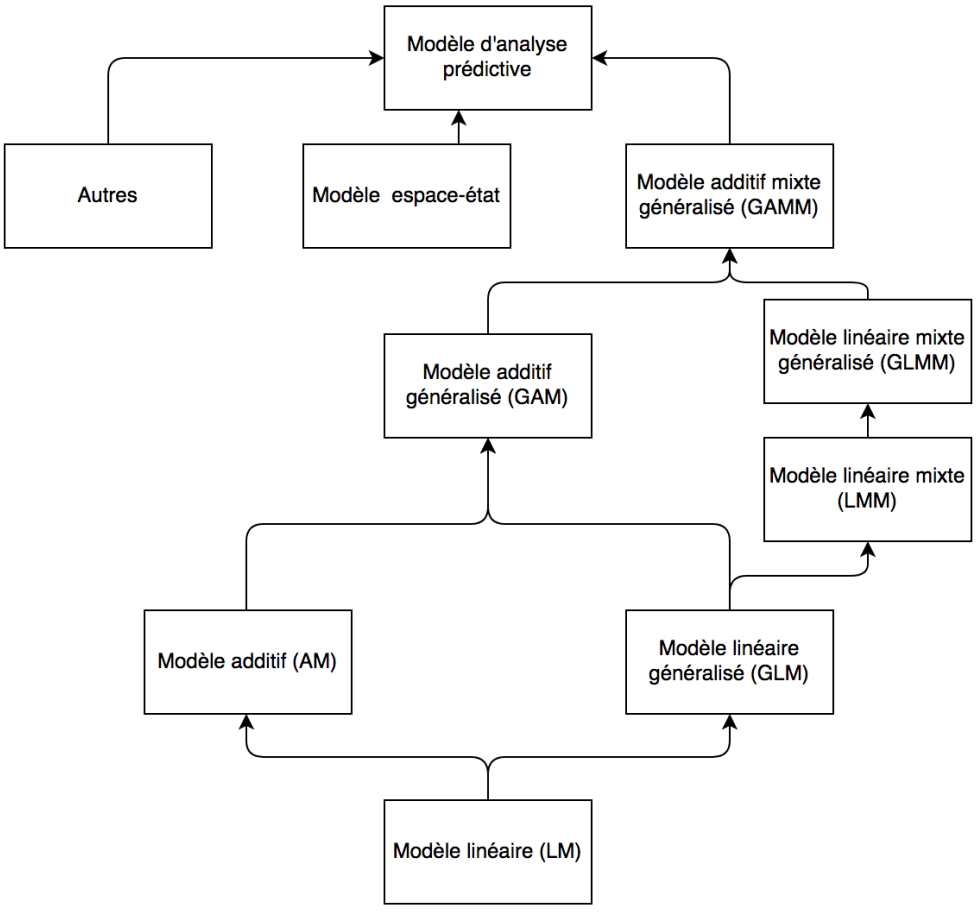

FIGURE 1.1 – Aperçu général et lien entre les modèles d'analyse prédictive.

1.1.1 Modèle linéaire

Le modèle linéaire (LM) est le modèle statistique de base que l'on utilise pour analyser une expérience où l'on étudie sur n unités expérimentales les variations d'une variable réponse y en fonction de facteurs qualitatifs ou quantitatifs, appelés aussi variables explicatives. Le modèle linéaire s'écrit :

$$\boldsymbol{Y_i = \mu_i + \epsilon_i}. \tag{1.1}$$

i est le numéro de l'unité expérimentale.
μ_i est l'espérance de Y_i et inclut l'effet de variables explicatives.
ϵ_i est une variable aléatoire résiduelle, appelée erreur, incluant la variabilité du matériel expérimental, celle due aux variables explicatives non incluses dans le modèle, et celle due aux erreurs de mesure.
Selon la nature des variables incluses dans la partie explicative μ_i du modèle, on distingue trois grandes catégories de modèle linéaire :
Lorsque les variables explicatives sont quantitatives, le modèle est appelé modèle de régression : simple s'il n'y a qu'une seule variable explicative, multiple sinon.
Lorsque les variables explicatives sont qualitatives, elles sont appelées facteurs et le modèle ainsi construit est un modèle d'analyse de la variance.
Lorsque les variables explicatives sont à la fois de nature quantitatives et qualitatives, le modèle ainsi construit est un modèle d'analyse de la covariance.

Structure aléatoire

Nous supposerons que la variable aléatoire Y_i a une distribution normale avec une moyenne μ_i et une variance σ^2, :

$$\boldsymbol{Y_i \sim N\left(\mu_i, \sigma^2\right)}$$

Nous supposons également que les observations sont mutuellement indépendantes. Cette hypothèse nous permet d'obtenir la distribution conjointe des données comme un simple produit des distributions de probabilité individuelles, qui sont à la base de la construction de la fonction de vraisemblance qui sera utilisée pour l'estimation des paramètres et les tests. Lorsque les observations sont indépendantes, elles sont également non corrélées et leur covariance est nulle, donc $\text{cov}\left(Y_i, Y_j\right) = 0$ pour $i \neq j$. Il sera pratique de collecter les n réponses dans un vecteur colonne Y, que nous considérons comme une réalisation d'un vecteur aléatoire Y de moyenne $\text{E}(Y) = \mu$ et matrice de variance-covariance $\text{var}(Y) = \sigma^2 I$, où I est la matrice identité. Les éléments diagonaux de $\text{var}(Y)$ sont égales à σ^2 et les éléments hors diagonale sont tous nuls, donc les observations n ne sont pas corrélées et ont les mêmes variances. Sous l'hypothèse de normalité, Y a une distribution normale multivarié $Y \sim N_n\left(\mu, \sigma^2 I\right)$.

Structure systèmatique

Supposons que nous ayons des données avec p prédicteurs $x_1, \ldots, x_p$ qui prennent des valeurs $x_{i1}, \ldots, x_{ip}$. Nous supposons que μ_i est une fonction linéaire des prédicteurs

$$\boldsymbol{\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta p x_{ip}}$$

Les coefficients β_j sont appelés coefficients de régression. Cette équation peut être écrite d'une manière plus compacte en utilisant la notation matricielle :

$$\boldsymbol{\mu_i} = \mathbf{x_i'}\boldsymbol{\beta}. \tag{1.2}$$

où x_i' est un vecteur ligne contenant les valeurs des prédicteurs. β est un vecteur colonne contenant les coefficients de régression. De manière encore plus compacte, nous écrivons :

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}. \tag{1.3}$$

où X est une matrice $n \times p$ contenant les valeurs des prédicteurs p pour les unités n. La matrice X est généralement appelée le modèle ou la matrice de conception. L'expression $X\beta$ est appelée le prédicteur linéaire.
Le modèle linéaire le plus simple possible suppose que $\mu_i = \mu$ pour tout i. Ce modèle est souvent appelé modèle nul, car il ne postule aucune différence systémique entre les unités. Le modèle nul peut être obtenu comme cas particulier de l'équation (1.2) en définissant $p = 1$ et $x_i = 1$ pour tout i.

Nous pouvons également définir un modèle où chaque unité a sa propre valeur μ_i. Ce modèle est appelé le modèle saturé car il a autant de paramètres dans le prédicteur linéaire (ou paramètres linéaires) que d'observations. Le modèle saturé peut être obtenu comme cas particulier de l'équation (1.2) en définissant $p = n$ et en laissant x_i prendre la valeur 1 pour l'unité i et 0 sinon. Dans ce modèle, les x sont des variables indicatrices pour les différentes unités, et il n'y a plus de variation aléatoire.
Évidemment, les modèles nul et saturé ne sont pas très utiles. La plupart des modèles statistiques se situent entre les deux.

Estimation des paramètres

Considérons pour l'instant un modèle abstrait où $\mu_i = x_i'\beta$.
L'objectif est d'estimer les paramètres β et σ^2 à partir des données. Pour ce faire, nous utilisons la fonction de vraisemblance (ou plus simplement vraisemblance), qui est une fonction des paramètres d'un modèle statistique calculée à partir de données observées.
La fonction de vraisemblance est définie en fonction d'un vecteur de paramètres $\theta = (\beta, \sigma^2)$ comme la densité des données observées par rapport à une mesure de probabilité discrète ou continue.

Estimation de $\boldsymbol{\beta}$

Le principe de vraisemblance nous demande de choisir les valeurs des paramètres qui maximisent la vraisemblance, ou de manière équivalente, le logarithme de la fonction de vraisemblance. Si les observations sont indépendantes, alors la fonction de vraisemblance est un produit des densités normales. En introduisant le logarithme, nous obtenons la log-vraisemblance normale [14] :

$$\log \boldsymbol{l}\left(\boldsymbol{\beta}, \boldsymbol{\sigma^2}\right) = -\frac{\boldsymbol{n}}{\mathbf{2}} \log\left(\mathbf{2}\boldsymbol{\pi}\boldsymbol{\sigma^2}\right) - \frac{\mathbf{1}}{\mathbf{2}} \sum \left(\boldsymbol{y_i} - \boldsymbol{\mu_i}\right)^{\mathbf{2}} / \boldsymbol{\sigma^2}. \tag{1.4}$$

où $\mu_i = x_i'\beta$. Maximiser la log-vraisemblance par rapport aux paramètres linéaires β pour une valeur fixe de σ^2 est équivalent à minimiser la somme des carrés des différences entre les valeurs observées y_i et μ_i, appelée également la somme des carrés résiduels (RSS)

$$\mathbf{RSS}(\boldsymbol{\beta}) = \sum \left(\boldsymbol{y_i} - \boldsymbol{\mu_i}\right)^{\mathbf{2}} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{1.5}$$

En d'autres termes, nous devons choisir des valeurs de β qui rapprochent le plus possible les valeurs ajustées $\mu_i = x_i'\beta$ aux valeurs observées y_i.
Prendre des dérivées de la somme des carrés résiduels (RSS) par rapport à β et définir la dérivée égale à zéro conduit aux équations dites normales pour l'estimateur du maximum de vraisemblance $\hat{\beta}$:

$$\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X'y}$$

Si la matrice du modèle X est de rang complet, de sorte qu'aucune colonne n'est une combinaison linéaire exacte des autres, alors la matrice $X'X$ est de rang complet et peut être inversé pour résoudre les équations normales. Cela donne une formule explicite pour l'estimateur des moindres carrés ordinaires (MCO) ou du maximum de vraisemblance des paramètres linéaires :

$$\hat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1}\,X'y.} \tag{1.6}$$

Si X n'est pas de rang complet, on peut utiliser des inverses généralisés, mais l'interprétation des résultats est beaucoup plus simple si l'on élimine simplement les colonnes redondantes.
Il existe plusieurs méthodes numériques pour résoudre les équations normales, y compris des méthodes qui opèrent sur $X'X$, comme l'élimination de Gauss ou la décomposition de Choleski, et des méthodes qui tentent de simplifier le calculs en factorisant la matrice du modèle X, y compris les réflexions de Householder, les rotations de Givens et l'orthogonalisation de Gram-Schmidt. Pour plus de détails, voir [15].

Les résultats précédents ont été obtenus en maximisant la log-vraisemblance par rapport à β pour une valeur fixe de σ^2. Le résultat (1.6) ne dépend pas de σ^2, et est donc un maximum global.

Propriétés de l'estimateur

Les estimateurs des moindres carrés sont non biaisés :

$$\mathbf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}. \tag{1.7}$$

On peut également montrer que si les observations ne sont pas corrélées et ont une variance constante σ^2, la matrice de variance-covariance de l'estimateur MCO s'écrit :

$$\mathbf{var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{(X'X)^{-1}\,\sigma^2}. \tag{1.8}$$

Une autre propriété de l'estimateur MCO est que sa variance est la plus faible parmi tous les estimateurs non biaisés, c'est-à-dire c'est le meilleur estimateur linéaire sans biais (BLEU). Puisqu'aucun autre estimateur sans biais ne peut avoir une variance plus faible pour une taille d'échantillon fixe, nous disons que les estimateurs MCO sont efficaces.

Estimation de σ^2

La substitution de l'estimateur MCO de β dans (1.4) donne une vraisemblance profilée pour σ^2

$$\log l\left(\sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\,\mathrm{RSS}(\hat{\beta})/\sigma^2$$

Différencier cette expression par rapport à σ^2 et fixer la dérivée à zéro conduit à l'estimateur du maximum de vraisemblance

$$\hat{\sigma}^2 = \mathrm{RSS}(\hat{\beta})/n$$

Cet estimateur se trouve être biaisé, mais le biais est facilement corrigé en divisant par $n - p$ au lieu de n. La situation est exactement analogue à l'utilisation de $n - 1$ au lieu de n lors de l'estimation d'une variance.

1.1.2 Modèle linéaire généralisé (GLM)

En statistiques, le modèle linéaire généralisé (GLM) est une généralisation flexible de la régression linéaire. Les modèles linéaires généralisés ont été formulés par John Nelder et Robert Wedderburn [16] comme un moyen d'unifier les modèles statistiques y compris la régression linéaire, la régression logistique et la régression de Poisson. Ils proposent une méthode itérative dénommée méthode des moindres carrés repondérés itérativement [17] pour l'estimation du maximum de vraisemblance des paramètres du modèle.

Présentation du modèle

Soit $y_1, \ldots, y_n$ n observations indépendantes. Nous traitons y_i comme une réalisation d'une variable aléatoire Y_i.
Dans le modèle linéaire généralisé, nous supposons que Y_i a une distribution normale de moyenne μ_i et variance σ^2

$$Y_i \sim \mathrm{N}\left(\mu_i, \sigma^2\right)$$

Nous supposons en outre que μ_i une fonction linéaire de p prédicteurs telle que

$$\mu_i = \mathbf{x}_i' \boldsymbol{\beta}$$

avec $\boldsymbol{\beta}$ un vecteur de paramètres inconnus. Nous généraliserons cela en deux étapes, en traitant les composantes aléatoires et déterministe du modèle.

Famille exponentielle

Nous supposons que les observations proviennent d'une distribution qui appartient à la famille exponentielle [18, 19] :

$$f\left(y_i\right) = \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{a_i(\phi)} + c\left(y_i, \phi\right)\right\}$$

θ_i et ϕ sont des paramètres et $a_i(\phi), b\left(\theta_i\right)$ et $c\left(y_i, \phi\right)$ sont des fonctions connues. La fonction $a_i(\phi)$ est définie par

$$a_i(\phi) = \phi / p_i$$

avec p_i des poids connus, généralement $p_i = 1$.
θ_i est appelé paramètre naturel de la famille exponentielle. ϕ est appelé paramètre de dispersion. Si Y_i est défini par une distribution appartenant à la famille exponentielle alors :

$$\begin{aligned} \mathrm{E}\left(Y_i\right) &= \mu_i = b'\left(\theta_i\right) \\ \mathrm{var}\left(Y_i\right) &= \sigma_i^2 = b''\left(\theta_i\right) a_i(\phi) \end{aligned}$$

avec $b'\left(\theta_i\right)$ et $b''\left(\theta_i\right)$ sont les dérivées premières et seconde de $b\left(\theta_i\right)$.
Quand $a_i(\phi) = \phi / p_i$, la variance est définie simplement par :

$$\mathrm{var}\left(Y_i\right) = \sigma_i^2 = \phi b''\left(\theta_i\right) / p_i$$

La famille exponentielle inclut comme cas particuliers les distributions gaussiennes normales, binomiales, de Poisson, exponentielles, gamma et inverses.

Exemple 1.1. Distribution gaussienne

La densité de probabilité de la loi normale s'écrit :

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(y_i - \mu_i)^2}{\sigma^2} \right\}$$

Nous avons : $(y_i - \mu_i)^2 = y_i^2 + \mu_i^2 - 2y_i\mu_i$ alors le coefficient de y_i est μ_i/σ^2.
Nous identifions θ_i avec μ_i et ϕ avec σ^2, où $a_i(\phi) = \phi$.
Nous réécrivons :

$$f(y_i) = \exp\left\{ \frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right) \right\}$$

Alors : $b(\theta_i) = \frac{1}{2}\theta_i^2$ où $(\theta_i = \mu_i)$.
La moyenne et la variance sont données par :

$$E(Y_i) = b'(\theta_i) = \theta_i = \mu_i$$
$$\text{var}(Y_i) = b''(\theta_i)\, a_i(\phi) = \sigma^2$$

La fonction lien

Le deuxième élément de la généralisation est qu'au lieu de modéliser la moyenne, nous allons introduire une transformation injective, différentiable et continue $g(\mu_i)$ et nous nous intéressons à :

$$\eta_i = g(\mu_i)$$

La fonction $g(\mu_i)$ est appelée fonction lien. Elle exprime une relation fonctionnelle entre la composante déterministe et la composante aléatoire.
Des exemples de la fonction lien incluent l'identité, le log, logit i.e. $ln(\pi/(1-\pi))$, etc.

Nous supposons en outre que la moyenne transformée suit un modèle linéaire tel que :

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$$

η_i est appelé prédicteur linéaire.

Puisque la fonction lien est injective alors :

$$\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$$

Le modèle de μ_i est généralement plus compliqué que le modèle de η_i. Notez que nous ne transformons pas la réponse y_i, mais μ_i.

1.1.3 Modèle additif (MA)

En statistique, un modèle additif (MA) est une méthode de régression non paramétrique. Il a été suggéré par Friedman et Werner [20]. Il peut être considéré comme une généralisation de la régression linéaire [21].
Le modèle additif utilise un lissage unidimensionnel pour créer une classe restreinte de modèles de régression non paramétriques. Il est plus flexible qu'un modèle linéaire standard,

tout en étant plus interprétable qu'une surface de régression générale au prix d'erreurs d'approximation.
Étant donné un ensemble de données $\{y_i, x_{i1}, ..., x_{ip}\}_{i=1}^{n}$, le modèle additif a la structure suivante :

$$E[y_i|x_{i1}, ..., x_{ip}] = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij})$$

ou bien

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon$$

Avec $\mathrm{E}(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$ et $\mathrm{E}[f_j(X_j)] = 0$.
Les fonctions $f_j(x)$ sont des fonctions de lissage inconnues ajustées à partir des données. Ajuster le modèle additif, c-à-d les fonctions $f_j(x_{ij})$, peut se faire en utilisant l'algorithme de backfitting [21].

1.1.4 Modèle additif généralisé (GAM)

Les modèles additifs généralisés (GAM), proposés par Hastie et Tibshirani [22], sont une extension des modèles linéaires généralisés (GLM). Le GAM fournit une structure pour généraliser un modèle linéaire général en permettant l'additivité des fonctions non linéaires des variables. Le concept du modèle additif (MA) avec le GLM peut être combiné pour obtenir le GAM.
Particulièrement, on suppose que la loi de Y appartient à la famille exponentielle avec une moyenne $\mu = E(y|x_1, x_2, ..., x_p)$ liée au prédicteur par :

$$g(\mu) = \sum_{j=1}^{p} f_j\left(x_j\right)$$

où $f_1(.), ..., f_p(.)$ des fonctions de lissage.
Hastie et al. [22], ont montré que l'algorithme Backfitting converge toujours pour ces fonctions.
Les GAM permettent une plus grande flexibilité que les GLM [23]. Le prédicteur reste sous forme linéaire, mais nous souhaitons désormais expliquer la variable étudiée par des fonctions de variables explicatives. Ces fonctions représentent une manière de traiter la relation non linéaire que peuvent avoir certaines variables avec la variable à expliquer.

1.1.5 Modèle linéaire à effets mixtes (LMM)

1.1.5.1 Présentation générale et hypothèses

Un modèle à effets mixtes [24] est un modèle qui considère à la fois des effets fixes et des effets aléatoires. Le mélange entre les deux est à l'origine du nom. Les effets fixes décrivent les relations entre les covariables et la variable dépendante pour une population entière tandis que les effets aléatoires sont spécifiques à l'échantillon. En d'autres termes, un effet aléatoire [25] est un effet dont nous ne voulons pas généraliser les propriétés et un effet fixe est un effet dont on veut généraliser les propriétés et en tirer des conclusions. Les effets aléatoires doivent nécessairement être des variables catégorielles. Les effets fixes sont représentés par des coefficients de régression. Ces effets décrivent les relations entre la variable dépendante et les prédicteurs. Nous supposons que les effets fixes sont inconnus et que nous les estimons sur la base des données. Les estimateurs aléatoires représentent

une déviation de la relation décrite par ces effets fixes. Les coefficient des effets aléatoires ne sont pas explicitement estimés. Néanmoins, il est possible de le faire. L'intérêt d'une telle estimation est de pouvoir faire des inférences sur la variabilité des effets aléatoires. Mathématiquement, le modèle à effets mixtes est défini par :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{e}. \tag{1.9}$$

- y est le vecteur des observations de dimension $n \times 1$.
- X est une matrice $n \times p$ connue associée aux effets fixes.
- β est le vecteur des effets fixes de dimension $p \times 1$.
- $Z = [Z_1, \ldots, Z_b]$, avec Z_i la matrice $n \times q_i$ associée au i^{em} effet aléatoire.
- $u = [u_1', \ldots, u_b']'$ est un vecteur $q \times 1$ des effets aléatoires avec u_i de dimension $q_i \times 1$ tel que $q = \sum_{i=1}^{b} q_i$ avec $\mathrm{E}(u) = 0$.
- e est un vecteur des erreurs de dimension $n \times 1$, avec $\mathrm{E}(e) = 0$.
- u et e suivent des distributions gaussiennes indépendantes et multivariées telles que $\begin{bmatrix} u \\ e \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} G(\gamma) & 0 \\ 0 & R(\rho) \end{bmatrix}\right)$ avec γ et ρ des vecteurs de paramètres de variance inconnus correspondant à u et e, de dimension $r \times 1$ et $s \times 1$ ($s \leqslant n(n+1)/2$) respectivement.

D'après Patterson and Thompson [26], nous pouvons réécrire la matrice de variance-covariance comme suit :

$$\boldsymbol{var(y) = \sigma^2 \left(ZGZ' + R\right) = \sigma^2 H.} \tag{1.10}$$

avec

$$\boldsymbol{H = \mathbf{Z}G\mathbf{Z}' + R}$$

1.1.5.2 Estimation jointe des effets fixes et aléatoires : Paramètres de la matrice de covariance connus

Une fois le modèle formulé, des méthodes sont nécessaires pour estimer les paramètres du modèle mixte. Dans cette section, nous traitons d'abord l'estimation conjointe des effets fixes β et aléatoires u. L'estimation des paramètres de variance $\gamma, \rho.$ et σ^2 sera étudiée dans la section suivante. Il existe de nombreuses méthodes pour obtenir les estimations des effets fixes et aléatoires simultanément [27]. Ces méthodes comprennent les équations mixtes de Henderson [28], l'approche de Goldberger [29] pour la prédiction d'une observation future, les techniques basées sur la régression en deux étapes, la linéarité en y, le partitionnement de y et l'estimation de Bayes.

Dans cette section, nous étudions l'estimation des paramètres via les équations mixtes de Henderson. Notre choix est motivé par le fait que cette méthode, produit un lien avec l'estimation du maximum de vraisemblance des paramètres de variance.

Henderson a supposé que u et y sont conjointement normalement distribuées :

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} \sim \boldsymbol{N}\left(\begin{bmatrix} \mathbf{0} \\ \boldsymbol{X\beta} \end{bmatrix}, \boldsymbol{\sigma^2} \begin{bmatrix} \boldsymbol{G} & \boldsymbol{GZ'} \\ \boldsymbol{ZG} & \boldsymbol{H} \end{bmatrix}\right). \tag{1.11}$$

y est définie par la densité marginale $N\left(X\beta, \sigma^2 H\right)$, G et R sont supposées connues.
Henderson [28] a maximisé le log de la densité conjointe de y et u pour estimer β et u.

La distribution marginale de u est définie par :

$$\boldsymbol{u \sim N\left(0, \sigma^2 G\right)}$$

La distribution conditionnelle de y sachant u est définie par :

$$\boldsymbol{y \mid u \sim N\left(X\beta + Zu, \sigma^2 R\right)}$$

Alors le log de la densité conjointe de y et u est donné par :

$$\begin{aligned}
\log f(y,u) &= \log f(y \mid u) + \log f(u) \\
&= -\frac{1}{2}\left\{n \log \sigma^2 + \log R + (y - X\beta - Zu)' R^{-1} (y - X\beta - Zu)/\sigma^2\right\} \\
&\quad -\frac{1}{2}\left\{q \log \sigma^2 + \log G + u' G^{-1} u/\sigma^2\right\} \\
&= -\frac{1}{2}\left\{(n+q) \log \sigma^2 + \log R + \log G + (y - X\beta)' R^{-1} (y - X\beta)/\sigma^2\right\} \\
&\quad -\frac{1}{2\sigma^2}\left\{u'\left(ZR^{-1}Z' + G^{-1}\right) u - 2(y - X\beta)' R^{-1} Zu\right\}
\end{aligned}$$

Les estimations de β et u sont obtenues en résolvant les équations :

$$\left\{\begin{array}{rl}
\boldsymbol{X'R^{-1}(y - X\hat{\beta}) - X'R^{-1}Z\tilde{u} =} & \boldsymbol{0} \\
\boldsymbol{Z'R^{-1}(y - X\hat{\beta}) - \left(Z'R^{-1}Z + G^{-1}\right)\tilde{u} =} & \boldsymbol{0}
\end{array}\right. \tag{1.12}$$

(1.12) représente les équations du modèle mixte proposées par Henderson [28, 30]. Elles peuvent être exprimées sous forme matricielle par :

$$\left[\begin{array}{cc}
\boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}Z} \\
\boldsymbol{Z'R^{-1}X} & \boldsymbol{Z'R^{-1}Z + G^{-1}}
\end{array}\right]\left[\begin{array}{c}
\boldsymbol{\hat{\beta}} \\ \boldsymbol{\tilde{u}}
\end{array}\right] = \left[\begin{array}{c}
\boldsymbol{X'R^{-1}y} \\ \boldsymbol{Z'R^{-1}y}
\end{array}\right] \tag{1.13}$$

Nous avons $\mathrm{E}(y) = X\beta$ et $\mathrm{var}(y) = \sigma^2 H$. En supposant que H est connue, le paramètre β est estimé par les moindres carrés généralisés (GLS).

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

est le meilleur estimateur linéaire non biaisé (BLUE) de β.
Si X n'est pas de plein rang, nous pouvons utiliser n'importe quel inverse généralisé $(X'H^{-1}X)^-$ à la place de $(X'H^{-1}X)^{-1}$ pour obtenir une solution de β.
Dans ce cas, β n'est pas unique et est biaisé. En revanche, $X\hat{\beta}$ est unique et est non biaisé.
Dans le cas où G et R sont connues, les estimations de β et u sont données par :

$$\left\{\begin{array}{rl}
\hat{\beta} = & \left(X'H^{-1}X\right)^{-1} X'H^{-1}y \\
\tilde{u} = & GZ'H^{-1}(y - X\hat{\beta})
\end{array}\right.$$

et :

$$\left\{\begin{array}{rl}
\mathrm{var}(\hat{\beta}) = & \sigma^2 \left(X'H^{-1}X\right)^{-1} \\
\mathrm{var}(\tilde{u}) = & \sigma^2 GZ'PZG
\end{array}\right.$$

avec :

$$\boldsymbol{\mathrm{P} = H^{-1} - H^{-1}X\left(X'H^{-1}X\right)^{-1} X'H^{-1}.} \tag{1.14}$$

Jusqu'à présent, les résultats obtenus supposent que les paramètres de variance sont connus. En revanche, si les paramètres γ, ρ et σ^2 sont inconnus, G, R et σ^2 doivent être remplacés par leurs estimations $\hat{G}, \hat{R}$ et $\hat{\sigma}^2$ pour obtenir les estimations des effets fixes, des effets aléatoires et les erreurs standards.
Dans la section suivante, nous discutons les méthodes d'estimation de paramètres de variance γ, ρ et σ^2.

1.1.5.3 Estimation des paramètres de variance

Le maximum de vraisemblance (ML) et le maximum de vraisemblance restreint (REML) sont des méthodes standard d'estimation des paramètres de variance. Dans cette section, nous décrivons et comparons les estimations ML et REML pour l'estimation des paramètres de variance dans les modèles à effets mixtes.

Maximum de vraisemblance (ML)

Le maximum de vraisemblance est une méthode qui permet d'obtenir des estimations de paramètres inconnus en optimisant la fonction de vraisemblance. Le principe de la vraisemblance est d'estimer la probabilité d'observer des données de manière itérative. Ainsi, Nous partons d'une valeur "aléatoire" que nous déplaçons de sorte à augmenter la probabilité d'observer cette distribution de données. Quand cette probabilité a atteint un maximum (moyennant une certaine tolérance), on dit que le modèle a convergé.
Dans un modèle à effets mixtes, la distribution marginale de y est donnée par $N\left(X\beta, \sigma^2 H\right)$. La fonction de log-vraisemblance marginale de y s'écrit [31] :

$$\boldsymbol{l_{\mathrm{ML}}(\beta, \phi; y) = -\frac{1}{2}\left\{ n \log(2\pi) + n \log \sigma^2 + \log |H| + \frac{(y - X\beta)' H^{-1} (y - X\beta)}{\sigma^2} \right\}.} \tag{1.15}$$

avec $\phi = \left(\kappa', \sigma^2\right)'$, $\kappa = (\gamma', \rho')'$.

En dérivant la log-vraisemblance marginale de y par rapport à β, σ^2 et κ_j, $j = 1, \ldots, r+s$, nous obtenons :

$$\boldsymbol{\frac{\partial l_{\mathrm{ML}}(\beta, \phi; y)}{\partial \beta} = -\frac{1}{\sigma^2}\left(X' H^{-1} X\beta - X' H^{-1} y\right).} \tag{1.16}$$

$$\boldsymbol{\frac{\partial l_{\mathrm{ML}}(\beta, \phi; y)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{y - X\beta)' H^{-1} (y - X\beta)}{2\sigma^4}.} \tag{1.17}$$

$$\boldsymbol{\frac{\partial l_{\mathrm{ML}}(\beta, \phi; y)}{\partial \kappa_j} = -\frac{1}{2} \operatorname{tr}\left(H^{-1} \dot{H}_j\right) + \frac{(y - X\beta)' H^{-1} \dot{H}_j H^{-1} (y - X\beta)}{2\sigma^2}.} \tag{1.18}$$

avec $\dot{H}_j = \partial H / \partial \kappa_j$
En annulant les équations (1.16), (1.17) et (1.18), nous obtenons :

$$\boldsymbol{X' \hat{H}^{-1} X \hat{\beta} = X' \hat{H}^{-1} y.} \tag{1.19}$$

$$\boldsymbol{n \hat{\sigma}^2 = (y - X\hat{\beta})' \hat{H}^{-1} (y - X\hat{\beta}).} \tag{1.20}$$

$$\boldsymbol{\operatorname{tr}\left(\hat{H}^{-1} \widehat{H}_j\right) = \frac{1}{\hat{\sigma}^2} (y - X\hat{\beta})' \hat{H}^{-1} \hat{H}_j \hat{H}^{-1} (y - X\hat{\beta}).} \tag{1.21}$$

$\hat{H}$ et $\widehat{H}_j$ fournissent les estimateurs ML de $\kappa_j, j = 1, \ldots, r+s$.
Le nombre de paramètres de variance dans le modèle, y compris σ^2 est $t = r + s + 1$.
Les estimateurs ML de $\hat{\beta}$ et $\hat{\sigma}^2$ sont donnés par :

$$\boldsymbol{\hat{\beta} = \left(X' \hat{H}^{-1} X\right)^{-1} X' \hat{H}^{-1} y.} \tag{1.22}$$

$$\hat{\boldsymbol{\sigma}}^{\mathbf{2}} = \frac{\mathbf{1}}{\boldsymbol{n}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{H}}^{-\mathbf{1}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{1.23}$$

Les paramètres κ_j dépendent de $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}^2$, et donc sont calculés itérativement comme suit :
Étape 0 : initialisation $\kappa_{(0)} = (\kappa_1, \ldots, \kappa_{r+s})'$
Étape 1 : pour chaque itération remplacer $\hat{\kappa}_{(m-1)}$ dans l'expression de $\hat{\boldsymbol{\beta}}_{(m)}$ et $\hat{\sigma}^2_{(m)}$.
Étape 2 : Utiliser les résultats de l'étape 0 et 1, i.e. remplacer $\hat{\kappa}_{(m-1)}, \hat{\boldsymbol{\beta}}_{(m)}$ et $\hat{\sigma}^2_{(m)}$ dans $\partial l_{\mathrm{ML}}/\partial\kappa_j$ pour calculer $\hat{\kappa}_{(m)}$.
Étape 3 : Répéter les étapes 0, 1 et 2 jusqu'à convergence.

Maximum de vraisemblance restreinte (REML)

Un des biais du maximum de vraisemblance est que les estimations de certains paramètres pour les effets aléatoires ne prennent pas en compte la perte de degrés de liberté (ddl) qui résulte de l'estimation des paramètres des effets fixes. Un moyen d'éliminer ce biais est d'utiliser le maximum de vraisemblance restreinte (REML) [32].
L'avantage du REML est qu'il produit des estimations non biaisées des paramètres de covariances en prenant en compte la perte de ddl qui résulte de l'estimation des effets fixes.
Dans le contexte du modèle linéaire mixte, l'estimation ML de σ^2 est RSS$/n$, où RSS désigne les sommes résiduelles des carrés, tandis que l'estimation REML est égale à RSS $/(np)$.
D'un point de vue bayésien, Harville [33] a montré que n'utiliser que des contrastes d'erreur pour faire des inférences sur les paramètres de variance équivaut à ignorer toute information préalable sur les paramètres des effets fixes.
Verbyla [34] a montré que le REML peut également être considérée comme une vraisemblance marginale, alors que Barndoff-Nielsen [35] l'a considérée comme une log-vraisemblance profilée modifiée. Lee et [36] ont considéré le REML comme une vraisemblance conditionnelle en supposant une distribution gaussienne asymptotique (multivariée) pour les estimations des effets fixes, sachant les valeurs des paramètres de variance.
La fonction de log-vraisemblance REML (en ignorant les constantes) pour le modèle mixte est donnée par :

$$\boldsymbol{l}_{\mathbf{R}}(\boldsymbol{\phi}; \boldsymbol{y}) = -\frac{\mathbf{1}}{\mathbf{2}}\left\{(\boldsymbol{n} - \boldsymbol{p})\log\boldsymbol{\sigma}^{\mathbf{2}} + \log|\boldsymbol{H}| + \log\left|\boldsymbol{X}'\boldsymbol{H}^{-\mathbf{1}}\boldsymbol{X}\right| + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\boldsymbol{\sigma}^{\mathbf{2}}}\right\}. \tag{1.24}$$

avec $\hat{\beta}$ l'estimation GLS de β, et P définit par (1.14).

En dérivant (1.24) par rapport à σ^2 et κ_j, $j = 1, \ldots, r+s$, nous obtenons [37] :

$$\frac{\partial \boldsymbol{l}_{\mathbf{R}}(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \boldsymbol{\sigma}^{\mathbf{2}}} = -\frac{\boldsymbol{n} - \boldsymbol{p}}{\mathbf{2}\boldsymbol{\sigma}^{\mathbf{2}}} + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\mathbf{2}\boldsymbol{\sigma}^{\mathbf{4}}}. \tag{1.25}$$

$$\frac{\partial \boldsymbol{l}_{\mathbf{R}}(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \boldsymbol{\kappa}_{\boldsymbol{j}}} = -\frac{\mathbf{1}}{\mathbf{2}}\left\{\mathbf{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) - \frac{\mathbf{1}}{\boldsymbol{\sigma}^{\mathbf{2}}}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}\right\}. \tag{1.26}$$

D'où :

$$\hat{\boldsymbol{\sigma}}^{\mathbf{2}} = \frac{\boldsymbol{y}'\hat{\boldsymbol{P}}\boldsymbol{y}}{\boldsymbol{n} - \boldsymbol{p}}. \tag{1.27}$$

(1.27) doit être calculé itérativement car elle dépend de $\hat{\kappa}$ via $\hat{P}$ [37, 38].

1.1.5.4 Algorithmes de maximisation du Maximum de vraisemblance

Nous décrivons quatre procédures itératives connexes, qui sont utilisées pour le calcul des estimations ML ou REML des paramètres de variance, à savoir : les algorithmes de Newton-Raphson (NR), Fisher Scoring (FS) et le Expectation Maximization algorithm (E-M).

L'algorithme de Newton-Raphson (NR)

L'algorithme NR [39] est l'algorithme le plus souvent utilisé. Cet algorithme minimise $-2 * log\ vraisemblance$. NR utilise le développement du premier ordre de la fonction score autour d'une estimation $\phi_{(m)}$ pour prédire $\phi_{(m+1)}$.
Cet algorithme suppose la concavité de la fonction log-vraisemblance pour obtenir l'approximation quadratique de la fonction. Chaque itération NR nécessite le calcul de la fonction score et de sa dérivée.
La procédure NR peut être décrite comme suit :

Considérons la fonction log-vraisemblance $l(\phi)$ que nous cherchons à maximiser :

$$\boldsymbol{\frac{\partial l(\phi)}{\partial \phi} = 0.} \tag{1.28}$$

Le développement du premier ordre de $\frac{\partial l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}$ est :

$$\frac{\partial l(\phi)}{\partial \phi} = U(\phi) \approx U\left(\phi_{(0)}\right) + \frac{\partial^2 l(\phi)}{\partial \phi \partial \phi'}\left(\phi - \phi_{(0)}\right)$$

Alors :

$$U\left(\phi_{(0)}\right) + \frac{\partial^2 l(\phi)}{\partial \phi \partial \phi'}\left(\phi - \phi_{(0)}\right) = 0$$

Nous obtenons :

$$\boldsymbol{\phi = \phi_{(0)} - \left[\frac{\partial^2 l(\phi)}{\partial \phi \partial \phi'}\right]^{-1} U\left(\phi_{(0)}\right).} \tag{1.29}$$

(1.29) peut être utilisée de manière itérative pour trouver l'estimation du maximum jusqu'à la $(m+1)$ ème l'itération :

$$\begin{aligned}\phi_{(m+1)} &= \phi_{(m)} - \left[\frac{\partial^2 l(\phi)}{\partial \phi \partial \phi'}\right]^{-1} U\left(\phi_{(m)}\right) \\ &= \phi_{(m)} + \left[I_{\mathcal{O}(m)}\right]^{-1} U\left(\phi_{(m)}\right)\end{aligned}$$

$\mathcal{I}_{\mathcal{O}(m)}$ est la matrice d'information observée évaluée en $\phi_{(m)}$

L'algorithme Fisher Scoring (FS)

L'algorithme FS [40] est un variant de l'algorithme NR. Il remplace la matrice d'information observée dans l'algorithme NR par la matrice $\mathrm{E}\left[-\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\right]$ (expected information matrix).

L'algorithme Expectation Maximization (E-M)

Dempster et al. [41] ont introduit l'algorithme EM pour l'estimation des paramètres dans les modèles avec des données incomplètes (manquantes).
Dempster et al. [42] ont également montré comment cet algorithme est utilisé pour obtenir des estimations du maximum de vraisemblance des composantes de la variance dans le modèle à effets mixtes.
La mise en oeuvre de cette méthode dans le modèle à effets mixtes est basée sur la visualisation des effets aléatoires comme des données non observées ou manquantes.
L'algorithme EM se compose essentiellement de deux étapes : une étape d'espérance (étape E) et une étape de maximisation (étape M). Ces étapes sont décrites comme suit :
Étape E : Utiliser $\phi_{(m)}$, pour évaluer la log-vraisemblance de la distribution conditionnelle $u \mid y$ et calculer l'espérance de la log-vraisemblance pour une nouvelle valeur de ϕ.
Étape M : Maximiser les espérances obtenues de l'étape E par rapport à ϕ pour obtenir $\phi_{(m+1)}$.
La procédure EM est répétée jusqu'à convergence.

Comparaison des méthodes itératives pour obtenir les estimations des composantes de la variance

Méthodes	Avantages	Inconvénients
NR	• Converge plus rapidement que E-M. • Donne les erreurs standards asymptotiques pour les estimations.	• Calcul intensif par rapport au E-M (instable loin du maximum).
FS	• Converge plus rapidement que E-M. • Donne les erreurs standards asymptotiques pour les estimations. • Valeurs de départ robustes à faibles comparé à NR.	• Tendance à converger vers des valeurs en dehors de l'espace des paramètres. • Intensif en calcul par rapport à E-M (instable loin du maximum).
E-M	• Stable numériquement.	• Faible taux de convergence. • Ne donne pas les erreurs standards pour les estimations.

1.1.6 Modèle linéaire mixte généralisé (GLMM)

En statistique, un modèle linéaire mixte généralisé (GLMM) [43, 44, 45] est une extension du modèle linéaire généralisé (GLM) dans lequel le prédicteur linéaire contient des effets aléatoires en plus des effets fixes habituels. Le GLMM étend également le modèle mixte linéaire (LMM) aux données non normales.
Dans les GLMM, y est défini par une distribution appartenant à la famille exponentielle [46].

$$\boldsymbol{ln\ p(y|u) = \sum \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)}$$

Soit le prédicteur linéaire, η, défini comme la combinaison des effets fixes et aléatoires excluant les résidus.

$$\boldsymbol{\eta = X\beta + Zu}$$

Notons $g(\cdot)$ la fonction de lien. Cette fonction relie y au prédicteur linéaire η :

$$\begin{aligned} &\eta = X\beta + Zu \\ &g(\cdot) = \text{fonction lien} \\ &h(\cdot) = g^{-1}(\cdot) = \text{ inverse de la fonction lien} \end{aligned}$$

Notre modèle pour l'espérance conditionnelle de y est défini par :

$$g(E(y|u)) = \eta$$

Nous pourrions également modéliser l'espérance de y :

$$E(y|u) = h(\eta) = \mu$$

avec y définie par :

$$y = h(\eta) + \varepsilon$$

La vraisemblance $ln\ p(y,u) = ln \int p(y|u)p(u)du$ n'a pas une forme analytique explicite, et l'intégration sur les effets aléatoires est généralement intensive en calcul.
En plus d'approximer numériquement cette intégrale (par exemple via la méthode de quadrature de Gauss-Hermite), des méthodes motivées par l'approximation de Laplace ont été proposées [43].

1.1.7 Modèle mixte additif généralisé (GAMM)

Les modèles mixtes additifs généralisés (GAMM) [47] sont une extension des modèles additifs généralisés incorporant des effets aléatoires.
Supposons que nous avons des observations y_i, des covariables $x_i = (1, x_{i1}, \ldots, x_{ip})'$ associées à des effets fixes et un vecteur z_i de dimension $q \times 1$ associé à des effets aléatoires. Etant donné un vecteur u d'effets aléatoire de dimension $q \times 1$, nous supposons que $E(y_i \mid u) = \mu_i^u$ et $\text{var}(y_i \mid u) = \phi p_i^{-1} v(\mu_i^u)$, les p_i sont des poids connus et ϕ un paramètre de dispersion.
Le modèle mixte additif généralisé (GAMM) s'écrit :

$$\boldsymbol{g\left(\mu_i^u\right) = \beta_0 + f_1\left(x_{i1}\right) + \ldots + f_p\left(x_{ip}\right) + z_i'u}$$

avec $g(\cdot)$ une fonction lien différentiable monotone. $f_j(\cdot)$ des fonctions de lissage deux fois différentiables, les effets aléatoires u sont distribués suivant une $N\{0, D(\theta)\}$ et θ est un

vecteur des composantes de la variance de dimension $c \times 1$. Une caractéristique clé du GAMM est que des fonctions non paramétriques additives sont utilisées pour modéliser les effets des covariables, et les effets aléatoires sont utilisés pour modéliser la corrélation entre les observations. Si $f_j(\cdot)$ est une fonction linéaire, le GAMM se réduit au GLMM [43].

1.1.8 Modèle espace-état

Dans cette partie, nous rappelons et discutons les principaux résultats des modèles espace-état. Nous commençons par définir le modèle espace-état et ses généralités. Ensuite nous présentons les algorithmes, Filtre de Kalman et algorithme dit E-M (Expected-Maximisation), permettant d'estimer les variables cachées et les paramètres de ce modèle.

1.1.8.1 Présentation générale et hypothèses

Les modèles espace-état [48] distinguent entre les variables observées (le signal) et les variables cachées (l'état interne). Ils s'expriment sous forme :
- D'une ou plusieurs équation(s) de mesure décrivant la manière dont les variables observées sont générées par les variables cachées et les résidus.
- D'une ou plusieurs équation(s) d'état décrivant la manière dont les variables cachées sont générées à partir de leur retard et innovations.
Nous utilisons la terminologie suivante pour le reste de cette partie ; à la date t :

Y_t est appelé observation ou variable de mesure.

Z_t est la variable d'état.

ϵ_t est le vecteur des innovations.

η_t est le vecteur des erreurs de mesures.

A_t est la matrice de transition.

C_t est la matrice de mesure.

$X_{1,t}, X_{2,t}$ sont des variables exogènes, prédéterminées.

$C_t Z_t$ est le signal.

Nous appelons modèle espace-état, le système décrit par :

$$\mathbf{Z_{t+1}} = \mathbf{A_t Z_t} + \mathbf{B_t X_{1,t}} + \boldsymbol{\varepsilon_t} \tag{1.30a}$$
$$\mathbf{Y_t} = \mathbf{C_t Z_t} + \mathbf{D_t X_{2,t}} + \boldsymbol{\eta_t} \tag{1.30b}$$

avec :

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim NID\left(0, \begin{pmatrix} Q_t & S_t \\ S_t' & R_t \end{pmatrix}\right)$$

où les matrices A_t et C_t sont de taille $k \times k$ et $n \times k$ respectivement, B_t et D_t sont des matrices déterministes de taille $k_1 \times k$ et $k_2 \times k$ respectivement et Z_0 est un vecteur aléatoire de loi $N(m, p)$ indépendant du bruit blanc normal.

Hypothèses

Les modèles espace-état reposent sur un certain nombre d'hypothèses principales :
1) Les équations de mesure et d'état sont linéaires.
2) Les bruits d'observation et d'innovation sont des bruits blancs.
3) Les variables cachées suivent à un instant initial donné une loi gaussienne. À ces dernières, se sont ajoutées des hypothèses secondaires permettant de déterminer la forme canonique (voir définition 1.2).
4) L'indépendance entre les bruits d'observation et d'innovation.
5) L'indépendance entre la variable cachée initiale et ces bruits.
Notons que toutes ces hypothèses sont destinées à simplifier les procédures d'estimation.

Notons que Y_t n'admet pas une unique representation espace-état. En effet, s'il existe une représentation de vecteur d'état Z_t, nous pouvons facilement formuler une autre représentation Z_t^*. Pour illustrer ce fait, nous donnons l'exemple suivant :

Exemple 1.2. Le modèle AR(2) : $Y_t + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} = \varepsilon_t$ peut être écrit sous la formulation espace-état suivante :

$$\begin{pmatrix} Z_{1,t} \\ Z_{2,t} \end{pmatrix} = \begin{pmatrix} -\phi_1 & 1 \\ -\phi_2 & 0 \end{pmatrix} \begin{pmatrix} Z_{1,t-1} \\ Z_{2,t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \varepsilon_t$$
$$Y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} Z_{1,t} \\ Z_{2,t} \end{pmatrix}$$

ou bien :

$$\begin{pmatrix} Z_{1,t}^* \\ Z_{2,t}^* \end{pmatrix} = \begin{pmatrix} -\phi_1 & -\phi_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} Z_{1,t-1}^* \\ Z_{2,t-1}^* \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \varepsilon_t$$
$$Y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} Z_{1,t}^* \\ Z_{2,t}^* \end{pmatrix}$$

Définitions et propriétés

Définition 1.1. *Forme développée*
Le système défini par les équations (1.30a) et (1.30b) peuvent s'écrire sous la forme dite développée. Cette forme est particulièrement utile lorsqu'on s'intéresse à l'estimateur des moindres carrés généralisés du vecteur d'état ou à l'initialisation du filtre de Kalman :

$$Z_{t+1} = \prod_{j=0}^{'} A_{t-j} Z_0 + \sum_{j=1}^{'} \left(A_{t-j} \prod_{k=0}^{j-1} B_{t-k} \right) X_{1,t-j} + \sum_{j=1}^{'} \left(\prod_{k=0}^{-1} A_{t-k} \right) z_{t-j} + \varepsilon_t$$

$$Y_t = C_t \left[\prod_{j=0}^{'} A_{t-j} Z_0 + B_t X_{1,t} + \sum_{j=1}^{'} \left(A_{t-j} \prod_{k=0}^{j-1} B_{t-k} \right) X_{1,t-j} + \sum_{j=1}^{'} \left(\prod_{k=0}^{j-1} A_{t-k} \right) \xi_{t-j} + \varepsilon_t \right] + D_t X_{2,t} + \eta_t$$

Les variables d'état et de mesure s'écrivent donc en fonction de la variable d'état initiale, du passé des erreurs de mesure et des innovations ainsi que des variables exogènes.

Définition 1.2. *Forme canonique*
Le système défini par les équations (1.30a) et (1.30b) est dit sous forme canonique si et seulement si :

$$E(\varepsilon_t \eta_s) = E(\varepsilon_t Z_0) = E(\eta_t Z_0) = 0 \quad \forall t, s = 1, \ldots, T$$

Le modèle espace-état est alors dit causal et inversible.

Définition 1.3. Le modèle espace-état décrit par les équations (1.30a) et (1.30b) est dit stationnaire si les matrices A_t, B_t, C_t et D_t ne dépendent pas de t.

1.1.8.2 Estimation des variables d'état par le filtre de Kalman

Dans cette section, et pour plus de simplicité, nous supposons que B_t et D_t (les variables exogènes) sont nulles.
Les matrices A_t, C_t, R_t, Q_t et p et m sont supposées connues.
L'objectif est d'estimer à chaque instant t les variables cachées conditionnellement aux variables observées jusqu'à l'instant t.

Présentation de l'algorithme

Pour calculer des estimations filtrées du vecteur d'état, l'algorithme, appelé filtre de Kalman [11, 49, 50], est utilisé.
L'algorithme est structuré en deux étapes itératives. Les deux premières équations (1) et (2) sont des équations de mises à jour des mesures (update) et les deux suivantes (3) et (4) de (mise à jour du temps).
La première étape concerne les lois de probabilité a posteriori qui tiennent compte de l'information à l'instant t.
La seconde étape, à la différence de la première, ne dépend pas des observations à l'instant t : le calcul peut être fait sans utiliser Y_t.
Enfin, la dernière équation (5) actualise la matrice de gain K_t qui intervient dans les équations précédentes.
Pour chaque, nous avons les équations suivantes :

$$\begin{cases} (1) : Z^*_{t,t} = Z^*_{t-1,t} + K_t \left(Y_t - C_t Z^*_{t-1,t}\right) \\ (2) : \Sigma_{t,t} = \left(I - K_t C_t\right) \Sigma_{t-1,t} \\ (3) : Z^*_{t,t+1} = A_t Z^*_{t,t} \\ (4) : \sum_{t,t+1} = A_t \Sigma_{t,t} A'_t + Q_t \\ (5) : K_t = \Sigma_{t-1,t} C'_t \left(C_t \Sigma_{t-1,t} C'_t + R_t\right)^{-1} \end{cases}$$

Initialisation : $Z^*_{-1,0} = m, \Sigma_{-1,0} = P$ où P est symétrique semi-définie positive
• $Z^*_{t,t}$ est l'estimation courante du vecteur d'état.
• $\Sigma_{t,t} = V\left(Z_{t,t} - Z^*_{t,t}\right)$ est l'erreur quadratique moyenne sur Z_t.
• $Z^*_{t-1,t}$ est la prévision du vecteur d'état faite à la date $t-1$.
• $\Sigma_{t-1,t} = V\left(Z_{t-1,t} - Z^*_{t-1,t}\right)$ est l'erreur quadratique moyenne de prévision correspondante.
• K_t est la matrice de gain de Kalman.

Description de l'algorithme

L'équation (1) calcule l'estimation actuelle du vecteur d'état $Z^*_{t,t}$ comme la somme pondérée de la prévision à l'instant $t-1$ du vecteur d'état Z_t et de l'erreur de prévision calculée à partir de la dernière valeur observée Y_t.
La pondération K_t, appelée matrice de gain, est actualisée à chaque itération par l'équation (5).
L'équation (3) permet de calculer la prévision de Z_t à l'instant $t+1$.
$Z^*_{t,t+1}$ est la projection de $Z_{t,t+1}$ sur son passé.

Les équations (2) et (4) permettent de calculer la suite des gains de Kalman K_t et ce calcul peut être fait sans utiliser Y_t.
La matrice de covariance a posteriori $\Sigma_{t,t}$ gagne en précision par rapport à la matrice de

covariance a priori $\Sigma_{t-1,t}$ grâce au terme $K_t C_t \Sigma_{t-1,t}$ (équation 2).

La matrice de covariance a priori en $t+1$, noté $\Sigma_{t,t+1}$, prend en compte les erreurs liées aux innovations de l'état avec la matrice Q_t, mais est aussi augmentée d'un terme $A_t \Sigma_{t,t} A_t'$ associé aux erreurs sur l'état à la date t (équation 4).

1.1.8.3 Estimation des paramètres par le maximum de vraisemblance : Algorithme EM

Jusqu'à présent, nous avons supposé que les matrices A_t, C_t, Q_t, R_t, P ainsi que le vecteur m étaient supposés connus. En pratique, ces matrices sont souvent inconnues et doivent être estimées. L'algorithme EM est couramment utilisé pour déterminer les Estimateurs du Maximum de Vraisemblance des paramètres d'un modèle espace-état. Cet algorithme itératif a été introduit par Dempster et al. [41] pour estimer le maximum de vraisemblance de modèles stochastiques à variables cachées.

Pour procéder à une estimation par maximum de vraisemblance des paramètres d'un modèle espace-état, il est nécessaire d'avoir l'expression de la fonction de vraisemblance. Soit θ l'ensemble des paramètres, la log-vraisemblance associée à un échantillon $Y_1, \ldots, Y_T$ d'un modèle espace-état s'exprime à partir des valeurs prévues de l'état $Z^*_{t-1,t}$ et des matrices de covariance associées $\Sigma_{t-1,t}$:

$$\ln l\left(Y_{0:T};\theta\right) = cte - \frac{1}{2}\sum_{t=0}^{'} \ln \det M_{t-1,t}(\theta) - \frac{1}{2}\sum_{t=0}^{'} \tilde{Y}'_{t-1,t}(\theta) M^{-1}_{t-1,t}(\theta) \tilde{Y}_{t-1,t}(\theta)$$

avec : $\tilde{Y}_{t-1,t}(\theta) = Y_t - C_t Z^*_{t-1,t}$ et $M_{t-1,t}(\theta) = C_t \Sigma_{t-1,t} C'_t + R_t$

L'algorithme EM est alors un algorithme itératif qui génère une séquence d'estimations $(\theta_i)_{i=1,2,\ldots}$ à partir d'une condition initiale θ_0. Chaque itération se décompose en deux étapes qui s'écrivent :

Étape E : $\ln l\left(Y_{0:T};\theta\right)$ se déduit de $Z^*_{t-1,t}\left(\theta_i\right)$ et de $\Sigma_{t-1,t}\left(\theta_i\right)$ calculés par un filtre de Kalman.
Étape M : la maximisation de $\ln l\left(Y_{0:T};\theta\right)$ par rapport à θ conduit à Σ_{i+1}

Présentation de l'algorithme

La première étape E (Espérance) calcule une vraisemblance à partir de la formule précédente sur la vraisemblance d'un modèle espace-état. Ces formules mobilisent en particulier l'application d'un filtre de Kalman pour connaitre l'espérance conditionnelle de l'état $Z^*_{t-1,t}$ et de sa covariance $\Sigma_{t-1,t}$ à paramètres θ_i et observations $Y_{0:T}$ fixés.
La seconde étape M (Maximisation), consiste à rechercher un jeu de paramètres maximisant la vraisemblance estimée dans l'étape E.
Cette maximisation peut être analytique ou numérique selon la complexité du problème.
Après un cycle (Étape E/ Étape M), nous obtenons θ_{i+1}.
En iterant les étapes E et M, les paramètres estimés par l'algorithme convergent généralement vers le maximum de vraisemblance.

1.2 Sélection de variables et techniques de régularisation

Dans la section Modèles et modélisation, nous avons vu que le modèle linéaire standard décrit la relation entre l'observation Y et l'ensemble des covariables $X_1, X_2, ..., X_p$. Nous avons également montré que ce modèle est généralement ajusté en utilisant la méthode des moindres carrés (MCO). Dans cette partie, nous examinons certaines approches permettant d'étendre le cadre du modèle linéaire. Nous discutons les moyens par lesquels le modèle linéaire simple peut être amélioré, en remplaçant les moindres carrés simples par des procédures d'ajustement alternatives. La question qui se pose est pourquoi voudrions-nous utiliser une autre procédure d'ajustement au lieu des moindres carrés ? Comme nous le verrons, d'autres procédures d'ajustement peuvent donner une meilleure précision de prédiction et une meilleure interprétabilité du modèle.

Précision de la prédiction : à condition que la relation entre l'observation et les prédicteurs soit approximativement linéaire, les estimations des moindres carrés auront un biais faible. En effet, Si $n \gg p$ c'est-à-dire, n (le nombre d'observations) est bien plus grand que p (le nombre de variables), alors les estimations des moindres carrés ont tendance à avoir une faible variance, et donc fonctionneront bien sur les observations de test. Cependant, si n n'est pas beaucoup plus grand que p, alors il peut y avoir beaucoup de variabilité dans l'ajustement des moindres carrés, ce qui entraîne un sur-ajustement et par conséquent de mauvaises prédictions sur des observations futures.
Si $p > n$, alors l'unicité du coefficient des moindres carrés estimé n'est pas assurée. En effet, la variance est infinie et donc la méthode des moindres carrés ne peut être utilisée. En rajoutant des contraintes ou en réduisant les coefficients estimés (shrinkage) [51], nous pouvons réduire considérablement la variance au prix d'une augmentation négligeable du biais et ainsi améliorer la précision de prédiction du modèle sur des observations futures.

Interprétabilité du modèle : Souvent, certaines variables utilisées dans un modèle de régression multiple ne sont pas associées à l'observation. L'inclusion de ces variables non pertinentes conduit à une complexité inutile dans le modèle résultant. En supprimant ces variables, c'est-à-dire en fixant à zéro les estimations de coefficients correspondantes, nous pouvons obtenir un modèle plus facile à interpréter. Or, il est improbable que les moindres carrés donnent des estimations de coefficients exactement nulles.
Dans cette section, nous étudions quelques approches pour effectuer automatiquement la sélection de variables, c'est-à-dire pour exclure des variables non pertinentes d'un modèle de régression multiple.
De nombreuses approches existent pour ajuster la méthode des moindres carrés à un problème de régression linéaire [52]. Citons à titre d'exemple, la sélection de sous ensemble (Subset selection), la réduction de dimension (Dimension-reduction), et la régularisation. Nous nous intéressons principalement à la régularisation. C'est dans ce sens que nous passons en revue les principales méthodes de régularisation et de sélection de variables existantes dans la littérature. Nous commençons par rappeler les définitions de ces méthodes et les conditions idoines pour leur utilisation. Ensuite, nous décrivons l'algorithme de gradient de descente, connu pour son adaptabilité aux méthodes de régularisation, et nous l'appliquons à différentes méthodes de régularisation. Nous évoquons également, de manière concise et non exhaustive, d'autres méthodes de régularisation figurant dans la littérature.

1.2.1 Méthodes de régularisation

1.2.1.1 La régression ridge

La régression ridge [4] est un terme utilisé pour désigner un modèle de régression linéaire dont les coefficients ne sont pas estimés par les moindres carrés ordinaires (MCO), mais par un estimateur, appelé estimateur Ridge, qui est biaisé mais a une variance plus faible que l'estimateur MCO. Dans certains cas, l'erreur quadratique moyenne de l'estimateur ridge est plus petite que celle de l'estimateur MCO.

L'estimateur ridge

Soient $x_i = (x_{i,1}, ..., x_{i,p})'$ le vecteur contenant les variables explicatives associées à l'individu $i = 1, ..., N$, y_i la réponse associée et $\beta = (\beta_1, \beta_2, ..., \beta_p)'$.

L'estimateur ridge $\hat{\beta}_\lambda$ résout le problème de minimisation suivant :

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\beta}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{(y_i - x_i\beta)^2} + \boldsymbol{\lambda} \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\beta_j^2}. \tag{1.31}$$

avec λ une constante positive. Ainsi, dans l'estimateur ridge, nous ajoutons une pénalité au critère des moindres carrés : nous minimisons la somme des carrés des résidus

$$SSR = \sum_{i=1}^{N} (y_i - x_i\beta)^2$$

plus la norme au carré du vecteur des coefficients

$$\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2$$

En d'autres termes, le problème de la régression ridge pénalise les grands coefficients de régression. En effet, si le paramètre λ est grand, le terme de pénalité est dominant. Nous discuterons ci-dessous comment choisir le paramètre de pénalité λ.

proposition 1.1.

La solution du problème de minimisation (1.31) est donnée par :

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = \boldsymbol{(X'X + \lambda I)^{-1} X'y}. \tag{1.32}$$

avec I la matrice identité $p \times p$.

Démonstration.

La fonction objective à minimiser peut être écrite sous forme matricielle comme suit :

$$\begin{aligned} F &= \sum_{i=1}^{N} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\ &= (y - X\beta)'(y - X\beta) + \lambda\beta'\beta \end{aligned}$$

D'après la condition d'optimalité du premier ordre :

$$\nabla_\beta F = 0$$

Alors :

$$-2X'(y - X\beta) + 2\lambda\beta = 0$$

or

$$(X'X + \lambda I)\,\beta = X'y$$

La matrice $X'X + \lambda I$ est définie positive pour tout $\lambda > 0$. Par conséquent, la matrice est de plein rang et est inversible. En conséquence, la condition du premier ordre est satisfaite par

$$\widehat{\beta}_\lambda = (X'X + \lambda I)^{-1} X'y$$

□

proposition 1.2.

La solution du problème (1.31) est un minimum global.

Démonstration.

Pour vérifier si la solution est un minimum global, nous calculons la matrice Hessienne i.e les dérivées secondes de F :

$$\nabla_{\beta\beta} F = 2\,[X'X + \lambda I]$$

La matrice Hessienne est définie positive (c'est un multiple positif d'une matrice définie positive). Alors, F est strictement convexe en β, ce qui implique que β est un minimum global. □

Nous remarquons que contrairement à l'estimation des moindres carrés, nous n'avons pas besoin de supposer que la matrice de X est de plein rang. En d'autres termes, l'estimateur ridge existe même si X n'est pas de plein rang.

Biais et Variance de l'estimateur Ridge

Nous calculons le biais et la variance de l'estimateur ridge sous l'hypothèse habituellement formulée dans le modèle de régression linéaire :

$$\mathrm{E}[\varepsilon \mid X] = 0 \qquad \mathrm{Var}[\varepsilon \mid X] = \sigma^2 I$$

Avec σ^2 une constante positive et I est la matrice identité $N \times N$. En d'autres termes, nous supposons que les erreurs de la régression ont une espérance nulle et une variance constante σ^2 et ne sont pas corrélées.

Définition 1.4. Biais

Si $\hat{\theta}$ est l'estimateur de θ, $bias(\hat{\theta}) \equiv E[\hat{\theta}] - \theta$

proposition 1.3.

L'espérance conditionnelle de $\widehat{\beta}_\lambda$ est définie par :

$$\mathrm{E}\left[\hat{\beta}_\lambda \mid X\right] = (X'X + \lambda I)^{-1} X'X\beta$$

qui est différente de β sauf si $\lambda = 0$ (cas des moindres carrés).

Le biais de cet estimateur est donné par :

$$\mathrm{E}\left[\widehat{\beta}_\lambda \mid X\right] - \beta = \left[(X'X + \lambda I)^{-1} - (X'X)^{-1}\right] X'X\beta$$

La matrice de covariance de l'estimateur ridge est donnée par :

$$\mathrm{Var}\left[\widehat{\beta}_\lambda \mid X\right] = \sigma^2 (X'X + \lambda I)^{-1} X'X (X'X + \lambda I)^{-1}$$

La variance de l'estimateur ridge est toujours plus petite que la variance de l'estimateur des MCO. En d'autres termes $(\mathrm{Var}[\hat{\beta} \mid X] - \mathrm{Var}\left[\hat{\beta}_\lambda \mid X\right])$ est définie positive. Rappelons que la covariance de deux estimateurs sont comparées en vérifiant si leur différence est définie positive.

Erreur quadratique moyenne

L'erreur quadratique moyenne de l'estimateur ridge est égale à la trace de sa matrice de covariance plus la norme carrée de son biais, appelée la décomposition biais-variance :

$$\mathrm{MSE}\left(\widehat{\beta}_\lambda \mid X\right) = \quad \mathrm{trace}\left(\mathrm{Var}\left[\widehat{\beta}_\lambda \mid X\right]\right) + \left\|\mathrm{bias}\left(\widehat{\beta}_\lambda \mid X\right)\right\|^2$$

Le biais de l'estimateur des moindres carrés est égale à 0. Son erreur quadratique moyenne s'écrit :

$$\mathrm{MSE}(\widehat{\beta} \mid X) = \quad \mathrm{trace}(\mathrm{Var}[\hat{\beta} \mid X])$$

Il existe toujours une valeur du paramètre λ tel que l'estimateur ridge a une erreur quadratique moyenne plus faible que l'estimateur des moindres carrés. Ce résultat est très important d'un point de vue pratique et théorique. Bien que, d'après le théorème de Gauss-Markov, l'estimateur OLS a la variance la plus faible (et la plus faible erreur quadratique moyenne (MSE)) parmi les estimateurs sans biais, il existe un estimateur biaisé (estimateur ridge) dont l'erreur quadratique moyenne est plus faible que celle de l'OLS.

Choix du paramètre de pénalisation λ

Nous avons montré qu'il existe toujours un λ tel que l'estimateur ridge est meilleur que l'estimateur des moindres carrés en terme d'erreur quadratique moyenne.
La question qui se pose est comment choisir le paramètre λ ?
L'une des méthodes les plus célèbres pour choisir λ est la (leave-one-out cross-validation) [53] :
1. Nous choisissons une grille de valeurs possibles $\lambda_1, \ldots, \lambda_P$;
2. Pour $i = 1, \ldots, n$, on exclut la i ème observation (y_i, x_i) de l'échantillon :
a. Nous utilisons $n - 1$ observations pour calculer P estimations ridge de β, notées $\widehat{\beta}_{\lambda_s,i}$
b. Nous calculons P prédictions hors échantillon de l'observation exclue.

$$\widehat{y}_{\lambda_p,i} = x_i \widehat{\beta}_{\lambda_p,i}$$

Pour $p = 1, \ldots, P$
3. Nous calculons l'erreur quadratique moyenne des prédictions :

$$MSE_{\lambda_\rho} = \sum_{i=1}^{N} \left(y_i - \widehat{y}_{\lambda_p,i}\right)^2$$

Pour $p = 1, \ldots, P$
4. Nous choisissons comme paramètre de pénalité optimal λ^* qui génère la plus petite erreur quadratique moyenne.

$$\lambda^* = \arg\min_{\lambda_p} MSE_{\lambda_p}$$

1.2.1.2 Lasso

Dans cette section, nous définissons le "Least Absolute Shrinkage and Selection Operator (lasso)". Le lasso est une méthode de régularisation introduite par Tibshirani [54]. L'estimateur lasso, $\hat{\beta}$, pour le modèle de régression linéaire est donné par :

$$\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\arg\min} \left\{ \frac{\mathbf{1}}{\boldsymbol{n}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_{\mathbf{2}}^{\mathbf{2}} + \boldsymbol{\lambda}\|\boldsymbol{\beta}\|_{\mathbf{1}} \right\}. \tag{1.33}$$

où $\lambda \geq 0$ le paramètre de régularisation. La géométrie de la norme ℓ_1 réduit quelques coefficients exactement à zero et permet une procédure de sélection de variables.

Le problème de sélection de variable

La sélection de variable fait référence au problème de sélection de sous-ensemble, c'est-à-dire, sélectionner un sous-ensemble de prédicteurs optimal pour le modèle. La sélection de sous-ensembles de prédicteurs est un problème majeur, en particulier lorsque le nombre de prédicteur p est grand et que des covariables sont redondantes ou non pertinentes. Dans le cadre de la régression linéaire, le problème de sélection de sous-ensemble s'exprime par le modèle de la forme :

$$Y = X_S \beta_S + \epsilon$$

Avec $S \subset \{1, \ldots, p\}$ est l'ensemble actif (active set), X_S sont les colonnes et β_S est le vecteur des coefficients de régression correspondant au sous-ensemble S. Puisque S n'est pas connu, il y a toujours une incertitude sur ce sous-ensemble (2^p sous-ensembles) à utiliser. Certaines méthodes standard de sélection de sous-ensembles sont la sélection vers l'avant (forward-selection), l'élimination vers l'arrière (backward selection) et la combination de ces deux méthodes (c'est-à-dire, forward selection steps suivie par backward elimination steps). Une littérature abondante existe sur les méthodes de sélection de variables pour les modèles linéaires [55] et pour les problèmes de grande dimension [56].

Nous considérons le modèle de régression linéaire pour les cas de grande dimension, où le nombre de paramètres inconnus à estimer est beaucoup plus élevé que le nombre d'observations. Pour $p > n$, le modèle de régression linéaire, est un problème mal posé (un problème qui peut avoir plus d'une solution). Afin de résoudre ce problème, nous devons introduire des contraintes ou des régularisations dans le processus d'estimation. Comme mentionné précédemment, le lasso est une méthode de régression régularisée. Il estime les coefficients de régression en résolvant le problème des moindres carrés avec les contraintes suivantes :

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\textbf{minimize}} \left\{ \frac{\mathbf{1}}{\boldsymbol{n}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_{\mathbf{2}}^{\mathbf{2}} \right\} \quad \text{subject to } \|\boldsymbol{\beta}\|_{\mathbf{1}} \leq \boldsymbol{t}. \tag{1.34}$$

Si t est égale ou supérieur à la norme ℓ_1 de l'estimateur des moindres carrés, alors l'estimateur lasso est équivalent à l'estimateur des moindres carrés. Si t est inférieur à la norme ℓ_1 de l'estimateur des moindres carrés, alors le lasso réduit les coefficients de régression

estimés vers ou exactement à zero.
Il est possible de montrer qu'il existe une correspondance entre t et λ. En d'autres termes, pour $\lambda > 0$ il existe un t tel que les deux problèmes (1.33) et (1.34) ont la même solution [57].
La forme Lagrangienne du problem du lasso est définie par :

$$\underset{\beta_1 \in \mathbb{R}}{\text{minimize}} \left\{ \frac{1}{n} \|Y - X\beta_1\|_2^2 + \lambda |\beta_1| \right\}$$

En général, le lasso n'a pas une solution explicite car la fonction objective n'est pas différentiable. Cependant, il est possible d'obtenir des solutions pour le cas particulier d'une matrice orthonormée. Différentes interprétations pour différents scénarios de la solution lasso sont discutées comme suit.

L'estimateur lasso : cas d'un modèle de régression simple

Nous commençons par illustrer la solution du lasso pour un modèle de régression simple, où $p = 1$ et $Y = X_1\beta_1 + \epsilon$.
Considérons le problème de minimisation suivant :

$$\underset{\beta_1 \in \mathbb{R}}{\text{minimize}} \left\{ \frac{1}{n} \|Y - X_1\beta_1\|_2^2 + \lambda |\beta_1| \right\}$$

Supposons que $\hat{\beta}_1$ est solution du problème de minimisation. D'après les conditions de KKT :

$$-\tfrac{2}{n} X_1' \left(Y - X_1\hat{\beta}_1\right) + \lambda \operatorname{sign}\left(\hat{\beta}_1\right) = 0$$
$$\tfrac{1}{n} X_1' \left(Y - X_1\hat{\beta}_1\right) = \tfrac{\lambda}{2} \operatorname{sign}\left(\hat{\beta}_1\right)$$

Notons que $\frac{1}{n} X_1' X_1 = 1$, car nous supposons que les prédicteurs sont standardisés.

$$\hat{\beta}_1 = \tfrac{1}{n} X_1' Y - \tfrac{\lambda}{2} \operatorname{sign}\left(\hat{\beta}_1\right)$$
$$\hat{\beta}_1 = \begin{cases} \frac{1}{n} X_1' Y + \frac{\lambda}{2} & \text{si } \frac{1}{n} X_1' Y < -\frac{\lambda}{2} \\ 0 & \text{si } \frac{1}{n} |X_1' Y| \leq \frac{\lambda}{2} \\ \frac{1}{n} X_1' Y - \frac{\lambda}{2} & \text{si } \frac{1}{n} X_1' Y > \frac{\lambda}{2} \end{cases}$$

qui est équivalent au terme $\frac{x_1' Y}{n}$ limité par $\frac{\lambda}{2}$

$$\hat{\beta}_1 = \mathbb{S}_{\frac{\lambda}{2}} \left(\frac{X_1' Y}{n} \right)$$

Par conséquent, l'estimateur lasso pour le cas de la variable unique peut également être calculé en limitant légèrement l'estimateur OLS par $\frac{\lambda}{2}$

$$\hat{\beta}_1 = \mathbb{S}_{\frac{\lambda}{2}} \left(\hat{\beta}_{OLS} \right)$$

avec $\hat{\beta}_{OLS} = \frac{X_1' Y}{n}$

L'estimateur lasso : cas orthonormé

Nous dérivons l'estimateur lasso pour le cas orthonormé. Nous supposons alors que les variables ne sont pas corrélées, ce qui implique $X_i' X_j = 0 \; \forall \; i \neq j$ et $\frac{1}{n} X' X = I_p$.
Supposons que $\hat{\beta}$ est une solution du lasso, la condition de stationnarité KKT nous donne :

$$-\tfrac{2}{n}X'(Y - X\hat{\beta}) + \lambda \operatorname{sign}(\hat{\beta}) = 0$$
$$\tfrac{1}{n}X'(Y - X\hat{\beta}) = \tfrac{\lambda}{2}\operatorname{sign}(\hat{\beta})$$

où $\frac{1}{n}X'X = I_p$. Il s'ensuit que,

$$\hat{\beta} = \frac{1}{n}X'Y - \frac{\lambda}{2}\operatorname{sign}(\hat{\beta})$$
$$\hat{\beta}_j = \begin{cases} \frac{1}{n}\left(X'Y\right)_j + \frac{\lambda}{2} & \text{si } \frac{1}{n}\left(X'Y\right)_j < -\frac{\lambda}{2} \\ 0 & \text{si } \frac{1}{n}\left|\left(X_1'Y\right)_j\right| \leq \frac{\lambda}{2} \\ \frac{1}{n}\left(X'Y\right)_j - \frac{\lambda}{2} & \text{si } \frac{1}{n}\left(X'Y\right)_j > \frac{\lambda}{2} \end{cases}$$

Le coefficient $\hat{\beta}_j$ est calculé en limitant la j^{me} ligne de $\left(\hat{\beta}_{OLS}\right)_j = \left(\frac{1}{n}X'Y\right)_j$, par $\frac{\lambda}{2}$.
Nous avons montré qu'en général, l'estimateur lasso n'a pas de solution explicite.

L'estimateur lasso : cas de plusieurs prédicteurs

Nous essayons de trouver une solution pour une composante et nous montrons qu'elle dépend de toutes les autres composantes.
Nous supposons que X est de plein rang, alors $X'X$ est inversible. X_{-j} désigne toutes les colonnes sauf la j^{me} colonne, et β_{-j} désigne le vecteur de paramètres sauf β_j. Supposons que $\hat{\beta}_j$ est une solution de la j^{me} composante β_j, alors, d'après la condition de stationnarité (KKT) nous avons :

$$-\frac{2}{n}X_j'\left(Y - X_{-j}\beta_{-j} - \hat{\beta}_j X_j\right) + \lambda \operatorname{sign}\left(\hat{\beta}_j\right) = 0$$

Une simplification supplémentaire donne :

$$-\frac{2}{n}X_j'Y + \frac{2}{n}X_j'X_{-j}\beta_{-j} + \hat{2}\beta_j\frac{X_j'X_j}{n} + \lambda \operatorname{sign}\left(\hat{\beta}_j\right) = 0$$

Comme $\frac{\mathbf{x}_j'\mathbf{x}_j}{n} = 1$, nous avons

$$\hat{\beta}_j = \frac{1}{n}X_j'\left(Y - X_{-j}\beta_{-j}\right) - \frac{\lambda}{2}\operatorname{sign}(\hat{\beta})$$

Nous remarquons que la solution d'un β_j dépend de toutes les autres composantes $\beta_{i\neq j}$, donc une solution explicite n'existe pas. Pour le cas orthonormé, le terme $X_j'X_{-j}$ s'annule en raison de l'orthogonalité et nous obtenons donc une forme explicite. En général, le lasso n'a pas de solution explicite, par contre il peut être résolu efficacement grâce à sa forme d'optimisation convexe [58, 59, 60].

Interprétation Bayesienne de la méthode du lasso

Les valeurs estimées des coefficients de régression par lasso peuvent également être interprétées comme une estimation du maximum a posteriori (MAP) ; c'est-à-dire que si nous supposons la loi double exponentielle (Laplace) a priori sur les coefficients de régression, les estimations MAP bayésiennes sont les mêmes que les estimations lasso.
Ici, nous considérons un modèle bayésien hiérarchique $Y \sim N_n\left(X\beta, \sigma^2 I_n\right)$ et $\beta_i \sim$ Double Exp $\left(\lambda\left|\beta_i\right|\right)$. Sans perte de généralité, nous pouvons supposer $\sigma^2 = 1$. Il est décrit comme suit (pour plus de détailles voir [61]) :

$$Y \mid X, \beta \sim N_n\left(X\beta, I_n\right)$$
$$\beta_1, \beta_2, \ldots, \beta_p \mid \lambda \overset{iid}{\sim} \frac{\lambda}{2}\exp\left(-\lambda\left|\beta_i\right|\right)$$
$$P(\beta \mid X, Y, \lambda) \propto P(Y \mid X, \beta)P(\beta \mid \lambda)$$

proposition 1.4.

L'estimateur $\hat{\beta}_{MAP}$ est égale à l'estimateur lasso, estimé à $2\lambda/n$.

Démonstration.

Considérons l'estimation MAP de β sous ce modèle.

$$\begin{aligned}
\hat{\beta}_{MAP} &= \arg\max_{\beta\in\mathbb{R}^p}\{\log P(\beta \mid X, Y, \lambda)\} \\
&= \arg\max_{\beta\in\mathbb{R}^p}\{\log[P(Y \mid X, \beta)P(\beta \mid \lambda)]\} \\
&= \arg\max_{\beta\in\mathbb{R}^p}\left\{\log\left[P(Y \mid X, \beta)\prod_{i=1}^{p} P\left(\beta_i \mid \lambda\right)\right]\right\} \\
&= \arg\max_{\beta\in\mathbb{R}^p}\left\{\log P(Y \mid X, \beta) + \sum_{j=1}^{p}\log P\left(\beta_j \mid \lambda\right)\right\}
\end{aligned}$$

Nous avons :

$$P(Y \mid X, \beta) = \frac{1}{(2\pi)^{(n/2)}}\exp\left(-\frac{1}{2}\|Y - X\beta\|_2^2\right)$$

En appliquant le log et en ignorant les termes constants, nous obtenons :

$$\arg\max_{\beta\in\mathbb{R}^p}\{\log P(Y \mid X, \beta)\} = \arg\max_{\beta\in\mathbb{R}^p}\left\{-\frac{1}{2}\|Y - X\beta\|_2^2\right\}$$

Nous pouvons simplifier le deuxième terme :

$$\arg\max_{\beta\in\mathbb{R}^p}\left\{\sum_{j=1}^{p}\log P\left(\beta_j \mid \lambda\right)\right\} = -\lambda\sum_{j=1}^{p}|\beta_j| = -\lambda\|\beta\|_1$$

Alors :

$$\begin{aligned}
\hat{\beta}_{MAP} &= \arg\max_{\beta\in\mathbb{R}^p}\left\{-\frac{1}{2}\|Y - X\beta\|_2^2 - \lambda\|\beta\|_1\right\} \\
&= \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\} \\
&= \hat{\beta}\left(\frac{2\lambda}{n}\right)
\end{aligned}$$

$\hat{\beta}_{MAP}$ est l'estimateur lasso, estimé à $2\lambda/n$. □

Pour λ fixé, le lasso est un problème d'optimisation quadratique en β.
Pour tout λ, nous obtenons une solution du problème. C'est dans ce sens que nous avons besoin d'une méthode efficace pour choisir λ, telle que la validation croisée, le *bootstrapping*, etc. Si $\lambda = 0$ la solution du lasso est la solution des moindres carrés ordinaire. Si λ croit, le nombre des composantes non nulles $\hat{\beta}$ décroit. Si $\lambda = \infty$, le lasso donne le modèle nul où $\hat{\beta} = 0$.
En règle générale, nous choisissons la valeur de λ qui minimise l'erreur de prédiction [62].

1.2.1.3 La régression Elastic-net

Zou et Hastie [5] ont remarqué que la méthode du lasso est limitée. En effet, le nombre de variables sélectionnées par le lasso est limité par la taille de l'échantillon n qui ne peut être dépassée. De plus, le lasso a tendance à ne sélectionner qu'une seule variable d'un groupe de variables très corrélées. Et finalement, le lasso n'est pas très performant en terme de prédiction par comparaison avec la méthode ridge et ce, quand une forte corrélation existe entre les variables explicatives.
Pour surpasser ces trois problèmes, Zou et Hastie [5] ont proposé l'elastic-net. Cette méthode combine deux pénalités, l_1 et l_2.
L'estimateur elastic-net est donné par :

$$\hat{\boldsymbol{\beta}}_{\text{Enet}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right)$$

avec λ et $\alpha \in [0,1]$ deux paramètres de régularisation. Un bon choix de ces deux paramètres permettra de bénéficier des point forts du lasso et du Ridge et minimiser leurs inconvénients. Le Ridge ($\alpha = 0$) et lasso ($\alpha = 1$) sont deux cas particuliers de l'estimateur elastic-net.

1.2.1.4 La régression Weighted fusion

Inspiré par la méthode Elastic-net, Daye and Jeng [63] ont proposé la méthode Weighted fusion :

$$\hat{\boldsymbol{\beta}}_{\text{W}-\text{fusion}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \frac{\mu}{p} \sum_{j=1}^{p-1} \sum_{j>i} \omega_{ij} \left(\beta_i - s_{ij}\beta_j \right)^2$$

avec $\omega_{ij} = \frac{|\rho_{ij}|^\gamma}{1-|\rho_{ij}|}, \rho_{ij} = x_i' x_j$ est la corrélation empirique entre les variables x_i et x_j $s_{ij} = \text{sign}(\rho_{ij})$ le signe de ρ_{ij} et $\lambda \geq 0, \mu \geq 0, \gamma > 0$ sont des paramètres de régularisation. Le critère s'écrit sous la forme suivante :

$$\hat{\boldsymbol{\beta}}_{\text{W}-\text{fusion}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \mu \beta' Q \beta$$

où

$$p \times Q = \begin{pmatrix} \sum_{k \neq 1} w_{1k} & -s_{12}w_{12} & \cdots & -s_{1p}w_{1p} \\ -s_{12}w_{12} & \sum_{k \neq 2} w_{2k} & \cdots & \vdots \\ \vdots & \vdots & \ddots & -s_{p-1,p}w_{p-1,p} \\ -s_{1p}w_{1p} & \cdots & -s_{p-1,p}w_{p-1,p} & \sum_{k \neq p} w_{pk} \end{pmatrix}$$

Notons que Q est symétrique et semi-définie positive.

1.2.1.5 Adaptive-lasso

L'adaptive lasso a été proposé par Zou [3]. C'est une version pondérée du lasso. L'estimateur adptive-lasso est donné par :

$$\hat{\boldsymbol{\beta}}_{\text{Ad-lasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|$$

où $\hat{\omega} = (\hat{\omega}_1, \ldots, \hat{\omega}_p)'$ est un vecteur de poids donnés par un estimateur initial, et λ est un paramètre de régularisation. Cette méthode a été proposée pour résoudre le problème de la non-consistance de la méthode lasso, en pénalisant différemment les coefficients de $\boldsymbol{\beta}$. Notons que la consistance de l'adaptive lasso n'implique pas qu'elle est toujours plus performante que la méthode lasso en terme de prédiction.

1.2.1.6 La régression Fused lasso

Fused lasso a été proposé par Tibshirani et al. [64]. Cette méthode est performante dans le cas où une structure d'ordre entre les variables existe. L'estimateur fused-lasso est donné par :

$$\hat{\boldsymbol{\beta}}_{\text{Fused-lasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \mu \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$$

La première pénalité encourage la parcimonie (sparsity) des coefficients, tandis que la deuxième pénalité encourage la parcimonie de leurs différences. Notons que lorsque le nombre de variables p est grand, ce problème est difficile à résoudre.

1.2.1.7 La régression Smooth lasso

Le Smooth lasso a été développé par Hebiri and van De Geer [65]. Cette méthode est performante dans le cas où les coefficients de régression successifs varient lentement ou bien lorsque les variables sont ordonnées. L'estimateur Smooth lasso est donné par :

$$\hat{\boldsymbol{\beta}}_{\text{Sm-lasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \mu \sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2$$

De la même manière que le Weighted fusion, ce critère s'écrit sous la forme :

$$\hat{\boldsymbol{\beta}}_{\text{Sm-lasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \mu \beta' R' R \beta$$

où

$$R = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \ddots & \vdots \\ 0 & 1 & -1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}$$

1.2.2 Chemins de régularisation

1.2.2.1 Algorithme "*Coordinate Descent*"

Dans cette section, nous nous intéressons particulièrement à l'algorithme "*Coordinate Descent*"[66]. C'est un algorithme qui permet de retrouver efficacement les chemins de régularisation et a été appliqué à plusieurs méthodes de régularisation. Cet algorithme est très compétitif à l'algorithme LARS [58] et a fait récemment l'objet de plusieurs travaux. L'algorithme "*Coordinate Descent*" consiste à optimiser chaque paramètre séparément, tout en fixant les autres, et répéter la procédure jusqu'à ce que la convergence soit assurée. En effet, considérons une fonction objective, f qui est convexe mais pas nécessairement

différentiable. Alors cette fonction peut être divisée en une fonction différentiable et non différentiable : $f = f_d + f_c$, où f_d est convexe et différentiable, et f_c est convexe mais non différentiable. Lorsque la fonction objective f est différentiable ou bien sa fonction non différentiable est séparable i.e ($f_c(x) = \sum_{i=1}^{p} f_{c_i}(x_i)$ où chaque f_{c_i} est convexe), la fonction objectif f peut être minimisée par coordonnées.

1.2.2.2 Algorithme "*Coordinate Descent*" pour le lasso

La fonction objective du lasso peut être scindée en deux parties, une partie différentiable $f_d = \|Y - X\beta\|_2^2$ et non différentiable $f_c = \sum_{j=1}^{p} |\beta_j|$. $f_c = \sum_{j=1}^{p} |\beta_j|$ est strictement convexe en chaque coordonnée et par conséquent, nous pouvons minimiser ces coordonnées. Dans le cas d'un seul prédicteur, la solution du lasso est explicite, et est un seuil doux (*soft threshold*) de l'estimation des moindres carrés. Nous exploitons cette propriété pour implémenter l'algorithme de descente pour le lasso.
L'algorithme de "*Coordinate Descent*" est un algorithme itératif qui résout exactement pour une variable, le problème du lasso, tout en fixant les variables restantes [67].
Nous fixons tous les paramètres de β à l'exception de la j ème composante β_j.
Soit X_j la j ème colonne de X et X_{-j} toutes les colonnes à l'exception de la j ème colonne, alors le problème s'écrit :

$$\arg\min_{\beta_j \in \mathbb{R}} \{ \frac{1}{n} ||Y - X_{-j}\beta_{-j} - \beta_j X_j||_2^2 + \lambda\beta_j + \lambda \sum_{l \neq j} ||\beta_l|| \}$$

Soit $r_j := Y - X_{-j}\beta_{-j}$, le résidu partiel (qui est la différence entre l'observation Y et la partie du modèle ajusté qui n'implique pas la variable X_j). Alors le problème peut être vu comme un problème du lasso univarié où r_j joue le rôle de l'observation.

$$\arg\min_{\beta_j \in \mathbb{R}} \{ \frac{1}{n} ||r_j - \beta_j X_j||_2^2 + \lambda\beta_j + \lambda \sum_{l \neq j} ||\beta_l|| \}$$

Supposons que $\hat{\beta}_j$ est la solution du problème, d'après la condition de stationnarité KKT, nous avons :

$$-\tfrac{2}{n} X_j' \left(r_j - \hat{\beta}_j X_j \right) + \lambda \operatorname{sign}\left(\hat{\beta}_j \right) = 0$$
$$\tfrac{1}{n} r_j' X_j - \hat{\beta}_j = \tfrac{\lambda}{2} \operatorname{sign}(\hat{\beta})$$

L'estimateur des moindres carrés de la j^{me} variable est donné par $\left(\hat{\beta}_{OLS} \right)_j = \frac{1}{n} r' X_j$. Par conséquent, une solution du lasso univariée peut être calculée en appliquant un seuil doux de l'estimateur OLS :

$$\hat{\beta}_j = \mathbb{S}_{\frac{\lambda}{2}} \left(\left(\hat{\beta}_{OLS} \right)_j \right)$$

Algorithm 1 Algorithme "*Coordinate Descent* " pour le lasso

Input : dataset (Y, X)
Output : $\hat{\beta}$:= lasso estimated vector of regression coefficients
Initialize $\beta = 0$
repeat
  for each $j \in 1, ..., p$ **do**
    Compute the partial residual r_j, where

$$r_j = Y - \sum_{l \neq j} X^l \beta_l$$

    Compute OLS coefficient for single predictor

$$\left(\hat{\beta}_{OLS}\right)_j = \frac{1}{n} r_j' X_j$$

    Update β_j (lasso solution : single variable case)

$$\beta_j = \mathbb{S}_{\frac{\lambda}{2}} \left(\left(\hat{\beta}_{OLS}\right)_j \right)$$

  end for
until convergence ;
$\hat{\beta} = \beta$
return $\hat{\beta}$

1.2.2.3 Algorithme "*Coordinate Descent*" pour l'adaptive lasso

Zou [3] a montré que l'adaptive lasso peut être vu comme un problème lasso, en transformant les données originales. Donc, il est possible d'utiliser l'algorithme *Coordinate Descent* pour retrouver $\hat{\boldsymbol{\beta}}_{\text{Adalasso}}$. L'algorithme pour retrouver $\hat{\boldsymbol{\beta}}_{\text{Adalasso}}$ est le suivant :
1. Construire la nouvelle matrice des variables explicatives $X^{**} = \left(x_1^{**} | \ldots | x_p^{**}\right)$, avec $x_j^{**} = x_j / \hat{\omega}_j$ pour $j = 1, \ldots, p$
2. Pour tout λ, résoudre le problème lasso suivant en utilisant l'Algorithme 1 :

$$\hat{\boldsymbol{\beta}}^{**} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}^{**} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

3. L'estimateur $\hat{\beta}_{\text{Adalasso}}$ est donnée par : $\hat{\beta}_{\text{Adalasso}} = \left(\hat{\beta}_1^{**}/\hat{\omega}_1, \ldots, \hat{\beta}_p^{**}/\hat{\omega}_p\right)'$

1.2.2.4 Algorithme "*Coordinate Descent*" pour Elastic-net

Cet algorithme a été proposé par van der Kooij [68].
Pour l'estimateur Elastic-net, nous remplaçons l'étape de la mise à jour (update) par :

$$\boldsymbol{\beta}_j \leftarrow \frac{S\left(\frac{1}{n} r_j' X_j, \lambda\alpha\right)}{1 + \lambda(1-\alpha)} = \frac{S\left(\left(\hat{\beta}_{OLS}\right)_j, \lambda\alpha\right)}{1 + \lambda(1-\alpha)}$$

1.2.3 Aperçu des méthodes de régularisation existantes dans la littérature

Méthodes de régularisation	Définition	Paramètres
Ridge	$\arg\min_{\beta} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$	$\lambda \geq 0$
lasso	$\arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\lvert\beta_j\rvert$	$\lambda \geq 0$
Elastic-net	$\arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+$ $\lambda\left(\alpha\sum_{j=1}^{p}\lvert\beta_j\rvert+\frac{1}{2}(1-\alpha)\sum_{j=1}^{p}\beta_j^2\right)$	$\lambda \geq 0$, $\alpha \in [0,1]$
Adaptive-lasso	$\arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\hat{\omega}_j\lvert\beta_j\rvert$	$\lambda \geq 0$, $\hat{\omega}$ un vecteur de poids donnés par un estimateur initial
Weighted-fusion	$\arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\lvert\beta_j\rvert+$ $\frac{\mu}{p}\sum_{j=1}^{p-1}\sum_{j>i}\omega_{ij}\left(\beta_i-s_{ij}\beta_j\right)^2$	$\omega_{ij}=\frac{\lvert\rho_{ij}\rvert^{\gamma}}{1-\lvert\rho_{ij}\rvert}$, $\rho_{ij}=x_i'x_j$, $s_{ij}=\operatorname{sign}(\rho_{ij})$, $\lambda\geq 0,\mu\geq 0,\gamma>0$.
Fused lasso	$\arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\lvert\beta_j\rvert+$ $\mu\sum_{j=2}^{p}\lvert\beta_j-\beta_{j-1}\rvert$	$\mu \geq 0$, $\lambda \geq 0$
Smooth lasso	$\arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\lvert\beta_j\rvert+$ $\mu\sum_{j=2}^{p}\left(\beta_j-\beta_{j-1}\right)^2$	$\mu \geq 0$, $\lambda \geq 0$
Grouped lasso	$\arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+$ $\lambda\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}\beta_j^2}$	$\lambda \geq 0$, L désigne le nombre de groupes et G_l l'ensemble des indices des variables contenues dans le groupe l avec $l=1,\ldots,L$
Oscar	$\min_{\beta\in\mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2$ s.t $\sum_{j=1}^{p}\lvert\beta_j\rvert+$ $c\sum_{j<k}\max\left\{\lvert\beta_j\rvert,\lvert\beta_k\rvert\right\}\leq\lambda$	$c \geq 0$, $\lambda > 0$
Bridge	$\arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\lvert\beta_j\rvert^{\gamma}$	$\gamma \geq 1$ et $\lambda \geq 0$
Garrote	$\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}c_j\hat{\boldsymbol{\beta}}_j^{ols}x_{ij}\right)^2$ s.t $c_j\geq 0,\sum_{j=1}^{p}c_j\leq t$	$\hat{\beta}^{\text{ols}}$ l'estimateur des moindres carrés ordinaire

La régression grouped lasso a été proposé par Yuan and Lin [69] pour analyser le cas où les variables explicatives forment des clusters ou groupes.
La régression Bridge a été proposée par Frank and Friedman [70] et réétudié par Fu [71]. les régressions Ridge et lasso sont deux cas particuliers de cette méthode pour $\gamma = 2$ et

$\gamma = 1$ respectivement.
La régression Oscar, proposé par Bondell et Reich [72], permet la sélection de variables tout en les regroupant en clusters.
La méthode Garotte, proposé par Breiman [73], rétrécit l'estimateur des moindres carrés par des coefficients positifs dont la somme est contrainte. Cependant, elle n'est pas valable pour $p > n$ puisque l'estimateur des moindres carrés, utilisé dans cette méthode, n'est pas unique.

Les différents modèles et techniques de régularisation que nous avons abordés jusqu'à présent dans les sections Modèles et modélisation et Sélection de variables et techniques de régularisation, sont couramment utilisés pour construire des modèles de prédiction, en l'occurence les modèles de prédiction clinique. Dans ce qui suit, nous présentons les différentes démarches de construction d'un modèle de prédiction clinique. L'accent sera mis sur la prédiction de risque chez les patients victimes d'accident vasculaire cérébral.

1.3 Construction et conceptualisation d'un modèle de prédiction clinique avec application à l'accident vasculaire cérébral

Le modèle de prédiction clinique est un outil quantitatif d'évaluation des risques. Il fournit des informations objectives et précises pour la prise de décision des médecins, des patients et du personnel de santé.
Par exemple, prédire qu'un patient atteint d'une tumeur maligne résistera à un certain médicament de chimiothérapie, permettra d'éviter de lui donner ce médicament.
L'utilisation de modèle de prédiction clinique devient très fréquente. La valeur des données n'a jamais été aussi importante. Le développement rapide de l'acquisition de données, du stockage et de la technologie de prédiction à l'ère du big data a rendu possible la vision d'un traitement médical personnalisé.
Dans cette partie, nous présentons les différentes méthodologies de construction de modèles de prédiction clinique. Nous définissons le concept, les méthodes et processus de construction de ces modèles et nous les classifions. Nous discutons les conditions nécessaires à la conduite de cette méthodologie de recherche ainsi que ses limitations.
Comme application, nous avons choisi l'Accident Vasculaire Cérébral (AVC), un sujet d'importance significative pour la santé publique. Nous allons définir l'AVC et ces conséquences tout en mettant l'accent sur la prédiction de la déficience cognitive post AVC.
Finalement, nous allons passer en revue les travaux menés sur la modélisation prédictive du risque post AVC, en particulier la méthodologie des courbes de récupération.

1.3.1 Concept d'un modèle de prédiction clinique

Le modèle de prédiction clinique fait référence à l'utilisation d'un modèle mathématique (paramétrique / semi-paramétrique / non paramétrique) pour estimer la probabilité qu'un sujet soit malade ou bien exposé à une maladie dans le futur [74]. Les méthodes couramment utilisées comprennent le modèle de régression linéaire multiple, le modèle de régression logistique et le modèle de régression de Cox. L'évaluation et la vérification de l'efficacité des modèles de prédiction sont la clé de l'analyse statistique et de la modélisation des données [75].
D'un point de vue statistique, ces modèles peuvent être construits tant qu'une observation issue d'un problème clinique (Y) peut être quantifiée par une caractéristique (X).
Sur la base des problèmes cliniques, les modèles de prédiction comprennent les modèles diagnostic, pronostic et d'occurrence de maladies [74].
Le modèle diagnostic est courant dans les études transversales, se concentrant sur les symptômes cliniques, les caractéristiques des sujets de l'étude, et la probabilité de diagnostiquer une certaine maladie.
Le modèle pronostic se concentre sur la probabilité de résultats tels que la récidive, la mort, l'invalidité et les complications au cours de la période d'une maladie particulière. Ce modèle est courant dans les études de cohorte.
Le modèle d'occurrence de maladies prédit si une maladie particulière se produira dans l'avenir en fonction des caractéristiques générales du sujet, ce qui est également courant dans les études de cohorte.
De nombreuses similitudes existent entre le modèle diagnostic, le modèle pronostic et le modèle d'occurrence de maladie.
D'un point de vue technique, les chercheurs seront confrontés à la sélection des prédicteurs, à la mise en place de stratégies de modélisation, ainsi qu'à l'évaluation et à la

vérification des performances du modèle.

1.3.1.1 Méthodes et processus de construction de modèles de prédiction clinique

La construction d'un modèle de prédiction clinique se base en général sur six grandes étapes [76] :

1) Sélectionner un ensemble de données de prédicteurs comme facteurs potentiels.
2) Choisir un modèle statistique approprié pour analyser la relation entre les prédicteurs et l'observation.
3) Sélectionner des variables parmi les prédicteurs existants qui sont suffisamment significatives pour être incluses dans le modèle.
4) Construire le modèle.
5) Évaluer le modèle.
6) Expliquer les applications du modèle dans la pratique clinique.

Une bonne conception de l'étude et un bon protocole de mise en oeuvre sont nécessaires. À l'heure actuelle, il n'existe pas de modèle de prédiction pour un problème clinique spécifique.
Un ensemble d'apprentissage et de validation sont nécessaires pour la construction et la vérification de la capacité et la performance de prediction du modèle. Concernant ces deux ensembles de données, ils peuvent être collectées de manière prospective ou rétrospective. Les ensembles de données collectées prospectivement sont en général de meilleure qualité. La taille de l'échantillon est d'une très grande importance et doit être aussi large que possible.
Le contrôle de la qualité et la gestion de la collecte des données doivent également être effectués. Si les données sont collectées rétrospectivement, leur qualité doit être évaluée, les valeurs aberrantes identifiées et les valeurs manquantes correctement traitées.
Enfin, le jeu de données d'apprentissage pour la modélisation et la validation sont déterminés en fonction des situations réelles. Parfois, nous ne pouvons modéliser et valider que dans le même ensemble de données. Cette approche est autorisée, mais l'applicabilité externe du modèle sera bien évidement affectée.

1.3.1.2 Établissement, évaluation et validation de modèles de prédiction clinique

Avant d'établir un modèle de prédiction, il est nécessaire de définir les prédicteurs rapportés dans la littérature, de déterminer les principes et les méthodes de sélection de ces prédicteurs et de choisir le type du modèle mathématique.
Habituellement, un modèle paramétrique, semi-paramétrique ou des algorithmes d'apprentissage automatique sont utilisés pour construire les modèles. Il est nécessaire de déterminer à l'avance la forme du modèle de prédiction. Les chercheurs peuvent faire des choix en fonction de la nature de l'étude. Une fois le modèle construit, comment peut-on l'évaluer ?
L'évaluation et la vérification du modèle relèvent d'une analyse statistique avancée. A titre d'exemples, la discrimination, l'étalonnage, l'efficacité clinique ainsi que d'autres indicateurs de performances.
L'effet du modèle de prédiction est susceptible de changer à mesure que la population change. Par conséquent, une étude complète du modèle de prédiction devrait inclure sa validation. La validation comprend la validité interne et la validité externe du modèle.

La validité interne reflète la reproductibilité du modèle, qui peut être confirmée par la validation croisée ou bien le bootstrap, et ce avec les mêmes données de l'étude. La validité externe reflète la généralisabilité du modèle et doit être validée avec des ensembles de données qui sont temporellement et géographiquement indépendants. La validation interne et externe du modèle sont des étapes nécessaires pour évaluer sa stabilité et son applicabilité.
En effet, pour vérifier la validité interne, plusieurs méthodes sont disponibles :

1) Divisez au hasard les données en deux parties (*data splitting*), construction et validation. Les études avec des échantillons de petite taille ne conviennent pas à cette méthode.
2) La méthode de validation croisée par dix (*ten-fold-cross validation*), consiste à diviser les données en dix parties et à utiliser neuf parties pour construire le modèle et la partie restante pour le vérifier. Cette méthodologie permet de construire un modèle relativement stable.
3) La méthode classique d'analyse de validité interne (*Bootstrap*) consiste à échantillonner au hasard un certain nombre d'observations dans l'ensemble de données d'origine pour créer un modèle, puis à utiliser l'ensemble de données d'origine pour le vérifier. En effectuant un échantillonnage aléatoire (500 à 1000 fois), 500 à 1000 modèles sont obtenus et les distributions de paramètres peuvent donc être résumées. Par conséquent, les valeurs finales des paramètres sont déterminées. Il est prouvé que les modèles obtenus par cette méthode ont une stabilité plus élevée que les méthodes : (*data splitting*) et (*ten-fold-cross validation*).
Nous pouvons évaluer l'efficacité clinique en calculant la sensibilité et la spécificité du modèle. Nous évaluons généralement si les patients peuvent être classés comme un bon ou mauvais pronostic selon une certaine valeur seuil. L'analyse de la courbe de décision est également une méthode couramment utilisée pour prédire l'efficacité clinique des modèles. En statistique, les notions de sensibilité et spécificité sont d'une importance majeure en épidémiologie et en théorie de la détection du signal.
- La sensibilité (ou sélectivité) d'un test mesure sa capacité à donner un résultat positif lorsqu'une hypothèse est vérifiée.
- La spécificité, qui s'oppose à la sensibilité, mesure la capacité d'un test à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée.
- L'analyse de la courbe de décision (DCA) [77, 78] est une méthode permettant d'évaluer la valeur ajoutée des informations fournies par un test pronostique, sur un intervalle de risques et de bénéfices d'un patient. Cette méthode facilite la prise de décision clinique. La DCA est exprimée graphiquement par des courbes, avec le bénéfice-net clinique sur l'axe vertical et les seuils de probabilité sur l'axe horizontal. Lorsque la courbe atteint son niveau maximal sur l'intervalle des seuils de probabilité, l'intervention associée serait la meilleure décision. La figure 1.2 synthétise les différentes démarches et méthodes jugées importantes lors de la construction et l'évaluation d'un modèle de prédiction clinique.

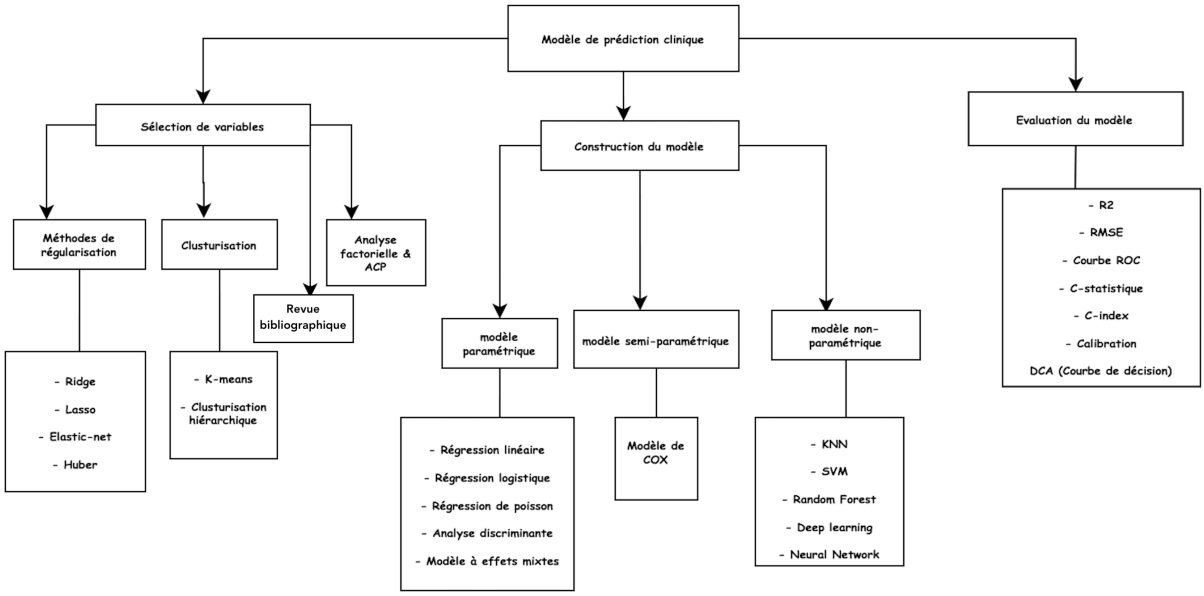

FIGURE 1.2 – Construction et évaluation d'un modèle de prédiction clinique

1.3.1.3 Les conditions nécessaires pour construire un modèle de prédiction clinique du point de vue des cliniciens

Du point de vue des cliniciens, de nombreuses conditions s'imposent pour construire un modèle de prédiction clinique :
(1) Construire une base de données de suivi d'une maladie et collecter des informations sur les patients, y compris mais sans s'y limiter : les caractéristiques démographiques, les antécédents familiaux, les antécédents personnels ; des informations relatives à la maladie telles que des résultats physiques et de laboratoire importants avant le traitement, la gravité de la maladie, le stade clinique, le stade pathologique ; informations sur le traitement : telles que les méthodes chirurgicales, les schémas de radiothérapie et de chimiothérapie, la dose et l'intensité.
(2) Une certaine taille d'échantillon est nécessaire pour atteindre une efficacité statistique suffisante pour discerner l'influence des facteurs de confusion sur le résultat [79].
(3) La construction d'un modèle de prédiction clinique consiste à résoudre des problèmes cliniques. La capacité de découvrir des problèmes cliniques de bon sens est cultivée à travers une large lecture et pratique clinique.

1.3.1.4 Problèmes actuellement rencontrés dans le développement du modèle de prédiction

Plusieurs problèmes sont rencontrés lors du développement d'un modèle de prédiction et peuvent être énumérés comme suit :
(1) Un faible taux de conversion clinique (utilisation du modèle par les cliniciens). En effet, la raison principale est que l'application clinique du modèle de prédiction doit être équilibrée entre la précision et la simplicité du modèle.
(2) La plupart des modèles de prédiction clinique sont construits et validés sur la base d'ensembles de données rétrospectives et la validation est rarement effectuée dans les données prospectives. Par conséquent, la stabilité des résultats prédits par les modèles est relativement faible.
(3) La validation de la plupart des modèles de prédiction clinique est basée sur des données internes. En général, il y a deux jeux de données, construction et validation. Les deux

ensembles de données proviennent souvent du même registre.
(4) Si le modèle de prédiction est validé par un ensemble de données externe, l'application du modèle sera considérablement élargie. En revanche, cette approche est difficile et nécessite une coopération multicentrique. De plus, la plupart des centres de recherche ne disposent pas d'une base de données complète pour la validation, d'où l'importance d'une base de données complète.

1.3.2 Application : Accident vasculaire cérébral (AVC)

1.3.2.1 L'accident vasculaire cérébral et ses conséquences cliniques

L'AVC est une maladie cérébrovasculaire décrite par l'organisation mondiale de la santé (OMS) comme suit :
"Un accident vasculaire cérébral (AVC) résulte de l'interruption de la circulation sanguine dans le cerveau, en général quand un vaisseau sanguin éclate ou est bloqué par un caillot. L'apport en oxygène et en nutriments est stoppé, ce qui endommage les tissus cérébraux".
Un accident vasculaire cérébral peut être défini comme une hémorragie ou un infarctus. Un accident vasculaire cérébral hémorragique survient en cas d'interruption de l'apport sanguin au cerveau suite à un saignement d'un vaisseau sanguin dans le cerveau, appelé hémorragie. Un infarctus survient lorsqu'un blocage, tel un caillot, est responsable de l'interruption de l'approvisionnement en sang (OMS, 2018). Ceci est communément appelé AVC ischémique. Les lésions cérébrales qui résultent d'un AVC hémorragique ou ischémique peuvent être mortelles ou avoir des conséquences graves [6].
Le registre des accidents vasculaires cérébraux du sud de Londres (SLSR), principale source de données utilisée dans cette thèse, décrit les AVC selon les critères de l'OMS. Le SLSR utilise la Classification internationale des maladies (CIM) pour définir l'AVC. La CIM est une classification médicale codifiée classifiant les maladies et une très vaste variété de signes, symptômes, lésions traumatiques, empoisonnements, circonstances sociales et causes externes de blessures ou de maladies.
Elle est publiée par l'Organisation mondiale de la santé (OMS) et est mondialement utilisée pour l'enregistrement des taux de morbidité et des taux de mortalité touchant le domaine de la médecine. Elle permet : le stockage, la récupération et l'analyse faciles des informations sur la santé pour une prise de décision fondée sur des preuves, le partage et la comparaison des informations sur la santé entre les hôpitaux, les régions, les milieux et les pays. Et aussi, la comparaison de données obtenues au même endroit sur différentes périodes.
L'AVC est causé par des facteurs tels que l'hypertension, le diabète et l'hyperlipidémie qui peuvent avoir un effet continu et cumulatif [80]. L'invalidité après un AVC est une conséquence qui entraîne des souffrances chez le patients et est une charge pesante sur leurs familles et soignants [81]. Une étude bibliographique épidémiologique a été entreprise par Fehey et al. [82], pour l'identification des conséquences potentiellement importantes de l'AVC. Cette étude a produit une liste de conséquences d'AVC et a été présentée et discutée avec le groupe "SLSR Stroke Research Patients and Family Group (SRPFG)".
Les conséquences post-AVC qui ont été sélectionnées pour des études plus approfondies et ce en partenariat avec le SRPFG comprennent : la récupération fonctionnelle, la mortalité, la déficience cognitive et la dépression.
Dans cette thèse, l'intérêt est porté principalement à la déficience cognitive post-AVC. Dans ce qui suit, nous commençons par justifier les motivations derrière ce choix. Ensuite nous passons en revue les facteurs de risque majeurs associés à la déficience cognitive post-AVC. Finalement nous discutons les travaux menés sur la méthodologie des "courbes de récupération".

1.3.2.2 Déficience cognitive

La déficience cognitive post-AVC (PSCI) est la deuxième cause la plus courante de déclin cognitif après la maladie d'Alzheimer [83]. Le risque de développer un accident vasculaire cérébral ou la démence à l'âge de 65 ans est d'un sur trois pour les hommes et d'un sur deux pour les femmes (Institut national d'excellence clinique (NICE), 2008). L'augmentation de l'espérance de vie, l'évolution démographique de la population et l'amélioration des taux de survie après un AVC signifient que le nombre absolu de patients atteints de PSCI augmentera.
Plusieurs études prospectives ont montré que des changements cognitifs surviennent après un AVC [84, 85]. La méta-analyse de ces études suggère que jusqu'à un tiers des survivants d'un AVC souffrent d'une forme de déficience cognitive. Cette statistique ne fait pas de différence entre les étiologies de l'AVC et peut inclure une proportion de patients (estimée à environ 10%) atteints de démence pré-AVC. La prévalence du PSCI est plus élevée chez ceux avec un AVC récurrent [86]. Ces données pourraient sous-estimer la prévalence réelle de PSCI, car ces études ont tendance à exclure les patients avec un état sévère et qui sont incapables de consentir à la collecte de données ou d'assister à des visites de suivi [87]. Il existe également des problèmes de faisabilité des tests cognitifs et d'attrition au suivi qui contribuent à une sous-estimation du risque des troubles cognitifs [87]. Le PSCI évolue avec le temps, et est souvent reconnu dans les premières semaines à quelques mois après un AVC. Les analyses groupées des cohortes longitudinales hospitalières et de population suggèrent une augmentation linéaire du PSCI d'environ 2% par an [86]. Ce fait est différent avec les changements cognitifs dans la population vieillissante normale, dans laquelle un taux de changement linéaire n'est pas observé. Les survivants d'un AVC montrent différentes trajectoires cognitives. Dans cette thèse, la déficience cognitive est définie à l'aide du Mini Mental State Examination (MMSE)[88]. Un seuil$>$ 15 indique une déficience cognitive. Une discussion sur ce point est donnée au chapitre 3.

1.3.2.3 Facteurs de risque

Patel et al ont constaté que les troubles cognitifs causées par un AVC étaient associés aux facteurs suivants, tous indépendants les uns des autres : un patient de 75 ans et plus ; ethnicité ; classe socio-économique inférieure ; accident vasculaire cérébral de l'hémisphère gauche ; et les indicateurs de la gravité initiale de l'AVC, y compris l'incontinence urinaire et les anomalies du champ visuel [89]. L'influence de l'âge sur le déclin cognitif a été démontrée non seulement chez les patients victimes d'un AVC, mais aussi dans la population générale [90]. L'ethnicité est également un facteur de risque. Les ethnies caribéenne / africaine et asiatique sont associées à de mauvais résultats cognitifs après un AVC [90]. Cela pourrait être dû à une variation culturelle en répondant au questionnaire MMSE. La corrélation identifiée entre l'origine ethnique et les troubles cognitifs s'est avérée indépendante de l'âge ou d'autres facteurs potentiels [90]. Nous avons également constaté que la déficience cognitive est associée à une classe socio-économique inférieure [91, 90]. Ces études ont rapporté qu'un niveau d'étude très bas était associé à une déficience cognitive après un AVC. Les patients victimes d'un AVC de l'hémisphère gauche étaient plus susceptibles de souffrir de troubles cognitifs que ceux ayant des lésions du côté droit, peut-être parce que l'hémisphère gauche contrôle le langage et les fonctions intellectuelles générales [89]. L'incontinence urinaire et le défaut du champ visuel sont deux marqueurs de la gravité initiale de l'AVC et sont significativement associés à une déficience cognitive à long terme, comme le rapporte Douiri, Rudd et Wolfe [92]. Une revue systématique et une méta-analyse menées par Pendlebury et al. [83] ont indiqué que la survenue d'un accident vasculaire cérébral semblait être le déclencheur d'un déclin cognitif rapide. Les

prédicteurs significatifs associés à la démence post-AVC comprenaient un AVC antérieur, un AVC récurrent, plusieurs lésions d'AVC, la sévérité de l'AVC, le type d'AVC (augmenté avec hémorragie), le volume de l'infarctus et l'emplacement de l'AVC (augmenté avec l'AVC dans l'hémisphère gauche) [83] .

1.3.2.4 Modélisation statistique pour la prédiction des conséquences post-AVC : "Courbes de récupération"

Historiquement, la prédiction du risque d'accident vasculaire cérébral concernait principalement le risque avant l'accident vasculaire cérébral : c'est-à-dire l'identification des facteurs de risque d'un accident vasculaire cérébral ou la construction de modèles qui prédisent le risque d'avoir un AVC en premier lieu. Des modèles tels que celui utilisé pour calculer le score de Framingham [93] sont couramment utilisés en pratique clinique et ont été adaptés pour prédire à la fois le risque à court et à long terme dans diverses populations.
Dans l'espace post-AVC, la recherche sur la prédiction du risque a identifié des facteurs de risque de mortalité, de mauvaise progression de récupération et d'AVC récurrent. De nombreux modèles ont été élaborés pour prédire les résultats à court et à long terme.
Ces modèles à long terme comprennent : The Essen Stroke Risk Score (ESRS) [94], Stroke Prognostic Instrument One (SPI I) [95], The Stroke Prognostic Instrument Two (SPI II) [96], Life Long After Cerebral ischémie (LiLAC) 1, LiLAC 2, LiLAC 3 [97] et le score Acute Stroke Registry and Analysis of Lausanne (ASTRAL) [98].
Ces modèles ont été développés d'une manière similaire aux modèles pré-AVC ; c'est-à-dire qu'ils prédisent la probabilité qu'une population, ou un patient, obtienne un résultat à des moments fixes et prédéfinis de façon unique. Contrairement à la prédiction du risque pré-AVC, les modèles de prédiction du risque post-AVC ne sont pas utilisés dans la pratique clinique.
Il a été théorisé que la raison en est qu'une prédiction longitudinale est plus significative que les estimations de risque à des moments prédéfinis [10, 7, 99, 8, 9].
Par exemple, une estimation de la fluctuation du risque de mortalité d'un patient sur une période d'un an après un AVC peut être plus utile pour les patients et les cliniciens qu'une estimation du risque de mortalité au 365^{me} jour après un AVC.
Cet argument est raisonnable car la progression de la récupération après un AVC s'est avérée très variable à la fois dans le temps et entre les individus. Compte tenu de cette variabilité, les données et les méthodes peuvent être plus appropriées pour capturer avec précision le rétablissement du patient, en particulier lorsque l'objectif est la planification simultanée de soins immédiats et à long terme.
Tous les modèles post-AVC identifiés dans la revue systématique [82] estiment le risque à des moments prédéfinis, et aucun n'est utilisé en pratique quelle que soit la qualité méthodologique du modèle. La seule exception à cela concerne les modèles dérivés par Tilling et ses collègues, Toschke et Douiri [99, 8, 9]. Ces chercheurs ont développé des méthodes pour la prédiction longitudinale du risque des conséquences post-AVC.

Tilling [10] a émis l'hypothèse que les modèles hiérarchiques pouvaient décrire et prédire les conséquences après une maladie. En effet, le nombre de patients dans une étude peut diminuer avec le temps, alors qu'en général, chaque modèle doit être basé sur tous les patients, et non seulement ceux qui survivent jusqu'à la fin de l'étude. Ces problèmes peuvent être traités avec des modèles hiérarchiques, qui prennent en compte le fait que les données sont emboîtées sous forme d'observations (premier niveau) au sein des individus (deuxième niveau) [7]. Tilling a appelé cette méthodologie *"outcomes curves"* [7].

Tilling et ses collègues ont proposé l'utilisation des *"outcomes curves"* pour prédire la récupération fonctionnelle après un AVC. Par cette méthode, un modèle de récupération a été quantifié jusqu'à 12 mois après l'AVC et le terme "courbe de récupération" (*recovery curve*) a été inventé [10, 7]. Cette méthodologie suppose que chaque patient a sa propre courbe de récupération réelle et que ces courbes de récupération réelles varient par rapport à la courbe de récupération moyenne. Cette propriété permet aux modèles de récupération de différer, même pour les patients ayant les mêmes caractéristiques de base. En plus d'estimer la courbe de récupération moyenne ainsi que la manière dont cette courbe est liée aux caractéristiques du patient, cette approche permet également de quantifier la variation entre les courbes de récupération de chaque patient et leur degré de variabilité.

L'avantage des courbes de récupération est qu'elles peuvent prédire la récupération fonctionnelle à tout moment après un AVC. Les prévisions générées par ces courbes de récupération ne sont pas aussi précises que celles produites par d'autres modèles de la littérature ; cependant, ils apportent une plus grande dimension à la prédiction car les prédictions ne sont pas limitées à un moment précis. Cela permet de modifier les prédictions au fur et à mesure de l'observation, et permet ainsi aux médecins d'identifier les patients qui présentent un risque plus élevé par rapport aux autres. Les courbes de récupération permettent également de quantifier les trajectoires de récupération après un AVC.

Toschke a évalué les effets des traitements fondés sur des données probantes, sur l'évolution des conséquences post-AVC. [8]. Toschke a réitéré que les courbes de récupération pouvaient être utilisées pour quantifier le modèle de récupération jusqu'à un an après l'AVC. Il a également montré que ces courbes pouvaient évaluer les effets des traitements et d'interventions complexes fondés sur des données probantes tout au long de la période de rétablissement du patient. Ainsi, les modèles ont contribué à l'étendue des connaissances cliniques sur les traitements efficaces [8].

Douiri et al [9] ont étendu les travaux de Toschke et Tilling en montrant que l'on pouvait également construire des courbes de récupération avec des prédicteurs disponibles seulement à l'admission du patient. De plus, Douiri et ses collègues ont construit des courbes de récupération intégrant la régularisation. Ils ont également été les premiers à évaluer l'utilité clinique des courbes de récupération en effectuant une analyse de la courbe de décision (DCA).

Cette thèse développe la méthodologie des courbes de récupération, en utilisant les travaux antérieurs susmentionnés comme base. Les courbes de récupération se sont avérées être la méthode la plus flexible permettant d'incorporer des informations de prédicteur longitudinal sans perte de qualité prédictive. Cette méthode est plus avantageuse par rapport aux autres en termes de flexibilité avec les valeurs manquantes, sa capacité à englober toutes les informations concernant le développement du prédicteur, et son utilité face à un grand nombre de mesures répétées [100].
Dans ce travail de thèse, cette méthodologie est utilisée pour développer et valider un modèle prédictif centré sur le patient pour estimer le risque de déclin cognitif à long terme (5 ans) après un AVC.
Ce travail intègre la régularisation dans la méthodologie des courbes de récupération. Enfin, il évalue également la performance de cette méthode en utilisant à la fois des para-

mètres traditionnels (discrimination et étalonnage) et non traditionnels (utilité clinique).

Chapitre 2

Une approche problème inverse pour les modèles de régression régularisés avec application à la prédiction de la récupération fonctionnelle après un AVC

2.1 Introduction

Les modèles de régression représentent, de nos jours, un outil majeur couramment utilisé dans les essais cliniques et les études épidémiologiques. Ces modèles ont été affinés grâce aux travaux de Fisher qui a combiné les travaux de Gauss et Pearson pour développer les propriétés de l'estimateur des moindres carrés [101]. Ces travaux ont permis l'utilisation de la régression pour ajuster les variables indépendantes et dépendantes et identifier les relations causales entre les prédicteurs et les observations [102]. Depuis la publication des travaux de Fisher, les modèles de régression ont connu un développement important, en particulier la régression paramétrique et non paramétrique [103], la régression bayésienne [104] et la régression régularisée [105]. Gauss a introduit la distribution normale [106], couramment utilisé comme distribution du terme d'erreur dans l'analyse de régression. Sous l'hypothèse que les erreurs sont normalement distribuées, les estimateurs des moindres carrés (MCO) [107] et ceux du maximum de vraisemblance (ML) sont identiques. Cependant, les hypothèses classiques sur les MCO [108] ou le ML sont souvent insatisfaites par les données réelles. Les MCO/ML peinent à bien prédire et à donner des interprétations significatives dans certains scénarios. Pour surmonter ces limitations, de nombreuses techniques de régularisation ont été développées telles que la régression ridge [4], le lasso [54], l'adaptive-lasso [3] et l'elastic-net [5]. D'un point de vue mathématique, l'ajustement d'un modèle de régression peut être formulé comme une minimisation d'un problème inverse qui mesure l'écart entre le résultat observé et les données produites par la représentation des prédicteurs. L'approche problème inverse, permet d'effectuer simultanément la sélection de variables et l'estimation des coefficients.
En particulier, pour un problème de régression linéaire généralisée [15] :
Soient $X \in M_{n,p}(\mathbb{R})$ et $Y \in \mathbb{R}^n$, nous cherchons $\beta \in \mathbb{R}^p$ le paramètre à estimer tel que : $g(Y) = X_1\beta_1 + X_2\beta_2 + \ldots + X_p\beta_p$, où $X_1, X_2, .., X_p$ les colonnes de X qui sont des variables aléatoires intégrables à valeur réelle.
En statistique, on dit que l'observation peut être expliquée par des covariables $X_1, X_2, ..., X_p$. Puisque les observations sont toujours accompagnées d'erreurs de mesure et dans certaines

situations des discontinuités, la minimisation de la distance entre les données observées et ajustées, donne des solutions insatisfaisantes. Dans ce cas, l'introduction d'une information préalable est nécessaire pour régulariser la solution.
Le problème de minimisation est énoncé comme suit :

$$\begin{cases} \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\text{Minimize}} & -\boldsymbol{log}(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{p}(\boldsymbol{y_i}|\boldsymbol{x_i},\boldsymbol{\beta},\boldsymbol{\sigma})) + \boldsymbol{\lambda}\sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{\psi}(\boldsymbol{\beta_i}) \\ \text{subject to} & \boldsymbol{\beta} \in \boldsymbol{C} \end{cases} \tag{2.1}$$

- λ est un paramètre de régularisation positif.
- C est un ensemble de contraintes sur les informations à priori de β.
- $\psi : \mathbb{R} \to \mathbb{R}$ est une fonction de régularisation.
- $-log(\sum_{i=1}^{n} p(y_i|x_i,\beta,\sigma))$ est appelé terme de fidélité.

Le problème (2.1) peut être exprimé par :

$$\begin{cases} \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\text{Minimize}} & \frac{\boldsymbol{1}}{\boldsymbol{2}}||\boldsymbol{Y}-\boldsymbol{X\beta}||^{\boldsymbol{2}} + \boldsymbol{\lambda}\sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{\psi}(\boldsymbol{\beta_i}) \\ \text{subject to} & \boldsymbol{\beta} \in \boldsymbol{C} \end{cases} \tag{2.2}$$

Ou sous la forme intégrale par :

$$\begin{cases} \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\text{Minimize}} & \frac{\boldsymbol{1}}{\boldsymbol{2}}||\boldsymbol{Y}-\boldsymbol{X\beta}||^{\boldsymbol{2}} + \boldsymbol{\lambda}\int \boldsymbol{\psi}(\boldsymbol{\beta})\ \boldsymbol{d\beta} \\ \text{subject to} & \boldsymbol{\beta} \in \boldsymbol{C} \end{cases} \tag{2.3}$$

Avec $\psi : \mathbb{R}^p \to \mathbb{R}$.
Choisir une fonction de régularisation optimale pour un problème inverse donné n'est pas simple. Dans ce sens, deux considérations importantes sont à prendre en compte : Une bonne précision de prédiction sur les données futures et une interprétation significative des résultats du modèle. Pour illustrer l'erreur liée au problème d'ajustement dans un modèle de régression, nous donnons un exemple simple (ci-dessous) montrant l'impact d'une observation perturbée sur l'exactitude de l'estimation.

$$X = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} Y = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \beta = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} Y+\delta Y = \begin{pmatrix} 32.1 \\ 23.9 \\ 33.1 \\ 31.9 \end{pmatrix} \beta+\delta\beta = \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix}$$

Dans ce travail, nous abordons les modèles de régression régularisés en utilisant le cadre théorique des problèmes inverses. Cette approche représente une nouvelle démarche et méthodologie pour analyser ces modèles statistique. Nous étudions le modèle de régression dans le cadre du problème inverse statistique, expliquant l'importance de la régularisation dans les modèles analytiques prédictifs. L'analyse théorique est étayée par une étude de simulation afin d'évaluer et de comparer différentes stratégies de régularisation dans le cas des données, de petite et grande dimension. En se basant sur la méthodologie statistique robuste de Huber [109] et le cadre théorique du problème inverse, nous caractérisons et proposons une nouvelle fonction de régularisation pour le problème de régression.
Nous évaluons et comparons les performances de la méthode proposée avec d'autres méthodes de régularisation (lasso, adaptive-lasso, ridge, elastic-net) par des simulations et des données réelles basées sur le registre des AVC du sud de Londres (SLSR).

2.2 Methodologie

2.2.1 Problème inverse statistique dans un contexte de régression

Le problème inverse consiste à trouver une estimation pour $\beta = (\beta_1, \beta_2, .., \beta_p)$ qui est liée par un opérateur L : $\beta \longrightarrow X\beta$ pour ajuster les données Y [110].
Les données mesurées $\bar{Y}$ sont approximées par des données observée Y avec une erreur $\delta > 0$ tel que $||\bar{Y} - Y|| \leq \delta$.
En se référant à la définition de Hadamard [111], on dit qu'un problème est mal posé si l'injectivité, la surjectivité et la stabilité ne peuvent être assurées.
Nous étudierons l'influence du terme de régularité pour que le problème ait une solution unique, stable, statistiquement significative et interprétable.
Dans un modèle de régression, nous supposons que nous observons un $Y \in \mathbb{R}^n$ et que nous avons une matrice de prédicteurs $X \in M_{n,p}(\mathbb{R})$, dont les colonnes $X_1, ..., X_p \in \mathbb{R}^n$ correspondent à des variables prédictives.
En utilisant la définition de Hadamard pour le problème de régression, nous établissons d'abord la surjectivité et l'injectivité à travers la sélection d'une solution significative en introduisant une information à priori. Ensuite nous formulons une généralisation du problème inverse régularisé.
Pour formuler une généralisation du problème de régression régularisée, nous utilisons une analyse heuristique basée sur quatre cas :
Le cas simple est quand X est bijective ($n = p$), le problème est analytiquement bien posé, la solution de $Y = X\beta$ vérifie les conditions d'Hadamard et est unique, mais reste très sensible aux petites perturbations. De plus, lorsque X est mal conditionnée, $X^{-1}Y$ amplifie l'erreur ce qui aboutit à une solution irrégulière.
Cependant, lorsque X n'est pas surjective ($Ran(X) < n \Leftrightarrow p < n$), l'observation Y peut ne pas appartenir à $Ran(X)$ i.e $Y \notin Ran(X)$. Nous remplaçons donc Y par sa projection sur $Ran(X)$ i.e $\bar{Y} = \underset{Ran(X)}{\text{Proj Y}}$ pour rétablir la surjectivité. Le problème devient $\bar{\beta} = argmin||Y - X\beta||^2$ avec $\bar{\beta}$ la solution unique de l'équation normale $X'Y = X'X\beta$ appelé solution des moindres carrés.
Lorsque X n'est pas injective ($Ran(X) < p$), nous avons une infinité de solutions formant une variété affine de dimension $(n - p)$ i.e. tous les vecteurs dans le sous-espace affine $X^{-1}\{Y\}$ sont des solutions. Par conséquent, nous devons sélectionner une solution significative dans le sous-espace $X^{-1}\{Y\}$. Pour ce faire, nous introduisons une information préalable impliquant le choix d'un principe d'inférence tel que la régularité de la solution ou d'autres critères de restauration. Cela permettra de réduire le sous-espace $X^{-1}\{Y\}$ et régulariser le problème inverse.

Alors le problème s'écrit :

$$\begin{cases} \underset{\boldsymbol{\beta} \in \mathbb{R}^{\boldsymbol{p}}}{\text{Minimize}} & \boldsymbol{\Psi(\beta)} \\ \text{subject to} & \boldsymbol{\beta \in X^{-1}\{Y\}} \end{cases} \tag{2.4}$$

La fonction $\Psi(\beta)$ peut être définie par :

$\Psi(\beta) = \sum_{j=1}^{p} |\beta_j|$ (lasso)
$\Psi(\beta) = \sum_{j=1}^{p} w_j|\beta_j|$ avec $w_j = \frac{1}{|\hat{\beta}_j|^v}$ et $v > 0$ (adaptive-lasso)
Les w_j sont des poids adaptifs. Comme suggéré par Zou [3], $\hat{\beta}_j$ peut être l'estimateur des moindres carrés ordinaires ou dans le cas de grande dimension, l'estimateur ridge.
$\Psi(\beta) = \sum_{j=1}^{p} \beta_j^2$ (Ridge)

$\Psi(\beta) = \sum_{j=1}^{p} |\beta_j - \beta_{j-1}|$ (Total variation)
$\Psi(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$ (elastic-net)

Lorsque X n'est ni injective ni surjective, la combinaison des deux stratégies précédentes permet de trouver une solution qui optimise la fonction $\Psi(\beta)$ sur le sous-espace affine $X^{-1}\{\bar{Y}\}$ avec $\bar{Y} = \underset{Ran(X)}{\text{Proj}}\ Y$.
Cela peut être exprimé comme suit :

$$\begin{cases} \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\text{Minimize}} & \boldsymbol{\Psi(\beta)} \\ \text{subject to} & \boldsymbol{\beta \in X^{-1}\{\bar{Y}\}} \end{cases} \tag{2.5}$$

Avec : $\Psi : F(\mathbb{R}, \mathbb{R}^p) \longrightarrow \mathbb{R}$ et $F(\mathbb{R}, \mathbb{R}^p)$ est l'espace des fonctions défini de $\mathbb{R}$ à $\mathbb{R}^p$.
En utilisant les multiplicateurs de lagrange, nous pouvons facilement montrer que le problème (2.5) est équivalent au problème (2.3).
D'un point de vue problème inverse, nous cherchons une fonction de régularisation Ψ et un paramètre λ qui optimise au mieux le problème (2.3).

2.2.2 Lien avec le cadre bayésien

L'approche bayésienne consiste à trouver le paramètre β sachant les données mesurées (observations) Y et le modèle de problème direct $X\beta$.
$P(\beta)$ représente la probabilité a priori du paramètre β (coefficient), et $P(Y|\beta)$ est la fonction de vraisemblance.
Le théorème de Bayes [104] défini par : $P(\beta|Y) = P(Y|\beta)P(Y)/P(Y)$ exprime la probabilité postérieure. $P(Y)$ est simplement une constante de normalisation une fois que les mesures Y sont données. Le maximum a postérieuri (MAP) permet de formaliser le problème inverse comme un problème statistique, où la probabilité postérieure $P(\beta|Y)$ est maximisée par rapport à β. Ceci revient à minimiser $E(\beta)$ =-$logP(\beta)$ -$logP(Y|\beta)$ (en intégrant la fonction log).
Pour écrire le modèle bayésien régularisé sous la forme (2.3), nous utilisons une distribution de Boltzmann ou Gibbs de la forme : $P(\beta)=\frac{1}{Z}\ exp(-\int \psi(\beta)d\beta)$ [112].
$P(\beta)$ est ensuite combiné avec le modèle basé sur des mesures et un bruit spécifique : $P(Y|\beta) = \frac{1}{Z'}exp(-\frac{1}{\lambda}||Y - X\beta||^2)$, où Z et Z' sont des constantes de normalisation, appelé également, fonction de partition. Par conséquent, la formulation du modèle bayésien régularisé est équivalente au problème (2.3).
Dans le cas où nous avons un modèle de bruit gaussien, nous présumons que $||.|| = ||.||_2$.

2.2.3 Méthode proposée : Nouvelle fonction de régularisation (hybride)

Dans cette section, nous considérons le problème (2.5) et nous définissons :

$$\boldsymbol{\Psi(\beta)} = \int_{\boldsymbol{\Omega}} \boldsymbol{\psi(|\beta(t)|)}\, \mathbf{d}\boldsymbol{t}. \tag{2.6}$$

Avec :

$$\Omega \subset \mathbb{R} \mapsto \mathbb{R}^p \mapsto \mathbb{R}$$
$$t \mapsto \beta(t) \mapsto \psi(|\beta(t)|)$$

Pour assurer que $\int_\Omega \psi(|\beta(t)|)\,\mathrm{d}t$ est convexe, nous avons besoin des hypothèses suivantes :
i) Ω est un ensemble convexe de $\mathbb{R}$
ii) ψ est une fonction par morceaux deux fois différentiables dans $\mathbb{R}$
iii) $\psi(0) = 0\,;\ \psi'(0) = 0$
iv) $\psi''(s) \geq 0$ et $\psi'(s) \geq 0\ \forall\ s \geq 0$

Rappel du problème :
Le problème consiste à trouver le minimum optimal sur un domaine convexe $\Omega \subset \mathbb{R}$ de la fonctionnelle suivante :

$$\underset{\boldsymbol{\beta}}{\text{Minimize}} \quad \boldsymbol{E(\beta) = \frac{1}{2}||Y(t) - X\beta(t)||^2 + \lambda \int_\Omega \psi(|\beta(t)|)\,\mathrm{d}t} \quad . \tag{2.7}$$

Supposons que $E(\beta)$ admet un minimum $\widehat{\beta}$, alors le problème satisfait la proposition suivante.

proposition 2.1. *$E'(\beta)$ satisfait l'équation suivante :*

$$\boldsymbol{X'(X\beta(t) - Y(t)) + \lambda \frac{\psi'(|\beta(t)|)}{|\beta(t)|}|\beta(t)| = 0.} \tag{2.8}$$

Démonstration. Supposons que toutes les hypothèses susmentionnées sont satisfaites, (iv) affirme que $\int_\Omega \psi(|\beta(t)|)\,\mathrm{d}t$ est convexe , ce qui assure la convexité de $E(\beta)$, alors le problème admet un minimum global.
Calculons la dérivée de $E(\beta)$:

Pour résoudre le problème de la non différentiabilité de la valeur absolue à $t = 0$, on considère : $|\beta| = \sqrt{|\beta|^2 + \epsilon}$

Nous avons $D(\beta \longrightarrow \int_\Omega \frac{1}{2}|X\beta(t) - Y|^2\ \mathrm{d}t) = \int_\Omega [X'(X\beta - Y)]$

et $D(\beta \longrightarrow \int_\Omega \psi(|\beta|)\mathrm{d}t) = \frac{d}{dt}[\int \psi(|\beta + th|)]_{t=0}$

On a : $\eta(t) = |\beta + th|^2 = |\beta|^2 + 2t\beta h + t^2.|h|^2$

et $\eta'(t) = 2\beta h + 2t|h|^2$

Alors $\frac{d}{dt}[\int \psi(\sqrt{\eta(t)})]_{t=0} = \frac{1}{2}\int \psi'(\sqrt{\eta(0)})\frac{\eta'(0)}{\sqrt{\eta(0)}} = \int \psi'(|\beta|)\frac{\beta}{|\beta|}h$

D'où, $\int_\Omega [X'(X\beta - Y)] + \lambda\psi'(|\beta|)\frac{\beta}{|\beta|}h = 0$

On conclut que $E'(\beta) = X'(X\beta - Y) + \lambda\frac{\psi'(|\beta|)}{|\beta|}\beta$

□

Inspiré par le cadre statistique robuste, nous définissons ψ par la fonction de Huber (2.9), une fonction de régularisation optimale qui vérifie les hypothèses ci-dessus :

$$\boldsymbol{\psi_\sigma(|\beta|) = \begin{cases} \frac{1}{2}\beta^2 & Si\ \ |\beta| \leq \sigma;\ \ \sigma > 0 \\ \sigma(|\beta| - \frac{\sigma}{2}) & sinon. \end{cases}} \tag{2.9}$$

Le sous gradient de ψ est donné par :

$$\frac{\psi'_{\sigma}(|\boldsymbol{\beta}|)}{|\boldsymbol{\beta}|} = \begin{cases} \dfrac{\boldsymbol{\beta}}{|\boldsymbol{\beta}|} & \boldsymbol{si} \quad |\boldsymbol{\beta}| \leq \boldsymbol{\sigma} \\ \dfrac{\boldsymbol{\sigma sign(\beta)}}{|\boldsymbol{\beta}|} & \boldsymbol{sinon.} \end{cases} \tag{2.10}$$

Dans la littérature, la norme de Huber est souvent utilisée comme une fonction de perte (loss function). Dans notre étude, la fonction de Huber est utilisée comme un terme de régularisation, ce qui rend notre méthode moins sensible aux valeurs aberrantes. Cette approche est fréquemment utilisée en tomographie optique [113], en résolution d'image [114] etc, et a montré son efficacité.
Le régulariseur Huber bénéficie des propriétés des normes L_1 et L_2. Il est quadratique près de l'origine et linéaire loin de l'origine (figure 2.1). À cette fin, σ a été spécifié pour assurer un passage fluide entre quadratique et linéaire. Huber a recommandé que la constante σ, qui régule le degré de robustesse, soit comprise entre 1 et 2. Généralement, on choisit $\sigma = 1.5$.
Notons que σ peut être adaptative ou choisie par la méthode dite "L-curve" [115].
La figure 2.1 est une illustration géométrique des régularisateurs étudiés, permettant ainsi d'avoir une intuition sur le comportement asymptotique de ces méthodes. Nous remarquons que la méthode Huber croît linéairement pour les vecteurs associés à de grands poids, tandis que l'elastic-net croît quadratiquement.

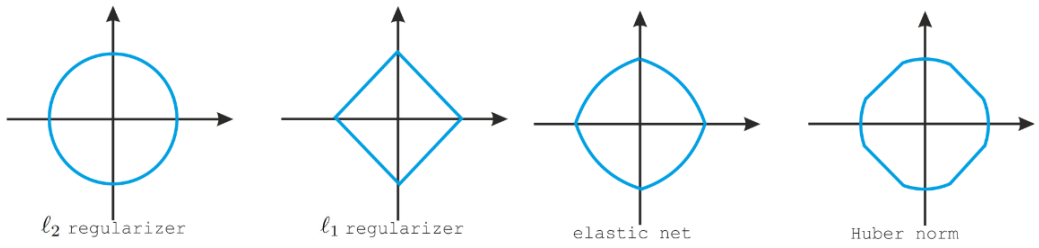

FIGURE 2.1 – Illustration géométrique des normes L_1, L_2, Elastic-net et Huber

Pour résoudre le problème (2.7) par la méthode proposée et les autres méthodes de régularisation (ridge, lasso, adaptive-lasso, elastic- net), nous avons utilisé une méthodologie basée sur l'optimisation convexe (*CVXR package*)[116].
CVXR est basé sur la méthode des points intérieurs (IPM), une classe d'algorithmes introduite par Karmarka [117]. Cette méthode utilise une fonction barrière logarithmique [118] pour décrire l'ensemble des solutions du problème.
En effet, résoudre le problème de minimisation (2.11) est équivalent à résoudre le problème de minimisation (2.12) appelé méthode barrière.

$$\begin{cases} \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\text{Minimize}} & \boldsymbol{f(\beta)} \\ \text{subject to} & \boldsymbol{h(\beta) = 0} \\ & \boldsymbol{g(\beta) \leq 0} \end{cases} \tag{2.11}$$

avec $h : \mathbb{R}^p \to \mathbb{R}^p$ et $g : \mathbb{R}^p \to \mathbb{R}^p$

$$\begin{cases} \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\text{Minimize}} & \boldsymbol{f(\beta) - \frac{1}{t}\sum_{i=1}^{p} log(-g_i(\beta))} \\ \text{subject to} & \boldsymbol{h(\beta) = 0} \end{cases} \tag{2.12}$$

Pour résoudre le problème (2.12), nous commençons avec une petite valeur pour t, nous calculons une solution pour cette valeur et ensuite nous utilisons la solution trouvée comme point de départ pour la prochaine itération, où la valeur de t est augmentée.

Algorithm 2 Algorithme "*Barrier method*"

1. START strictly feasible $\beta_k := \beta_0$, $t := t^0$, $\alpha > 0$, $\epsilon > 0$
2. Until $p/t \leq \epsilon$
a) Centring Step : Solve problem (Barrier Method) with $t := t^k$ starting from β_k
b) Update $\beta_{k+1} := \beta_k^*$
c) $t^{k+1} := \alpha t^k$
End

Les approximations s'améliorent en tendant t vers l'infini.
Le problème peut être résolu en utilisant les conditions de Kuhn-Tucker.
Dans le cas où le problème est convexe mais la fonction objectif est non différentiable, comme pour l'exemple du lasso, nous ne pouvons pas utiliser directement la méthode des points intérieurs. Nous transformons donc le problème en ajoutant éventuellement de nouvelles variables et contraintes. Le problème transformé a cependant plus de variables et de contraintes que le problème original. En revanche, nous pouvons le résoudre très efficacement via la méthode des points intérieurs.

2.2.4 Colinéarité, Conditionnement et test de Belsley, Kuh et Welsch

Le test de Belsley, Kuh et Welsch [119] est une méthode qui permet de classer les données affectées par la colinéarité en se basant sur le conditionnement de la matrice de données.
Notons C_i l'indice de conditionnement de la matrice avec : $C_i = d_i/d_{min}$ et $d_i = +\sqrt{\lambda_i}$ sont les valeurs singulières de la matrice.
Nous trions par ordre décroissant les d_i tel que $d_k = d_{max} > d_{k-1} > ... > d_2 > d_1 = d_{min}$.
Alors $d_1/d_{min} < d_2/d_{min} < ... < d_{k-1}/d_{min} < d_{max}/d_{min}$.
Et donc, $C_1 < C_2 < C_3 < ... < C_{k-1} < C_k$ avec C_k le conditionnement de la matrice.
À l'aide de simulations, Belsley, Kuh et Welsch ont montré que :

$$\begin{cases} \text{Si} \quad 30 < C_i < 100 & \textbf{existence d'une colinéarité} \\ \text{Si} \quad C_i > 100 & \textbf{existence d'une forte colinéarité} \ . \end{cases}$$

Pour les données relatives au secteur de santé, le conditionnement est souvent supérieur à 1000. Ce test permet de mesurer le degrés de colinéarité et calcule le nombre de relations de colinéarité existantes.

2.3 Simulations

Le cadre théorique proposé est appuyé et validé par une étude de simulations. Nous étudions cinq méthodes de régression linéaires régularisées ; Ridge, lasso, adaptive-lasso,

elastic-net et la méthode proposée (Huber). L'analyse statistique est réalisée à l'aide du logiciel *R*.
Les données simulées consistent sur des ensembles de données indépendants de petite ($n > p$) et grande dimension ($p >> n$). Nous générons ces ensembles de données sous différentes hypothèses : un conditionnement élevé de la matrice de données (forte colinéarité) et une amplitude d'erreur variée.
Chaque ensemble de données est divisé en ensembles d'entrainement et de validation. Les cinq méthodes de régularisation sont utilisées pour ajuster les modèles optimaux dans chaque ensemble d'entrainement. Les modèles ajustés sont utilisés pour évaluer les performances des prédictions dans les ensembles de validation correspondants. Pour évaluer la précision, nous calculons l'erreur quadratique moyenne (MSE), l'erreur standard (sd), et extrayons le nombre de coefficients $\hat{\beta}$ non nuls. La procédure est répétée 50 fois pour chaque exemple.
Toutes les variables de X ont une distribution normale multivariée continue. Les prédicteurs sont générés par échantillonnage à partir d'une distribution normale multivariée avec la fonction de densité de probabilité suivante :

$$p_X(x) = (\frac{1}{2\pi})^{\frac{n}{2}}.\frac{1}{\sqrt{det(\Sigma)}}.exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu))$$

avec une moyenne μ et une covariance Σ. Pour tout x nous supposons que $\mu = 0$ et $Var(x) = 1$. Par une décomposition en valeur singulière de X (SVD), nous définissons le conditionnement de X. Nous attribuons à chaque prédicteur une valeur β prédéterminée. Dans tous les exemples, l'ensemble de données simulé est divisé en 3 partitions, deux pour l'ensemble d'apprentissage et une pour l'ensemble de validation (test).

Cas $n > p$:

• Dans l'exemple 1, nous avons simulé 50 ensembles de données, n=20 pour chaque ensemble, générés par $y = X\beta + \sigma\epsilon$ avec $\epsilon \sim N(0, 1)$.
Nous fixons $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, $\sigma = 3$ et $cond(X) = 100$ (forte collinearité).

• Dans l'exemple 2, nous avons simulé 50 ensembles de données, n=20 pour chaque ensemble, générés par $y = X\beta + \sigma\epsilon$ avec $\epsilon \sim N(0, 1)$.
Nous fixons $\beta = (0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80)$, $\sigma = 3$ et $cond(X) = 100$.

• Dans l'exemple 3, nous avons simulé 50 ensembles de données, n=100 et p=40 pour chaque ensemble, générés par $y = X\beta + \sigma\epsilon$ avec $\epsilon \sim N(0, 1)$.
Nous fixons $\beta = (0, ..., 0, 2, ..., 2, 0, ..., 0, 2, ..., 2)$, les 10 premiers termes de β sont fixés à 0, les 10 termes suivants fixés à 2, les 10 termes suivants fixés à 0, et les 10 derniers termes fixés à 2. $\sigma = 15$ et $cond(X) = 100$.

Cas $p >> n$:

Dans ce cas, comme suggéré par Zou [3], nous considérons la norme l_2 au lieu de l'estimateur des moindres carrés pour calculer les poids adaptifs de la méthode adaptive-lasso.
Pour le cas des données de grande dimension, nous avons simulé 30 ensembles de données indépendants générés par $y = X\beta + \sigma\epsilon$ avec $\epsilon \sim N(0, 1)$ pour chacun des exemples suivants :

• Dans l'exemple 4, nous considérons n=200, p=400 et nous fixons tous les coefficients de

β à 0.85, $\sigma = 3$ et $cond(X) = 100$.

• Dans l'exemple 5, nous considérons n=30 et p=60, $\beta = (3, ..., 3, 0, ..., 0, 9, ..., 9)$, les premiers 20 coefficients de β sont fixés à 3, les 20 suivants à 0 et les 20 derniers à 9, $\sigma = 9$ et $cond(X) = 100$.

• Dans l'exemple 6, nous considérons n=100, p=400, $\beta = (1.5, 2.5)$ les premiers 200 coefficients sont fixés à 1.5 et les 200 suivants à 2.5, $\sigma = 9$ et $cond(X) = 100$.

Chaque exemple est considéré séparément. Ridge, lasso, adaptive-lasso, elastic-net et la méthode proposée (Huber) sont ajustés simultanément au même jeu de données. Le paramètre de régularisation λ est défini dans l'intervalle $[10^{-1}, 10]$. La valeur optimale de λ est déterminée en parcourant cet intervalle dans le processus de minimisation.

Example 1 :

Method	Parameters	Average MSE (sd)	Average of non-zero coefficients
Ridge	$\alpha = 0$	20.938 (2.177)	All
lasso	$\alpha = 1$	20.742 (2.188)	7.12
Ad-lasso	$\gamma = 0.5$	21.773 (2.278)	7.26
	$\gamma = 1$	23.022 (2.60)	7.20
	$\gamma = 1.5$	23.693 (2.883)	7.00
Elastic-net	$\alpha = 0.25$	20.813 (2.18)	7.70
	$\alpha = 0.5$	20.72 (2.182)	7.72
	$\alpha = 0.75$	20.707 (2.187)	7.40
Huber	$\sigma = 1$	20.82 (2.203)	7.96
	$\sigma = 1.5$	20.872 (2.196)	7.90
	$\sigma = 1.8$	20.878 (2.19)	7.98

Example 2 :

Method	Parameters	Average MSE (sd)	Average of non-zero coefficients
Ridge	$\alpha = 0$	20.945 (2.218)	All
lasso	$\alpha = 1$	21.398 (2.202)	7.32
Ad-lasso	$\gamma = 0.5$	22.398 (2.313)	7.38
	$\gamma = 1$	23.59 (2.636)	7.28
	$\gamma = 1.5$	24.13 (2.87)	7.20
Elastic-net	$\alpha = 0.25$	21.03 (2.221)	7.80
	$\alpha = 0.5$	21.103 (2.221)	7.72
	$\alpha = 0.75$	21.202 (2.212)	7.62
Huber	$\sigma = 1$	20.93 (2.212)	7.98
	$\sigma = 1.5$	20.812 (2.209)	All
	$\sigma = 1.8$	20.810 (2.209)	All

Example 3 :

Method	Parameters	Average MSE (sd)	Average of non-zero coefficients
Ridge	$\alpha = 0$	526.30 (20.635)	All
lasso	$\alpha = 1$	553.94 (22.642)	38.34
Ad-lasso	$\gamma = 0.5$	576.62 (24.668)	38.26
	$\gamma = 1$	596.38 (26.973)	37.74
	$\gamma = 1.5$	602.42 (28.041)	37.18
Elastic-net	$\alpha = 0.25$	531.03 (20.919)	39.66
	$\alpha = 0.5$	536.51 (21.257)	39.28
	$\alpha = 0.75$	543.57 (21.746)	39.00
Huber	$\sigma = 1$	532.36 (21.095)	39.94
	$\sigma = 1.5$	520.6 (20.452)	39.29
	$\sigma = 1.8$	515.44 (20.211)	39.88

Example 4 :

Method	Average MSE(sd)	Average of non-zero coefficients
Ridge	1049.9 (15.06)	All
lasso	1613.59 (26.03)	129.66
Ad-lasso	1627.17 (24.48)	120.30
Elastic-net	1208.14 (17.11)	253.93
Huber	1050.82 (15.11)	396.66

Example 5 :

Method	Average MSE (sd)	Average of non-zero coefficients
Ridge	39185.26 (3356.50)	All
lasso	62093.76 (5021.50)	18.96
Ad-lasso	68142.08 (5149.02)	19.00
Elastic-net	39045.61 (3448.09)	51.16
Huber	52078.32 (3771.15)	59.80

Example 6 :

Method	Average MSE (sd)	Average of non-zero coefficients
Ridge	6015.75 (379.099)	All
lasso	8071.26 (445.008)	65.73
Ad-lasso	8497.87 (464.066)	65.16
Elastic-net	6178.55 (398.025)	242.00
Huber	6140 (368.709)	398.00

À partir des exemples 1-2-3, les simulations ont confirmé que le lasso et l'adaptive-lasso ont séléctionné un petit nombre de prédicteurs pertinents. Nous avons observé que la méthode proposée (Huber) est d'une précision prédictive élevée. Lorsque le nombre de prédicteurs est très grand, la précision de prédiction de la méthode proposée (Huber) est plus élevée que celle du ridge, lasso, adaptive-lasso et de l'elastic-net. Ces exemples ont confirmés, de manière significative, la supériorité de Huber dans le cas d'une forte colinéarité et une amplitude d'erreur élevée. Notons que la méthode proposée bénéficie de la propriété de parcimonie dans certains cas, et ce en ramenant certains coefficients à zéro grâce à la norme l_1. Nous avons observé que les deux méthodes, ridge et elastic-net, sont généralement plus performantes que le lasso. Quant au lasso, il est plus performant que l'adaptive-lasso. Cependant, l'adaptive-lasso a tendance à sélectionner plus de variables que le lasso. Enfin, dans le cas $n > p$, une précision prédictive élevée a été observée avec la méthode huber suivie de la méthode ridge. Finalement, l'elastic-net, le lasso et l'adaptative-lasso incorporent plus que Huber, la propriété de sélection de variables, et leur modèle résultant est mieux interprétable que celui de la méthode proposée.
À partir des exemples 4-5-6, qui représentent le cas de grande dimension, nous avons observé que la régression ridge et huber surpassent toutes les méthodes dans les exemples 4 et 6 en terme de précision de prédiction. L'elastic-net est plus performante que les autres méthodes dans l'exemple 5. Le lasso et l'adaptative-lasso ont tendance à sélectionner plus de variables.
Dans le cas de grande dimension, nous avons observé qu'il n'y a pas de procédure de régularisation qui surpasse statistiquement toutes les autres méthodes. Dans ce cas, différentes procédures peuvent être utilisées et adaptées dans la pratique pour divers contextes.

2.4 Application : Prédiction de la récupération fonctionnelle après un AVC

2.4.1 Données et approche de modélisation

Nous avons utilisé les données d'AVC publiées par Douiri et al [9] pour évaluer et comparer la méthode de régularisation proposée (Huber) par rapport à d'autres méthodes (ridge, lasso, adaptive-lasso, elastic-net). Les données recueillies incluent 495 patients du registre des accidents vasculaires cérébraux du sud de Londres (SLSR) basé sur la population entre août 2002 et octobre 2004. Tous les patients présentant des infractions cérébrales (code 163 de la CIM-10) ont été inclus.
La progression de la récupération fonctionnelle a été évaluée à l'aide de l'indice de Barthel

(BI) avec des scores allant de 0 à 20. L'IB a été mesuré aux semaines un, deux, trois, quatre, six, huit, 12, 26 et 52 après un AVC. Un certain nombre de prédicteurs candidats ont été pris en compte dans la sélection des variables du modèle, y compris les données démographiques (âge, sexe, origine ethnique, invalidité prémorbide, statut socio-économique), les caractéristiques de l'AVC (sous-type basé sur la classification d'Oxford (infarctus lacunaire (LACI), infarctus total de la circulation antérieure (TACI), infarctus partiels de la circulation antérieure (PACI), infarctus de la circulation postérieure (POCI) et accident vasculaire cérébral hémorragique intra-cérébral (ICH), présence de symptômes cérébelleux, déficiences de base, variables (case-mix) : (score de Glasgow (GCS) et le National Institutes of Health Stroke Scale (NIHSS)). Les variables potentielles de la récupération fonctionnelle ont été examinées pour leur aspect pratique en fonction de leur prévalence dans la littérature, ce qui a donné lieu à 7 facteurs pronostiques candidats. Les détails de ces prédicteurs figurent dans [9]. Veuillez noter que toutes les méthodes de régularisation décrites dans ce travail, sont robustes pour la colinéarité.
En ce qui concerne les données manquantes et en partant de l'hypothèse que celles-ci étaient aléatoires, nous avons utilisé une méthode d'imputation multiple connue sous le nom de Monte-Carlo par chaînes de Markov (MCMC) [120].
Sept prédicteurs (sexe (sexe), groupes ethniques (ethgrp), score de Glasgow (glas _cs), sous-type d'AVC (TACI, PACI, LACI, POCI), âge du patient (âge), score de l'indice de Barthel à l'admission (batotw1) et NIHSS Stroke Scale (nihtot)) on été utilisés pour prédire le score de Barthel à 12, 26 et 52 semaines. Les méthodes ridge, lasso, adaptive-lasso, elastic-net et huber ont été appliquées aux données d'AVC. L'ajustement du modèle et la sélection des paramètres de régularisation qui optimisent le problème de minimisation ont été effectués sur les données d'entraînement. Nous avons ensuite comparé les performances des méthodes étudiées en calculant leur erreur quadratique moyenne de prédiction. L'estimation des paramètres, l'erreur quadratique moyenne (MSE) et l'écart type (sd) ont été calculés par la méthode du bootstrap avec 500 réplications. Comme nous l'avons mentionné précédemment, le conditionnement lié à la matrice de covariance pour les données du (SLSR) a été calculé et est supérieur à 100, ce qui indique l'existence d'une forte colinéarité dans nos données.

Stroke data (12 weeks)

Methods					
Predictor	**Ridge (sd)**	**lasso (sd)**	**Ad-lasso (sd)**	**Elastic-net (sd)**	**Huber (sd)**
coefficients estimations					
Sex	-1.13 (0.023)	-1.18 (0.032)	-1.34 (0.039)	-1.15 (0.026)	-0.90 (0.018)
ethgrp 1	1.18 (0.024)	1.36 (0.043)	2.18 (0.082)	1.25 (0.031)	0.88 (0.019)
ethgrp 2	0.06 (0.023)	0.15 (0.022)	1.20 (0.074)	0.10 (0.019)	-0.10 (0.015)
ethgrp 3	-0.15 (0.019)	-0.03 (0.008)	-0.01 (0.118)	-0.09 (0.012)	-0.09 (0.011)
ethgrp 4	-0.25 (0.015)	0.00 (0.002)	-0.60 (0.128)	-0.12 (0.009)	-0.16 (0.008)
glas_cs	0.59 (0.005)	0.58 (0.005)	0.50 (0.005)	0.58 (0.005)	0.59 (0.005)
TACI	-1.61 (0.022)	-1.90 (0.037)	-2.47 (0.045)	-1.72 (0.027)	-1.30 (0.021)
PACI	1.18 (0.016)	1.22 (0.032)	1.09 (0.042)	1.21 (0.021)	0.89 (0.011)
POCI	0.09 (0.035)	0.03 (0.043)	-0.51 (0.082)	0.07 (0.035)	0.11 (0.021)
LACI	-0.47 (0.027)	-0.50 (0.035)	-1.23 (0.062)	-0.47 (0.028)	-0.26 (0.018)
age	-0.04 (0.001)	-0.04 (0.001)	-0.02 (0.001)	-0.04 (0.001)	-0.04 (0.001)
batotw1	0.63 (0.003)	0.63 (0.003)	0.62 (0.003)	0.63 (0.003)	0.64 (0.002)
nihtot	0.21 (0.004)	0.21 (0.004)	0.19 (0.004)	0.21 (0.004)	0.21 (0.004)
Average MSE (sd)	30.523 (0.280)	31.222 (0.291)	33.721 (0.324)	30.673 (0.283)	30.168 (0.279)

Stroke data (26 weeks)

Methods					
Predictor	**Ridge (sd)**	**lasso (sd)**	**Ad-lasso (sd)**	**Elastic-net (sd)**	**Huber (sd)**
coefficients estimations					
Sex	-0.31 (0.025)	-0.27 (0.026)	-0.16 (0.030)	-0.29 (0.025)	-0.25 (0.019)
ethgrp 1	1.08 (0.028)	1.20 (0.045)	2.59 (0.102)	1.11 (0.033)	0.78 (0.020)
ethgrp 2	0.14 (0.020)	0.14 (0.022)	1.62 (0.091)	0.12 (0.018)	-0.01 (0.013)
ethgrp 3	0.38 (0.012)	0.01 (0.005)	3.55 (0.154)	0.15 (0.008)	0.18 (0.005)
ethgrp 4	-0.28 (0.018)	-0.03 (0.008)	-0.23 (0.140)	-0.17 (0.012)	-0.18 (0.010)
glas_cs	0.74 (0.006)	0.72 (0.007)	0.61 (0.006)	0.73 (0.006)	0.75 (0.006)
TACI	-0.68 (0.022)	-0.49 (0.027)	-0.65 (0.035)	-0.60 (0.024)	-0.55 (0.016)
PACI	1.49 (0.018)	2.06 (0.039)	2.49 (0.051)	1.68 (0.025)	1.05 (0.014)
POCI	0.13 (0.030)	0.17 (0.036)	0.12 (0.062)	0.14 (0.030)	0.10 (0.019)
LACI	-0.24 (0.026)	-0.12 (0.026)	-0.37 (0.049)	-0.19 (0.025)	-0.14 (0.017)
age	-0.06 (0.001)	-0.06 (0.001)	-0.05 (0.001)	-0.06 (0.001)	-0.06 (0.001)
batotw1	0.61 (0.003)	0.61 (0.003)	0.59 (0.003)	0.61 (0.003)	0.61 (0.002)
nihtot	0.19 (0.004)	0.18 (0.004)	0.14 (0.004)	0.18 (0.004)	0.18 (0.004)
Average MSE (sd)	30.87 (0.272)	31.06 (0.278)	33.22 (0.314)	30.89 (0.273)	30.38 (0.266)

Stroke data (52 weeks)

Methods					
Predictor	**Ridge (sd)**	**lasso (sd)**	**Ad-lasso (sd)**	**Elastic-net (sd)**	**Huber (sd)**
coefficients estimations					
Sex	-0.48 (0.0272)	-0.48 (0.0301)	-0.54 (0.0372)	-0.47 (0.027)	-0.35 (0.020)
ethgrp 1	1.01 (0.025)	1.06 (0.0420)	1.52 (0.0825)	1.02 (0.030)	0.74 (0.018)
ethgrp 2	0.17 (0.024)	0.17 (0.0264)	0.94 (0.084)	0.15 (0.022)	0.01 (0.016)
ethgrp 3	0.14 (0.006)	0.00 (0.00)	0.69 (0.072)	0.00 (0.000)	0.06 (0.002)
ethgrp 4	-0.50 (0.017)	-0.06 (0.011)	-2.50 (0.125)	-0.37 (0.014)	-0.28 (0.009)
glas_cs	0.54 (0.007)	-0.52 (0.007)	0.43 (0.007)	0.53 (0.007)	0.55 (0.007)
TACI	-0.61 (0.022)	-0.37 (0.024)	-0.32 (0.029)	-0.52 (0.023)	-0.52 (0.016)
PACI	1.61 (0.017)	2.37 (0.038)	3.13 (0.047)	1.84 (0.024)	1.12 (0.014)
POCI	0.94 (0.034)	1.37 (0.064)	1.90 (0.088)	1.04 (0.042)	0.64 (0.024)
LACI	-0.53 (0.031)	-0.35 (0.037)	-0.44 (0.062)	-0.47 (0.032)	-0.37 (0.020)
age	-0.04 (0.001)	-0.04 (0.001)	-0.02 (0.001)	-0.04 (0.001)	-0.04 (0.001)
batotw1	0.67 (0.003)	0.67 (0.003)	0.66 (0.003)	0.67 (0.003)	0.67 (0.003)
nihtot	0.21 (0.004)	0.20 (0.004)	0.17 (0.005)	0.21 (0.004)	0.21 (0.004)
Average MSE (sd)	31.73 (0.281)	32.43 (0.294)	33.66 (0.322)	31.94 (0.283)	31.32 (0.276)

Les résultats de ces simulations ont confirmé que la méthode proposée (huber) est robuste en terme de précision de prédiction. La méthode Ridge est plus performante que l'elastic-net, le lasso et l'adaptive-lasso. Par rapport à la méthode proposée, la précision de prédiction du lasso et de l'adaptive-lasso est clairement affectée comme prévu, par le conditionnement élevé de la matrice de covariance associée aux données. Cependant, la méthode proposée (huber) est la mieux adaptée pour résoudre ce problème.

2.5 Discussion

Dans ce travail, nous avons proposé une approche problème inverse pour le modèle de régression. Cette approche nous a permis de caractériser une fonction de régularisation permettant d'optimiser au mieux ce modèle. La fonction de régularisation proposée est hybride, simple et efficace. Elle a été conçue en se basant sur la méthodologie statistique robuste. Le modèle de régression régularisé résultant, s'est démarqué par de bonnes performances par rapport aux méthodes ridge, lasso, adaptive-lasso et l'elastic-net.
À notre connaissance, c'est la première fois que l'approche problème inverse, combinée avec la méthodologie statistique robuste est utilisée pour résoudre un problème de régression régularisée.
En effectuant des simulations, nous avons testé et comparé la méthode de régularisation proposée (Huber) par rapport à d'autres méthodes de régularisation, sous différents scénarios : l'existence d'une forte colinéarité dans les données simulées et une amplitude d'erreur variée, et ce pour les deux cas $(n > p)$ et $(p >> n)$.
La méthode proposée a montré de bonnes performances et moins de biais par rapport aux autres méthodes. Cela confirme que le cadre du problème inverse couplé avec la méthodologie statistique robuste, a permis de caractériser une fonction de régularisation optimale et efficace. Ce fait a été approuvé par les simulations conduites dans cette étude.

En utilisant des données cliniques basées sur la population londonienne, reflétant l'application sur le monde réel, la méthode proposée a également démontrée une bonne performance. Les paramètres de régression construits par notre méthode, correspondent le mieux aux données observées, aboutissant ainsi à une erreur de prédiction minimale. La méthode proposée établie une valeur clinique ajoutée dans la pratique et confirme que notre approche est optimale.
En recherche clinique, la méthodologie développée pourrait être appliqué comme un outil pour évaluer les effets bénéfiques des interventions fondées sur des données probantes et des structures de soins. En tant qu'outil de recherche, cela pourrait être utilisé pour tester de nouvelles interventions ou identifier des échantillons enrichis, ce qui réduirait le besoin d'essais randomisés contrôlés coûteux et souvent peu pratiques. Cette stratégie d'enrichissement prédictive est importante pour la conception d'essais futurs car elle permet le recrutement des patients les plus appropriés, permettant ainsi l'utilisation d'une population d'étude plus petite.
Cette approche pourrait être étendue à la résolution des modèles à effets mixtes et à la réduction de dimension. Elle pourrait également être utilisée dans des problèmes d'apprentissage automatique statistique pour optimiser la fonction de perte [121].
La méthode proposée, bénéficie de tous les avantages qu'offre le lasso et l'elastic-net pour l'ajustement du modèle et la sélection de variables. Comme pour le lasso, la méthode proposée peut produire des coefficients qui sont numériquement très petits, de sorte que le prédicteur lié peut être ignoré.
Les résultats rapportés dans ce travail suggèrent que l'approche problème inverse pourrait être utile dans une grande variété de problèmes d'estimation statistique. Cette méthodologie n'est pas bien explorée dans la communauté statistique et une étude plus approfondie est cependant nécessaire. Le cadre des problèmes inverses pourrait également fournir un outil utile pour la recherche clinique.
Par ce travail, nous avons prouvé que l'approche problème inverse, couplé avec la méthodologie statistique robuste, a permis de caractériser une fonction de régularisation robuste et moins biaisée que d'autres méthodes existantes dans la littérature, et pourrait être utilisée avec assurance dans la pratique.
En outre, la méthode proposée offre une alternative améliorée à la méthode ridge, lasso, adaptive-lasso et l'elastic-net, qui se sont déjà révélées être des méthodes très performantes dans des études antérieures.

Chapitre 3

Prédiction du risque de déclin cognitif post-AVC

3.1 Introduction

L'accident vasculaire cérébral (AVC) est une affection de longue durée (ADL), courante, dont l'incidence augmente à mesure que la population vieillit. Les patients ayant subi un AVC ont une probabilité accrue de déficit cognitif par rapport à ceux qui n'en ont pas subi [122]. La probabilité d'une déficience cognitive reste élevée de manière persistante jusqu'à quinze ans après la survenue d'AVC et peut être associée à une validité plus élevée, une qualité de vie inférieure et une dépression. La conjonction de la baisse de mortalité après l'occurrence d'un AVC et du vieillissement de la population [123] laissent présager que la déficience cognitive post-AVC deviendra de plus en plus fréquente, d'autant plus que le risque d'AVC [124] et la déficience cognitive [125] augmente de façon exponentielle avec l'âge.
Des études ont suggéré que le déclin cognitif pourrait être prévisible après la survenue d'un AVC [99, 6]. Un suivi longitudinal associé à des modèles prédictifs centrés sur les patients peut être plus approprié pour produire des bilans sanitaires, plus précis dans l'optique de planifier d'une manière simultanée des soins immédiats, à moyen et à long terme pour des patients ayant des bilans physiques et psychologiques médiocres. Malgré la disponibilité des médicaments préventifs et des programmes de rééducation permettant de mieux contrôler les risques, un outil plus centré sur le patient contribuerait à une bonne gestion des soins des patients et, par conséquent, leur permettrait de vivre une vie plus normale.
Le Mini Mental Test de Folstein (ou Mini-Mental State Examination dans la version anglaise) connu sous le sigle MMSE, est le test le plus utilisé permettant aux cliniciens de dépister les démences [88]. Il est significativement corrélé au déclin cognitif après un AVC. Suziki et. al [126] ont utilisé les scores MMSE enregistrés au début des symptômes de l'AVC afin de prédire les scores MMSE au fil du temps (modèle 1 R2=67.6%, modèle 2 R2=59.8%). Cependant, il se peut que cette approche ne permette pas prédire avec précision la récupération des patients, vu que les scores MMSE enregistrés au début de l'AVC sont souvent beaucoup plus faibles comparés aux périodes suivantes. En effet, de nombreux patients connaissent une certaine amélioration après la phase aiguë.
Ross et al [127] ont utilisé des paramètres d'imagerie issus de la spectroscopie par résonance magnétique du proton pour prédire le déclin cognitif chez des patients jusqu'à trois ans après un AVC (R2=54.6%). De la même manière, Saini et al [128] ont utilisé des paramètres issus de tomodensitométries pour prédire le déclin cognitif trois à six mois après un AVC ischémique. La présence d'une atrophie significative et de lésions de substance

blanche était associée à un déclin cognitif avec un odds ratio de 3.07 et 3.13 respectivement. Bien que ces modèles puissent être un outil d'assistance utile pour prédire le déclin cognitif, l'atrophie n'est pas systématiquement mesurée dans la pratique. Tang et al [129] [130], ont rapporté que plusieurs modèles ont été utilisés pour prédire la démence ou les troubles cognitifs. Plusieurs variables, notamment les scores démographiques, les résultats des test cognitifs et les marqueurs de neuroimagerie ont été incorporées dans différents modèles avec une précision prédictive jugée entre moyenne et élevée. Deux modèles ont été développés pour prédire les troubles cognitifs post-AVC et leur précision prédictive s'est avérée acceptable [131, 132]. Cependant, leur utilité est difficile à soulever, car ils incluent tous des variables de neuroimagerie qui ne sont pas facilement accessibles.
Dans cette étude, nous développons et validons un modèle prédictif centré sur le patient pour estimer le risque de déclin cognitif jusqu'à 5 ans après un AVC ischémique. Nous évaluons également les écarts entre la récupération observée et prévue ainsi que les différences dans les tendances de récupération. D'un point de vue mathématique, ce travail développe la stratégie de modélisation basée sur les courbes de récupération [7, 8, 9]. Il évalue la robustesse de cette méthodologie et sa capacité à prédire une des conséquences importantes d'AVC, qu'est la déficience cognitive. Une autre contribution inédite des travaux présentés dans ce chapitre est le développement des courbes de récupération régularisées via le modèle linéaire à effets mixtes (LMM)[45]. Nous évaluons également la performance de ces modèles en utilisant à la fois des métriques traditionnelles (discrimination et étalonnage) et non traditionnelles (utilité clinique).
Les résultats de cette recherche, d'un point de vue pratique, seront utilisés pour fournir des informations pronostiques aux survivants d'un AVC et à leurs familles, dans le but de faciliter leur suivi à long terme et aider à l'élaboration de plans de soins et de gestion plus adaptés.

3.2 Cadre théorique

Le choix du modèle linéaire mixte (LMM) pour la construction des "courbes de récupération" est à l'origine du mot "courbes de récupération régularisées". En effet, le modèle linéaire mixte (LMM/GLMM) incarne la notion de régularisation via la composante des effets aléatoires. La régularisation peut également être considérée explicitement en rajoutant une fonction de régularisation.

3.2.1 Modèle linéaire mixte et lien avec la régularisation

Considérons un modèle linéaire mixte (LMM) dont la distribution de y est gaussienne. La distribution conditionnelle de y sachant l'effet aléatoire u est définie par :

$$\boldsymbol{y} \mid \boldsymbol{u} \sim \mathbf{N}(\boldsymbol{\eta}, \boldsymbol{R})$$

où $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zu}$ est le prédicteur linéaire.
La distribution marginal des effets aléatoires est définie par :

$$\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{G})$$

L'effet aléatoire peut être vu comme faisant partie des résidus et explique une partie de la distribution résiduelle :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon^*} \quad , \quad \boldsymbol{\varepsilon^*} = \boldsymbol{Zu} + \boldsymbol{\varepsilon}$$

En combinant ces résultats, la distribution conjointe des données s'écrit :

$$\boldsymbol{y} \sim \boldsymbol{N}(\boldsymbol{X\beta}, \boldsymbol{V}) \quad , \quad \boldsymbol{V} = \boldsymbol{R} + \boldsymbol{ZGZ'}$$

Où X et Z sont les matrices liant les observations y au vecteur d'effets fixes β et d'effets aléatoires u respectivement.
Les résidus ont une distribution $\varepsilon \sim N(0, R)$, et V est la matrice de variance-covariance.

Considérons R et G comme connus. Puisque les deux distributions de $y \mid u$ et u sont gaussiens, et donc leur produit l'est aussi, leur distribution conjointe peut être décrite en combinant le modèle conditionnel et marginal :

$$p(y, u) = p(y \mid u)p(u) = L(\beta, u) = \exp \ell(\beta, u) = \exp \ell(y \mid u)\ell(u)$$

Les vraisemblances des modèles conditionnels et marginaux sont :

$$l(y \mid u) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|R|^{1/2}} \exp\left[\frac{-1}{2R}(y-\eta)'(y-\eta)\right]$$

$$\ell(y \mid u) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|R| - \frac{1}{2}(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu)$$

$$l(u) = \frac{1}{(2\pi)^{u/2}} \frac{1}{|G|^{1/2}} \exp\left[\frac{-1}{2G}u'u\right]$$

$$\ell(u) = -\frac{u}{2}\log(2\pi) - \frac{1}{2}\log|G| - \frac{1}{2}u'G^{-1}u$$

En retenant que les termes avec β et u, et en supprimant les constantes, nous obtenons la log-vraisemblance conjointe :

$$\ell(y, u) = -\frac{1}{2}(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) - \frac{1}{2}u'G^{-1}u$$

Maximiser cette log-vraisemblance revient à minimiser :

$$\mathbf{LS_{pen}} = \boldsymbol{(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u.} \tag{3.1}$$

(3.1) est connue sous le nom du critère des moindres carrés pénalisés, où $u'G^{-1}u$ est une certaine pénalité. Nous pouvons désormais voir les modèles mixtes comme une version régularisée des modèles linéaires.

3.2.2 Modèle linéaire mixte génralisé et lien avec la régularisation

Les modèles linéaires mixtes généralisés (GLMM) sont également une version régularisée des modèles linéaires généralisés (GLM).
En étendant le modèle linéaire mixte, nous pouvons permettre à la variable réponse y d'être non gaussienne. Dans le cas où y a une densité qui appartient à la famille exponentielle, un modèle mixte linéaire généralisé (GLMM) doit être utilisé au lieu d'un modèle mixte linéaire. La différence entre ces deux modèles est similaire à la relation entre le modèle linéaire (LM) et le modèle linéaire généralisé (GLM).
Comme pour le (GLM), nous avons une fonction lien $g(\cdot)$ tel que :

$$\eta = g(\mu) \quad , \quad \mu = h(\eta)$$

Ce qui distingue un (GLMM) d'un (GLM) est sa définition des prédicteurs, qui est la même pour un (LMM) :

$$\eta = X\beta + Zu$$

Contrairement au LMM mais similaire au GLM, $y \mid u$, a une distribution issue de la famille exponentielle :

$$l(y \mid u) = \exp\left[\frac{y\theta - b(\theta)}{\phi}w + c(y, \phi, w)\right]$$

Notons que, $l(y \mid u)$ est également connue sous le nom de quasi-vraisemblance (QL).
$y \mid u$ appartient àune famille exponentielle avec une moyenne
mu, $u \sim N(0, G)$
Nous supposons d'abord que la valeur de G est connue. Comme pour les LMM, nous souhaitons maximiser la probabilité :

$$l(\beta, u) = p(y, u) = p(y \mid u)p(u)$$

simultanément par rapport à β et u. En introduisant le log :

$$\log(l(\beta, u)) = \ell(\beta, u) = \ell(y \mid u)\ell(u)$$

Nous avons :

$$\ell(y \mid u) = \mathrm{QL}(y \mid u) = \frac{y\theta - b(\theta)}{\phi}w = y\frac{1}{\phi}\theta w - \frac{1}{\phi}b(\theta)w$$

et :

$$\ell(u) = -\frac{u}{2}\log(2\pi) - \frac{1}{2}\log|G| - \frac{1}{2}u'G^{-1}u$$

Par conséquent, la log-vraisemblance conjointe pour un GLMM s'écrit :

$$\ell(\beta, u) = y\frac{1}{\phi}\theta w - \frac{1}{\phi}b(\theta)w - \frac{u}{2}\log(2\pi) - \frac{1}{2}\log|G| - \frac{1}{2}u'G^{-1}u$$

En supprimant tous les termes qui ne contiennent ni β, ni u ou ne sont que des constantes, nous obtenons :

$$\boldsymbol{\ell(\beta, u) = y\frac{1}{\phi}\theta w - \frac{1}{\phi}b(\theta)w - \frac{1}{2}u'G^{-1}u.} \tag{3.2}$$

La log-vraisemblance pour un GLMM est connu sous le nom de quasi-vraisemblance pénalisée (PQL), car elle est équivalente à la quasi-vraisemblance d'un GLM avec une certaine pénalité $\frac{1}{2}u'G^{-1}u$ [44]. Ainsi, le modèle linéaire mixte généralisé (GLMM) peut également être vu comme une version régularisée du modèle linéaire généralisé (GLM).

L'expression de la log-vraisemblance $l(\beta, u)$ pour un GLMM, ne peut être simplifiée davantage. Nous pouvons l'approximer par linéarisation ou par la méthode de quadrature de Gauss [44].
La régularisation peut être considérée explicitement pour un (LMM/GLMM) en rajoutant une fonction de régularisation $\Psi(\beta)$ à la fonction de vraisemblance, afin de renforcer la parcimonie de notre estimateur et améliorer la précision de prédiction [13].
Pour ce faire, nous considérons la fonction objective suivante :

$$\boldsymbol{Q_\lambda(\beta, u) = \mathrm{LS}_{\mathrm{pen}} + \lambda\, \Psi(\beta).} \tag{3.3}$$

Où $\lambda \geq 0$ un paramètre de régularisation.
La fonction $\Psi(\beta)$ est une fonction de régularisation qui peut être définie par :

$\Psi(\beta) = \sum_{j=1}^{p} |\beta_j|$ (lasso)
$\Psi(\beta) = \sum_{j=1}^{p} \beta_j^2$ (Ridge)
$\Psi(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$ (elastic-net)

$$\Psi_\sigma(|\beta|) = \begin{cases} \frac{1}{2}\beta^2 & Si \quad |\beta| \leq \sigma; \quad \sigma > 0 \\ \sigma(|\beta| - \frac{\sigma}{2}) & sinon. \end{cases} \text{ (Huber)}$$

L'objectif est d'estimer le paramètre à effet fixe β et à effet aléatoire u par :

$$(\hat{\beta}, \hat{u}) := \arg\min_{\beta,u} Q_\lambda(\beta, u)$$

3.3 Stratégie de modélisation

La stratégie de modélisation basée sur les courbes de récupération est utilisée pour développer les résultats de ce chapitre. Cette approche permet d'obtenir des trajectoires de récupération au fil du temps, rendant possible une prédiction longitudinale du déficit cognitif post-AVC. Des travaux antérieurs [7, 8, 9] ont considéré les modèles de régression linéaires pour construire les courbes de récupération dans le but de prédire la récupération fonctionnel post-AVC. Dans ce travail, nous utilisons les modèles à effets mixtes pour développer des "courbes de récupération régularisées" permettant de prédire le déficit cognitif jusqu'à cinq ans post-AVC.
Pourquoi choisir un modèle mixte pour prédire le déficit cognitif post-AVC ?
Les patients ayant un déficit cognitif post-AVC ont tendance à décliner rapidement avec des sauts de discontinuités (jump discontinuities). Ce phénomène est clairement observé chez les patients avec des problèmes vasculaire. Un modèle linéaire standard est incapable de capturer cette dynamique non linéaire.
De nombreux chercheurs ont proposé certaines méthodes pour traiter le problème de non linéarité. Par exemple, utiliser le diagramme Q-Q pour les résidus, considérer quelques valeurs comme aberrantes et les omettre ou les retirer de l'analyse. D'autres ont proposé d'appliquer une transformation de variable. Cependant, cette transformation n'est pas toujours applicable pour les variables catégoriques ou ordinales, et rend difficile l'interprétation des résultats. De plus, considérer que des valeurs sont aberrantes indique que votre modèle est mal ajusté aux données. Bien qu'il puisse y avoir des raisons spécifiques pour transformer les données ou d'omettre des valeurs aberrantes, la principale raison pour laquelle les chercheurs semblent adopter cette approche est l'obtention d'une distribution normale et d'appliquer des méthodes de régression standard.
Une application courante de la régression pour traiter les relations non linéaires implique la régression polynomiale. Dans ce cas, pour un prédicteur X, des termes sont ajoutés afin d'obtenir un meilleur ajustement (quadratique X^2, cubique X^3,...). Cependant, comme pour la régression linéaire, les méthodes de régression polynomiale ne sont pas suffisamment complexes pour saisir les nuances de la relation entre le déficit cognitif (observation) et les caractéristiques des patients (prédicteurs). Le modèle à effets mixtes est plus robuste pour détecter ces nuances, comme les sauts de discontinuités existants dans les données. En revanche, contrairement à la régression polynomiale, il est peu probable qu'un modèle mixte puisse trouver une relation quadratique entre les variables. Le cas échéant est souvent associé à un sur-ajustement. Appliquer une régression polynomiale s'apparente à imposer notre vision sur les données, plutôt que de les laisser parler d'elles-mêmes. Bien évidemment, cela peut être une bonne approximation dans quelques cas, tout comme supposer une relation linéaire, mais souvent c'est un voeux pieux.

Une façon de résoudre les problèmes associés à la régression polynomiale est de diviser les données en morceaux à divers points (noeuds) et d'ajuster une régression linéaire ou polynomiale dans ce sous-ensemble de données. Une telle approche est connue sous le nom de régression par morceaux. Bien qu'elle soit notablement mieux appropriée qu'un modèle polynomial, encore une fois, elle n'est pas satisfaisante pour prédire le déficit cognitif post-AVC. Dans la régression par morceaux, les ajustements sont discontinus, ce qui conduit à des prédictions parfois différentes pour des valeurs approchées.
Dans le cas de la déficience cognitive post-AVC, la régression par morceaux ne peut être adoptée vu que le changement est progressif et non brusque, contrairement à la croissance microbienne où cette approche est couramment utilisée. Un tel changement brusque ne peut être remarqué chez les survivants d'un AVC.
Les approches de type *black-box*, telles que les modèles de réseaux de neurones, les réseaux/modèles graphiques, les forêts aléatoires, les machines à vecteurs de support etc, trouvent de nos jours, un grand intérêt sous le nom "d'apprentissage automatique". Ces modèles sont bien adaptés aux données de grande dimension, en particulier lorsque la réduction de dimension est un problème. Cependant, ils sont inadaptés à la pratique clinique.
Venables et Ripley [133] affirment que le modèle à effets mixte pourrait être considéré comme une approche se situant entre les modèles de régression linéaire entièrement paramétriques hautement interprétables, et les techniques de black-box .
Dans le contexte de la prédiction des conséquences d'AVC, en l'occurrence le déficit cognitif, les modèles mixtes ont le pouvoir de prédire avec précision la trajectoire des courbes de récupération d'AVC. Un autre avantage de ce modèle est qu'il incarne la notion de régularisation. En effet, les modèles mixtes sont une version régularisée des modèles linéaires généralisés (GLM). Le terme d'effets aléatoires peut être vu comme un terme de régularisation, permettant ainsi d'obtenir une bonne précision de prédiction. Le modèle à effets mixtes est un outil conceptuellement simple. Il considère des relations non linéaires et aléatoires d'une manière explicite, et permet de se maintenir dans les cadres de modélisation linéaire et généralisée dont les cliniciens sont déjà familiers.

3.4 Méthodologie

3.4.1 Source des données

Les données utilisées pour cette analyse proviennent du South London Stroke Register (SLSR), un registre permanent basé sur la population qui enregistre de manière prospective les premiers accidents vasculaires cérébraux chez des patients de tous les groupes d'âge vivant dans une zone géographiquement définie du sud de Londres depuis 1995. Dans cette analyse, nous avons utilisé des données collectées entre 1995 et 2018.
Les méthodes du SLSR ont été décrites en détail par Wolfe et al. [6, 99] et sont résumées comme suit : tous les patients ayant eu un premier accident vasculaire cérébral après le 1er janvier 1995 et qui résidaient dans une zone définie de la région central du sud de Londres étaient éligibles pour l'inclusion. Selon le recensement de 2011, avec des changements annuels prévus, le nord de Southwark et Lambeth (n=357,308) comprend une population multiethnique avec une grande proportion de résidents noir issu de Caraïbes et d'Afrique (25.3%). L'AVC est défini selon la définition adopté par l'OMS [99]. L'exhaustivité et la qualité des données est estimée complète à 88% par un modèle logit-multinomial de capture-recapture [99].

3.4.2 Les participants

Les patients admis dans les hôpitaux desservant la zone d'étude (2 hôpitaux universitaires à l'intérieur de la zone d'étude et 3 hôpitaux à l'extérieur de ladite zone) ont été identifiés par une revue régulière des services d'urgences accueillant des patients victimes d'un AVC, les données nationales sur les patients admis dans n'importe quel hôpital en Angleterre et au Pays de Galles avec un diagnostic d'AVC sont examinées pour être considérées comme source additionnelle de patients. Tous les médecins généralistes (N=699(2011)) à l'intérieur et aux frontières de la zone d'étude sont régulièrement contactés et invités à informer le SLSR des patients victimes d'un AVC. L'orientation des patients non hospitalisés, victime d'AVC, vers une clinique externe neurovasculaire (depuis 2003) ou une visite à domicile aux patients par l'équipe d'étude est également possible pour les médecins généralistes. Les thérapeutes communautaires sont contactés tous les 3 mois. Les certificats de décès sont vérifiés régulièrement. Les patients sont évalués à l'admission, après 3 mois et une fois par an après la survenue de l'AVC.

3.4.3 Résultat et prédicteurs

Nous nous intéressons à la prédiction du déficit cognitif jusqu'à cinq ans après un AVC. Le déficit cognitif est mesuré par le score du Mini Mental Test de Folstein (MMSE) ou bien par le test mental abrégé (Abbreviated Mental Test dans la version anglaise, Connu sous le sigle AMT) [134]. Les patients ont été évalués au bout de sept jours, après 3 mois et une fois par an après la survenue de l'AVC. Avant le 1er janvier 2000, l'état cognitif était évalué avec le Mini-Mental State Examination ; après cette date, le test mental abrégé (AMT) a été adopté. Les sujets ont été définis comme ayant une déficience cognitive selon des seuils prédéfinis (MMSE < 24 ou AMT < 8). Il a été démontré que le (MMSE) et le (AMT) sont insensibles aux troubles cognitifs légers et à la fonction exécutive [135, 136]. L'AMT montre une bonne concordance avec le MMSE (c-statistic de 0.83 à 0.87) [137]. La méta-analyse (73 études) menée par Pendlebury and Rothwell [83] a été utilisée pour identifier une liste initiale de prédicteurs candidats pour le déclin cognitif post-AVC. Ces prédicteurs candidats ont ensuite été examinés pour leur aspect pratique en fonction de la disponibilité clinique, de la facilité de mesure, de la prévalence dans la littérature universitaire et la concertation avec des experts (médecin traitant, statisticien et épidémiologiste). Cela a donné une liste initiale de 93 prédicteurs candidats disponibles dans le SLSR. Les données sont collectées par les agents de terrain du SLSR non impliqués dans cette étude et ce à l'admission, après 3 mois, un an et chaque année par la suite.

3.4.4 Données manquantes

La méthode de Monte Carlo par chaîne de Markov a été utilisée pour imputer les valeurs manquantes, sous l'hypothèse qu'elles sont manquantes au hasard, afin de réduire le biais et d'éviter d'exclure les participants de l'analyse [138].

3.4.5 Méthodologie et analyses statistique

3.4.5.1 Sélection de variables

L'algorithme des forêts d'arbres décisionnels [139] a été utilisée pour classer les prédicteurs possibles par ordre d'importance. Les prédicteurs considérés comme ayant une grande pertinence clinique ont été réinjecté dans le modèle. Des interactions cliniquement

significatives ont été incluses dans le modèle. Leur pertinence a été testée en groupe pour éviter d'accroître l'erreur de type I. Tous les termes d'interaction ont été supprimés en tant que groupe et le modèle a été réajusté quand les résultats n'étaient pas significatifs. Les interactions avec le temps ont également été examinées. Un modèle à effets mixte [23] a été ensuite adapté pour développer les "courbes de récupération régularisées" du déclin cognitif chez un patient présentant certains facteurs pronostiques spécifiques.

3.4.5.2 Mesures des performances du modèle

Nous avons évalué la validation interne avec la méthode de validation croisée pour une estimation de la performance du modèle de prédiction. Les mesures de performance comprennent l'aire sous la courbe (AUROC/AUC), la sensibilité et la spécificité, la calibration, le score de Brier et l'analyse de la courbe de décision (DCA) [77, 78].
La discrimination d'un modèle de prédiction clinique (AUROC/AUC), fait référence à la capacité du score de risque à différencier les patients cognitivement intacts et cognitivement déficients.
La DCA a été réalisée afin d'évaluer davantage l'utilité clinique des courbes de récupération régularisées dans le pronostic des troubles cognitifs à trois mois, un an et cinq ans post-AVC. L'analyse statistique a été réalisée à l'aide du logiciel R.

3.4.5.3 Courbes de récupération régularisées

Nous avons tracé des courbes de récupération régularisées pour examiner visuellement différents sous-groupes à risque bien définis. Les tendances moyennes prévues ont été analysées selon l'âge, le sous-type d'AVC, le score de Glasgow et la latéralisation de l'AVC (AVC à l'hémisphère gauche). Pour évaluer l'efficacité pronostique et l'utilité clinique des courbes de récupération, la déficience cognitive a été dichotomisée en utilisant une déficience cognitive légère (Seuil : 24/30 MMSE et 8/10 AMT) et une déficience cognitive sévère (Seuil : MMSE et 4/10 AMT) [140, 141, 142]. L'utilité clinique a été évaluée à trois mois, un an et cinq ans post-AVC.

3.4.5.4 Développement et validation du modèle

Un modèle à effets mixtes a été développé et validé. La méthode de validation croisée a été utilisée pour sélectionner le modèle avec les meilleurs paramètres. Une validation croisée interne a été utilisée pour évaluer les performances du modèle développé.
Le R2 et l'erreur quadratique moyenne (RMSE) ont été considérés pour estimer l'erreur de prédiction.
L'âge du patient, le sexe, le groupe ethnique, le score cognitif à l'admission du patient, le score de Barthel à l'admission, le score de Glasgow (GCS), le sous-type d'AVC (LACI, PACI, POCI, TACI) [143], le diabète, la latéralisation de l'AVC (AVC à l'hémisphère gauche), la dysphasie et les interactions entre les variables prédictives et le temps (années) ont été identifiés comme de bons prédicteurs indépendants.

3.4.5.5 Éthiques

Les patients, ou pour les patients ayant des problèmes de communication, leurs proches, ont donné leur consentement éclairé écrit pour participer à des études sur l'AVC dans le cadre du SLSR. La conception a été approuvée par les comités d'éthique du Guy's and St Thomas 'NHS Foundation Trust, du Kings College Hospital, du Queens Square et des hôpitaux Westminster (Londres).

3.5 Résultats

3.5.1 Caractéristiques des participants

Un total de 6504 patients avec leur premier AVC entre 1995 et 2018 ont été enregistrés dans le SLSR. Dont n = 3411 patients avaient une fonction cognitive mesurée à l'admission, parmi eux n = 1204 avaient une déficience cognitive. Au total, n = 1608 ont terminé une entrevue de suivi en une année et n = 846 ont terminé une entrevue de suivi en cinq ans. Au total, n = 2171 personnes sont décédées dans les trois mois post-AVC. Un total de n = 2000 personnes n'avaient pas de fonction cognitive mesurée à l'admission post-AVC, pour des raisons médicales. Lors de l'admission post-AVC, les raisons médicales étaient des troubles de communication n = 992 et le coma n = 737.
Le nombre restant était dû à un enregistrement tardif ou parce que le patient ne s'est pas présenté au contrôle médical. La cohorte de développement comprenait 2 468 participants de (1995-2010) et la cohorte de validation comprenait 940 participants de (2011-2018). Le tableau 1 résume les caractéristiques des patients dans les cohortes de développement et de validation. Les données manquantes représentent moins de 15% des données. Les principales caractéristiques sont réparties uniformément entre les deux cohortes.

	Development cohort (1995-2010)		**Validation cohort (2011-2018)**	
Cognitive Impairment	**Intact (%)** 1468	**Impaired (%)** 1000	**Intact (%)** 736	**Impaired (%)** 204
Age, mean (SD)	66.84 (14.61)	74.10 (12.90)	69.60 (15.40)	70.50 (15.31)
Sex				
female	824 (56.13%)	540 (54%)	279 (38%)	98 (48.04%)
Male	644 (43.87%)	460 (46%)	457 (62%)	106 (51.96%)
Ethnicity				
White	1044 (71.12%)	732 (73.2%)	412 (56%)	95 (46.57%)
Black	346 (23.57%)	219 (21.9%)	278 (38%)	95 (6.37%)
Other	63 (4.29 %)	44 (4.4%)	43 (0.6%)	13 (6.37%)
Missing	15 (1.02%)	5 (0.5%)	3 (0.41%)	1 (0.5%)
Socioeconomic group				
Manual	831 (56.61%)	621 (62.1%)	218 (29.62%)	59 (28.92%)
Non-manual	533 (36.31%)	207 (20.7%)	222 (30.16%)	47 (23.04%)
Unknown	2 (0.14%)	2 (0.2%)	1 (0.14%)	0 (0%)
Missing	102 (6.95%)	170 (17%)	295 (40.08%)	98 (48.04%)

Pre-stroke vascular risk factors				
Transient ischemic attack				
No	1285 (87.53%)	858 (85.8%)	666 (90.5%)	179 (87.75%)
Yes	173 (11.78%)	129 (12.9%)	57 (7.74%)	22 (10.78%)
Missing	10 (0.68%)	13 (0.13%)	13 (1.77%)	3 (1.47%)
Atrial fibrillation				
No	1310 (89.24%)	787 (78.7%)	602 (81.79%)	156 (76.47%)
Yes	148 (10.08%)	198 (19.8%)	115 (15.63%)	44 (21.57%)
Missing	10 (0.68%)	15 (0.15%)	19 (2.58%)	4 (1.96%)
Hypertension				
No	517 (35.22 %)	301 (30.1%)	245 (33.29%)	62 (30.39%)
Yes	944 (64.31 %)	691 (69.1%)	484 (65.76%)	140 (68.63%)
Missing	7 (0.48%)	8 (0.8%)	7 (0.95%)	2 (1%)
Diabetes mellitus				
No	1188 (80.93%)	775 (77.5%)	540 (73.37%)	133 (65.20%)
Yes	271 (18.46%)	217 (21.7%)	189 (25.68%)	67 (32.84%)
Missing	9 (0.61%)	8 (0.8%)	7 (0.95%)	4 (2%)
Hypercholesterolemia				
No	895 (60.97%)	512 (51.2%)	428 (58,15%)	113 (55.4%)
Yes	310 (21.12%)	163 (16.3%)	298 (40.5%)	86 (42.16%)
Unknow	263 (17.92%)	325 (32.5%)	10 (1.36%)	5 (2.45%)
Current Smoker				
No	491 (33.45%)	365 (36.5%)	306 (41.60%)	90 (44.12%)
Yes	481 (32.77%)	308 (30.8%)	246 (33.42%)	59 (28.92%)
Unknown	473 (32.22%)	270 (27%)	180 (24.46%)	48 (23.53%)
Missing	23 (1.57%)	57 (0.57%)	4 (0.54%)	7 (3.43%)
Stroke severity (Case-mix)				
Glasgow coma scale (GCS)				
Severe (<8)	21 (1.43%)	69 (0.69%)	6 (0.82%)	8 (3.92%)
Moderate (9-12)	38 (2.59%)	173 (17.3%)	31 (4.21%)	25 (12.25%)
Mild (13-15)	1375 (93.66%)	739 (73.9%)	678 (92.12%)	162 (79.41%)
Missing	34 (2.32%)	19 (0.19%)	21 (2.85%)	9 (4.41%)
Urinary incontinence				
No	1180 (80.38%)	424 (42.4%)	633 (86%)	130 (63.73%)
Yes	251 (17.10%)	555 (55.5%)	82 (11.14%)	67 (32.84%)
Missing	37 (2.52%)	21 (0.21%)	21 (2.85%)	7 (3.43%)
Stroke subtype				
Infarct	1286 (87.60%)	817 (81.7%)	629 (85.46%)	179 (87.75%)
Haemorrhagic	169 (11.51%)	161 (16.1%)	106 (14.40%)	25 (12.25%)
Missing	13 (0.89%)	22 (0.22%)	1 (0.14%)	0 (0%)

Tableau 1 : Caractéristiques des patients à l'admission post-AVC : socio-démographiques, antécédents médicaux, sévérité et sous-types d'AVC

3.5.2 Performance du modèle

Les courbes de récupération régularisées prédictive ont montré un bon ajustement et une bonne prédiction. Dans la validation croisée interne, l'erreur prédictive RMSE est de 0.12 et R2 est de 73%. Le score cognitif moyen est caractérisé par une amélioration initiale au cours des 3 premiers mois, puis une baisse progressive par la suite.
La figure 1 présente la moyenne des courbes prédictives du modèle et la moyenne observée du score cognitif jusqu'à 5 ans après un AVC.

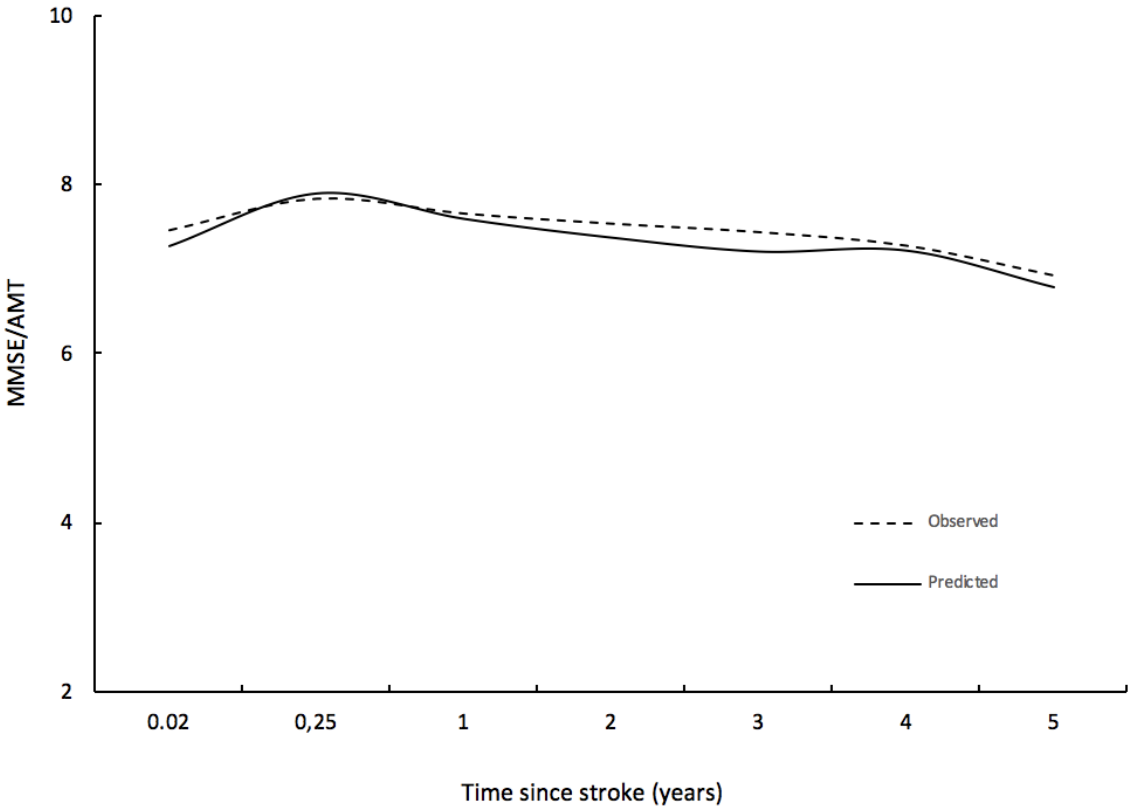

FIGURE 3.1 – La moyenne des courbes de récupération (modèle) Vs la moyenne observée du score cognitif jusqu'à 5 ans après un AVC

Les courbes prédictives montrent des similitudes entre les sous types d'AVC LACI et POCI au départ, mais une grande différence un an plus tard, le LACI ayant la plus forte baisse par rapport au POCI. Des différences sont observées entre les sous types d'AVC TACI et PACI au départ, mais sont comparables après 3 ans. Nous avons montré que le score cognitif varie selon les groupes d'âge. Nous avons observé une phase d'amélioration la première année chez les jeunes patients, mais une baisse significative chez les survivants d'AVC plus âgés, et ce, jusqu'à 5 ans post-AVC. Malgré la phase d'amélioration remarquée chez les patients plus jeunes, nous nous attendons à une légère baisse du score cognitif après 1 an. Un AVC sévère (c'est à dire score de Glasgow (GCS) modérée à sévère, ou un AVC dans l'hémisphère gauche) a montré une association significative avec le déclin cognitif.
La figure (3.2) présente la moyenne du score cognitif post AVC stratifié par groupe d'âge, sous-type d'AVC et GCS.

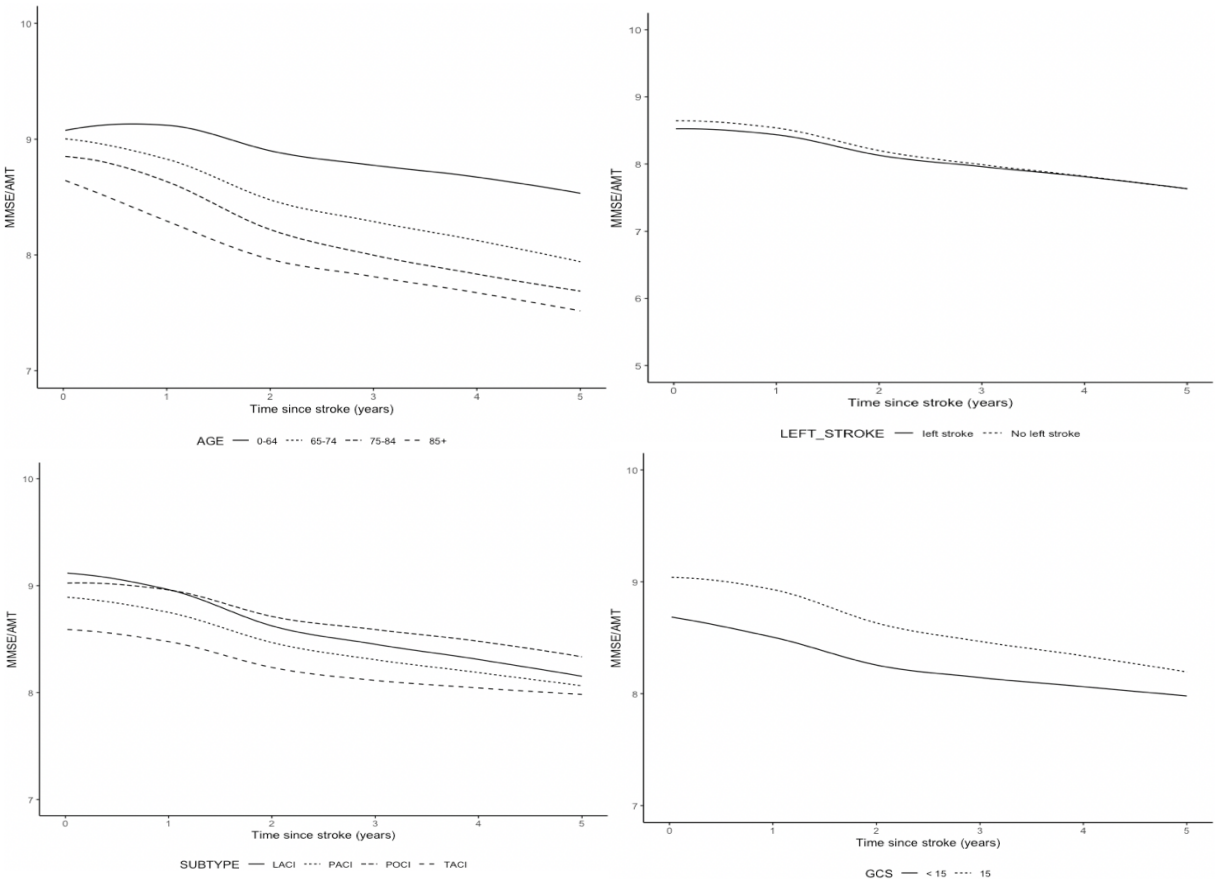

FIGURE 3.2 – la moyenne du score cognitif post AVC stratifié par groupe d'âge, sous-type d'AVC et GCS : score de Glasgow

Le modèle a en outre été évalué pour identifier la précision pronostique, la sensibilité, la spécificité et l'utilité du modèle à différents seuils du déclin cognitif permettant ainsi, la discrimination entre une déficience cognitive grave et légère à 3 mois, 1 an et 5 ans après un AVC. Le modèle a une bonne validité. En effet, pour un déficit cognitif sévère à 3 mois (la sensibilité est de 52 à 71% et la spécificité est de 91 à 94%) pour un déficit cognitif léger à 3 mois (la sensibilité est de 73 à 82% et la spécificité est de 68 à 75%). Le modèle a également montré une utilité clinique potentielle. En effet, les valeurs prédictives négatives étaient de (96%, IC 95% [94-97]), (96%, IC 95% [94- 97]), (97%, IC 95% [96-98]) pour un déficit cognitif sévère à 3 mois, 1 an et 5 ans respectivement. Le tableau 2 résume les valeurs prédictives et les taux de vraisemblance (likelihood ratio) permettant de classer chaque score de déficit cognitif d'interêt.

Measure	3 months	1 year	5years
(cut-off=4)			
Prevalence	10% [8- 12]	9% [7- 11]	6% [4-7]
Overall prognostic performance			
Overall performance (Brier)	7%	7%	8%
Discrimination (AUC)	88.5% [85-90]	89.6% [86-92]	87% [85-91]
Prognostic performance at a cut-off			
Sensitivity	62% [52-71]	58% [48-68]	59% [46-71]
Specificity	93% [91- 94]	92% [90-94]	90% [88-92]
Clinical utility at cut-off			
PPV	49% [41- 58]	42% [33-50]	27% [20- 35]
NPV	96% [94-97]	96% [94- 97]	97% [96 -98]
LR+	8.75 [6.68 -11.46]	7.29 [5.57-9.54]	6.10 [4.61-8.05]
LR-	0.41 [0.32- 0.52]	0.45 [0.36- 0.57]	0.46 [0.34- 0.61]
(cut-off=8)			
Prevalence	32% [29-35]	39% [36-42]	42% [39- 45]
Overall prognostic performance			
Overall performance (Brier)	17%	19%	20%
Discrimination (AUC)	80% [76-81]	77% [73-78]	75% [72-78]
Prognostic performance at a cut-off			
Sensitivity	78% [73-82]	74% [70-78]	72% [68-76]
Specificity	72% [68-75]	65% [61-68]	65% [61-69]
Clinical utility at cut-off			
PPV	56% [52-61]	58% [53 - 62]	60% [55-64]
NPV	87% [85-90]	79% [76 - 83]	76% [73- 80]
LR+	2.75 [2.42 - 3.12]	2.10 [1.87 - 2.36]	2.05 [1.82 - 2.31]
LR-	0.31 [0.25 - 0.37]	0.40 [0.34 - 0.47]	0.43 [0.37- 0.51]

AUC: area under the curve; LR: likelihood ratio; NPV: negative predictive value; PPV: positive predictive value.

Tableau 2 : Valeurs prédictives et taux de vraisemblance (likelihood ratio) permettant de classer les scores de déficit cognitif.

Le bénéfice-net exprimé en fonction des seuils de probabilités du déficit cognitif à 3 mois, 1 an et 5 ans est illustré dans la figure (3.3). La ligne grise est tracée pour refléter la stratégie qui suppose que tous les patients souffrent de troubles cognitifs (c'est-à-dire recommander une intervention pour tous les patients), et la ligne noire a été tracée pour refléter la stratégie qui suppose que tous les patients ne sont pas atteints de troubles cognitifs (c'est-à-dire ne recommander aucune intervention). Le bénéfice est maximisé par la courbe de déclin cognitif du modèle prédictif (ligne rouge) avec des seuils de probabilités de 15 à 80% à 3 mois, 15 à 79% à 1 an et 15 à 82% à 5 ans.
Pour les seuils plus élevés ($> 80\%$ à 3 mois, $> 79\%$ à 1 an et $> 82\%$ à 5 ans) l'option de ne pas intervenir est privilégiée.

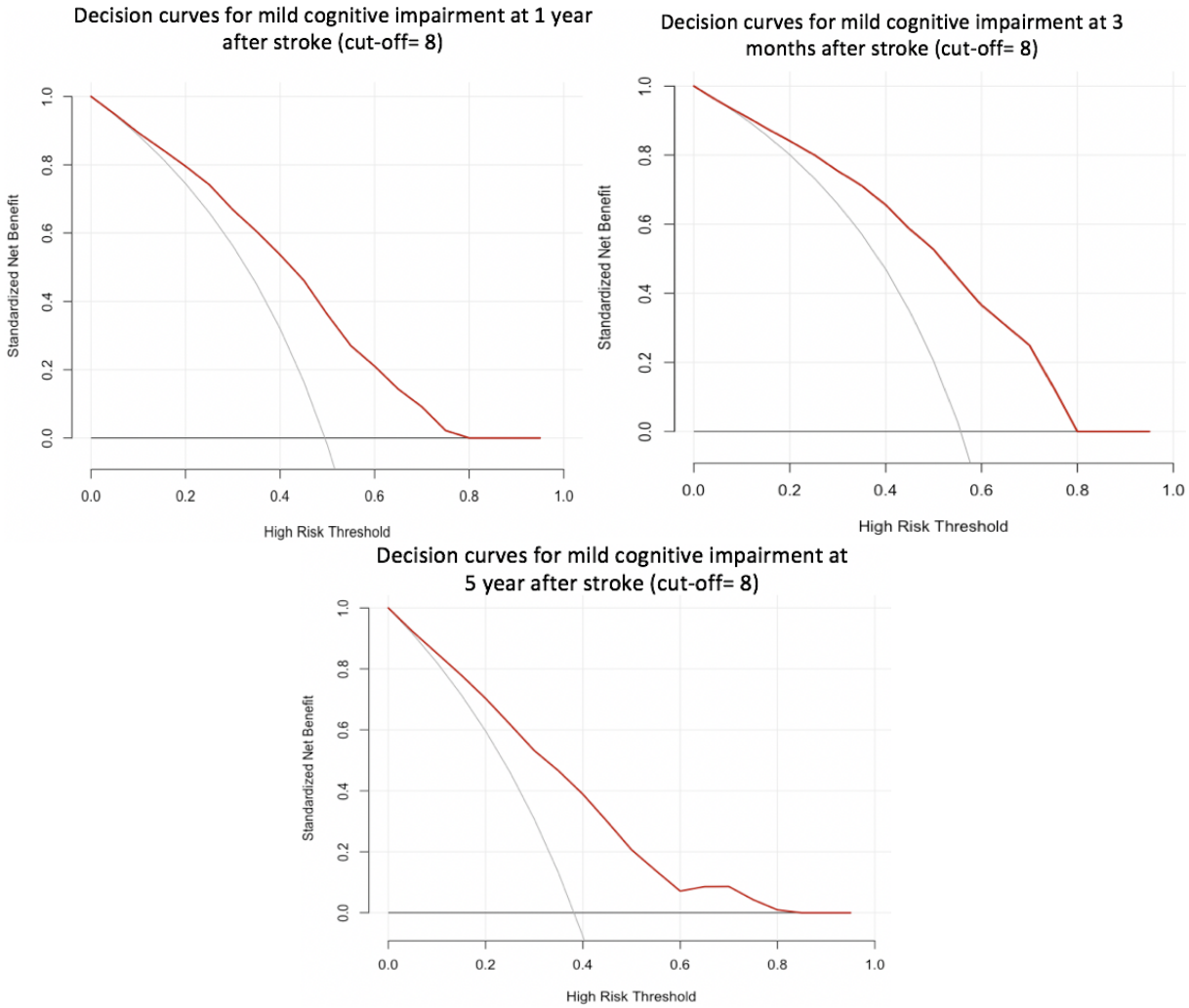

FIGURE 3.3 – Courbes de décision pour prédire un déficit cognitif léger chez les survivants d'un AVC à trois mois, 1 et 5 ans. Ligne rouge : modèle de prédiction. Ligne grise : suppose que tous les patients sont atteints de déficit cognitif. Ligne noire : suppose que tous les patients n'ont pas un déficit cognitif.

3.6 Discussion

Dans cette étude, nous avons développé et validé un outil pronostic permettant de prédire à long terme (5 ans), le déclin cognitif post- AVC dans une cohorte basée sur une population. Le modèle proposé est centré sur le patient et permet de prédire la déficience cognitive à l'aide d'un score continu. Il a également permis de prédire avec précision, les

trajectoires des courbes de récupérations régularisées jusqu'à 5 ans après un AVC. Plusieurs modèles ont été utilisés pour prédire le déficit cognitif [126, 127, 128, 129, 131, 132]. Cependant, la plupart de ces modèles ont une période prédictive relativement courte, et n'évalue pas le risque de déclin cognitif sur de longues périodes, en particulier chez les jeunes victimes d'AVC. De plus, ces modèles prédisent le risque de déficience cognitive uniquement à des moments prédéfinis.
À des moments prédéfinis, la précision du modèle proposé s'est avérée supérieure aux autre modèles existants dans la littérature.
La tendance générale du déclin cognitif post-AVC a été discuté et illustré au niveau de la population dans des études antérieures [6, 9]. Ceci s'est avéré utile pour la rééducation précoce et la planification de la sortie du patient, en prédisant si un patient est susceptible d'être dépendant, a besoin d'aide ou est indépendant, à un certain moment après l'AVC. Les facteurs influençant la récupération sont la latéralité de l'AVC et une diminution de la conscience à l'admission. Les patients présentant des lésions cérébrales du côté droit ont obtenu de meilleurs résultats que ceux présentant des lésions cérébrales du côté gauche, et ont montré une amélioration plus importante du score cognitif au fil du temps. La progression des troubles cognitifs chez les patients ayant une conscience réduite à l'admission est faible comparée aux patients sans diminution de la conscience au fil du temps. La réadaptation spécialisée après un AVC peut être bénéfique pour tous les âges, mais importante pour les plus de 65 ans. Elle confirme également que les patients plus âgés ont besoin d'une rééducation plus longue et sont moins susceptibles d'être libérés plus tôt de l'hôpital. Outre la vieillesse, des facteurs tels que la sévérité de l'apparition de l'AVC doivent également être pris en compte lors de la planification des interventions et de la rééducation après un AVC. Nous avons montré qu'en utilisant un modèle prédictif multivarié centré sur le patient, nous pouvons créer des profils de récupération individuels et classer avec précision les risques futurs de déclin cognitif. Le modèle que nous proposons, fait des prédictions sur des observations continues au lieu de se limiter aux abstractions binaires. De plus, les prévisions ne sont pas limitées par des périodes précises. Les paramètres finaux du modèle ont été sélectionnés à l'aide de la validation croisée. Cela signifie que ces coefficients reflètent la moyenne de plusieurs modèles construits sur des permutations aléatoires de sous-populations. Ainsi, les paramètres finaux du modèle reflètent des associations réelles et ne sont pas soumis à un sur-ajustement. De plus, la taille de l'échantillon est importante par rapport au nombre de variables pronostiques, ce qui augmente la puissance de l'étude. Les variables incorporées dans le modèle ont été sélectionnées pour leur association significative avec le déclin cognitif après un AVC, et testées à l'aide de plusieurs méthodes robustes, assurant ainsi une confiance totale dans les capacités prédictives des variables. De plus, ces variables sont régulièrement recueillies en phase aiguë de l'AVC et les évaluations de suivi, ce qui augmente la facilité d'utilisation du modèle.
Les prévisions de courbes de récupération régularisées centrées sur le patient, permettent de mieux comprendre le processus de récupération neurologique après un AVC. Cette information pronostique est importante pour les cliniciens et les survivants d'un AVC.
Dans la recherche clinique, cela pourrait également être appliqué pour aider à évaluer les effets bénéfiques des interventions et des milieux de soins fondés sur des données probantes. En tant qu'outil de recherche, cela pourrait être utilisé pour tester de nouvelles interventions ou pour identifier des échantillons enrichis, réduisant ainsi le besoin d'essais contrôlés randomisés coûteux et souvent peu pratiques. Cette stratégie d'enrichissement prédictif est importante pour la conception d'essais futurs, car elle permet le recrutement des patients les plus appropriés, permettant ainsi l'utilisation d'une population d'étude plus petite. Une autre application potentielle pourrait être de dériver un ensemble de

pondérations préliminaires des coûts sur l'utilisation des ressources, ce qui aiderait les commissaires à créer des modèles de financement pour les soins personnalisés des patients.
L'un des principaux atouts de la présente étude est que le modèle a été construit à l'aide d'une cohorte prospective non sélectionnée basée sur une population avec un premier AVC. Cela est préférable aux populations hospitalières, qui peuvent aboutir à des modèles casemix, ou à des modèles utilisant des données agrégées d'essais cliniques, qui représentent généralement des populations fortement sélectionnées et donc non-représentatives. Notre échantillon de données reflète réellement la population géographique d'intérêt et est donc optimal pour construire un modèle représentatif.
Une gestion appropriée du risque vasculaire est associée à une réduction à long terme du risque de déficit cognitif. Il faut donc soutenir la pharmacothérapie préventive optimale des facteurs de risque vasculaires et leur prise en charge [92]. Ce modèle peut potentiellement aider les cliniciens à organiser un programme de soins pour les patients après un AVC, qui est adapté à leur évolution cognitive prévue. Elle peut également aider à communiquer le risque aux patients et à leur famille de manière simple et claire, notamment grâce à l'utilisation des représentations graphiques du score cognitif en fonction du temps.
Nonobstant les points forts, les limites suivantes de cette étude doivent être prises en compte. Premièrement, l'étude pourrait être encore améliorée si le modèle est validé dans une population totalement indépendante, de préférence d'un autre pays et par des chercheurs indépendants. Deuxièmement, nous avons utilisé un modèle à effets mixtes régularisé. Cette stratégie de régularisation peut conduire à une sous-estimation des effets des prédicteurs dans l'échantillon de développement, mais permet d'augmenter la probabilité de réplication dans les études de validation.
Troisièmement, une étude d'impact doit être menée dans le cadre d'un essai contrôlé randomisé (ECR) pour confirmer si la capacité de prédire le rétablissement et l'intervention qui en résulte, pourrait avoir un impact sur le patient. Quatrièmement, un léger déclin cognitif est mesuré par la fonction exécutive. Cette dernière n'est pas mesurée avec les scores MMSE et AMT. La précision du modèle pourrait être augmentée en utilisant des mesures qui tiennent compte de la fonction exécutive, comme l'outil d'évaluation cognitive de Montréal (MoCA)[144]. MoCA est pratique et fiable, cependant, l'examen des tâches exécutives et linguistiques complexes est limité par rapport à l'évaluation en face-à-face. Les scores MMSE sont les plus utilisés par les cliniciens dans leur pratique quotidienne.
Enfin, la prédiction des troubles cognitifs sans thérapie efficace à portée de main soulève des préoccupations éthiques. Il est peu probable que de tels modèles soient déployés dans la pratique clinique avant qu'une validation et une évaluation supplémentaires ne soient entreprises.

3.7 Implications

Les implications de la recherche pour cette étude résident dans une meilleure compréhension des modèles de déclin cognitif post-AVC centrés sur les patients. Le modèle peut prédire le risque à long terme jusqu'à 23 ans après un AVC, ce qui, à la connaissance des auteurs, n'a pas été réalisé pour le déclin cognitif chez les patients victimes d'un AVC. Le modèle peut être utilisé comme outil de recherche pour tester l'influence de nouvelles interventions et de nouveaux médicaments sur le déclin cognitif. Les prévisions peuvent être modifiées à la lumière de la récupération observée, affinant ainsi le modèle et permettant des prédictions plus précises. Le seuil utilisé par le modèle peut être ajusté pour mettre l'accent sur la détection des troubles cognitifs légers ou sévères, permettant ainsi

une flexibilité en fonction des caractéristiques du patient. Le modèle peut détecter les personnes à risque plus élevé d'avoir un déficit cognitif, comme le démontrent nos analyses de sous-groupes. Cela permettra d'identifier les patients qui bénéficieront du traitement, et ainsi améliorer la rentabilité des soins de l'AVC. Les implications de cette étude sont vastes, fournissant un moyen d'organiser efficacement les soins post-AVC en déterminant les groupes à risque plus élevé, en communiquant le risque aux patients et à leurs familles et en prédisant, à des fins de recherche, les résultats des traitements médicamenteux sur le déficit cognitif.

3.8 Conclusion

La classification du risque pronostique basée sur des modèles prédictifs peut être cliniquement utile. Cet outil rend disponible des informations pronostiques qui pourraient soutenir le développement d'une prise en charge plus adaptée et aider à la mise en oeuvre de modèles de soins plus raffinés. L'outil des courbes de récupération régularisées pourrait fournir un cadre utile pour la pratique clinique et la santé publique.

Chapitre 4

Séries temporelles structurelles : de la formulation espace-état à la représentation en modèle à effets mixtes

Introduction

Les modèles espace-état interviennent généralement dans l'analyse des séries chronologiques [48]. Les séries chronologiques présentent des fluctuations tendancielles, saisonnières, cycliques, et peuvent être bien décrites par des modèles espace-état [49]. Elles sont utilisées dans une grande variété de domaines tels que la médecine, la biologie, la finance, le traitement du signal, etc. Leur principal objectif demeure la prédiction ou l'analyse de l'évolution d'une certaine quantité dans le temps. La modélisation espace-état appliquée aux séries chronologiques peut s'avérer complexe et compliquée. L'idée d'une approche équivalente simple et efficace, permettant d'obtenir de bons résultats pour un faible coût en efforts et complexité, semble être attrayante. Shumway2000 et al. [145, 146] ont montré que des modèles espace-état simple, générant une matrice de transition connue, appartiennent à la classe des modèles linéaires mixtes.
Dans cette étude, nous proposons une stratégie générale pour transformer un modèle espace-état en modèle mixte. Cette stratégie générale est ensuite appliquée aux séries chronologiques structurelles.
Cette procédure permettra de comptabiliser d'une manière explicite et flexible, les différentes sources de variation aléatoire. Des procédures d'estimation de paramètres, tel que le maximum de vraisemblance restreint (REML) [37, 147] peuvent alors être exploitées.
Dans ce chapitre, nous commençons par démontrer le passage générale d'un modèle espace-état à un modèle à effets mixtes. Ensuite, en se basant sur le livre [48] et l'article [49], nous passons en revue la classe des séries chronologiques structurelles (tendance, saisonnalité, cycle...). Nous explicitons en détail, pour chaque modèle appartenant à cette classe, sa formulation espace-état et sa représentation modèle à effets mixtes équivalente.
Via des simulations et des données réelles, utilisant le registre d'AVC du sud de Londres (SLSR)[6], nous montrons que les estimation basées sur le filtre de Kalman [11, 50], spécifique aux modèles espace-état, sont quasi-équivalentes à la meilleure prédiction linéaire sans biais (BLUP) [12], basée sur le (REML), dans un modèle mixte.
Comme limitation de la méthodologie proposée, nous étudions et discutons le cas de modèles espace-état générant une matrice de transition complexe (plus de deux paramètres).
Pour illustrer la méthodologie proposée, nous nous sommes restreint concernant les simulations et les applications, au *"local level model"* [48]. Les autres modèles étudiés et expliqués théoriquement en détail, peuvent être validés de la même manière. Ils feront

donc l'objet de travaux ultérieurs, ouvrant ainsi de nouvelles perspectives d'applications à des sujets de santé publique, en l'occurrence les AVC.
Par ce travail, nous proposons une méthodologie optimale permettant l'utilisation des modèles à effets mixtes pour des cas ne nécessitant pas un cadre de modélisation complexe, tel les modèles espace-état.

4.1 Modèle espace-état linéaire gaussien

Le modèle espace-état linéaire gaussien est défini dans différents livres et articles par différentes notations. Nous considérons pour ce chapitre, les notations utilisées dans le livre de Durbin et Kopman. Le modèle espace-état linéaire gaussien est défini par :

$$\begin{aligned} \boldsymbol{y_t} &= \boldsymbol{Z_t\alpha_t} + \boldsymbol{\varepsilon_t}, \qquad \boldsymbol{\varepsilon_t} \sim \mathbf{N}\left(\mathbf{0}, \boldsymbol{H_t}\right) \\ \boldsymbol{\alpha_{t+1}} &= \boldsymbol{T_t\alpha_t} + \boldsymbol{R_t\eta_t}, \quad \boldsymbol{\eta_t} \sim \mathbf{N}\left(\mathbf{0}, \boldsymbol{Q_t}\right), \quad \boldsymbol{t = 1, \ldots, n} \end{aligned} \tag{4.1}$$

Avec y_t le vecteur d'observation de dimension $p \times 1$; α_t le vecteur d'état de dimension $m \times 1$; Z_t la matrice de mesure de dimension $p \times m$; T_t la matrice de transition de dimension $m \times 1$ dont tous les éléments sont supposés connus ; ε_t est le vecteur d'erreur de dimension $p \times 1$; η_t est le vecteur de dimension $r \times 1$; R_t est une matrice de dimension $m \times r$; H_t est une matrice de dimension $p \times p$.
Les erreurs ε_t et η_t sont indépendantes.
Pour plus de simplicité, nous pouvons définir $\eta_t^* = R_t\eta_t$ et $Q_t^* = R_tQ_tR_t'$ sans inclure explicitement R_t. Notons que dans plusieurs applications, la matrice R_t est la matrice identité.
La première équation est appelée l'équation de l'observation et la deuxième équation est appelée l'équation de l'état.
Ce modèle peut être utilisé pour une analyse classique et bayésienne.
L'équation de l'observation a une structure d'un modèle de régression linéaire où le vecteur de coefficient α_t varie en fonction du temps. La deuxième équation représente un modèle vectoriel autorégressif du premier ordre, dont la nature markovienne explique plusieurs propriétés du modèle espace-états.

4.2 Formulation générale

Considérons l'équation d'état définissant le modèle espace-état.
Nous réécrivons l'équation d'état par récursions :

$$\begin{aligned} \boldsymbol{\alpha_t} &= \boldsymbol{T_t\alpha_{t-1}} + \boldsymbol{\eta_t^*} \qquad t = 1, \cdots, n \\ &= \boldsymbol{T_t}\left(\boldsymbol{T_{t-1}\alpha_{t-2}} + \boldsymbol{\eta_{t-1}^*}\right) + \boldsymbol{\eta_t^*} \qquad t = 1, \cdots, n \\ &= \vdots \\ &= \left(\prod_{i=1}^{'} \boldsymbol{T_i}\right)\boldsymbol{\alpha_0} + \left[\sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{'} \boldsymbol{T_j}\right)\boldsymbol{\eta_i^*}\right] + \boldsymbol{\eta_t^*} \qquad t = 1, \cdots, n \end{aligned}$$

En remplaçant α_t dans l'équation d'observation y_t :

$$\boldsymbol{y_t} = \boldsymbol{Z_t}\left(\prod_{\boldsymbol{i=1}}^{'} \boldsymbol{T_i}\right)\boldsymbol{\alpha_0} + \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{t-1}} \boldsymbol{Z_t}\left(\prod_{\boldsymbol{j=i+1}}^{'} \boldsymbol{T_j}\right)\boldsymbol{\eta_i^*}\right] + \boldsymbol{Z_t\eta_t^*} + \boldsymbol{\varepsilon_t} \qquad \boldsymbol{t = 1, \cdots, n}$$

Posons :

$$\boldsymbol{X_t = Z_t \left(\prod_{i=1}^{'} T_i \right)}$$

et :

$$\boldsymbol{V}_{it} = \begin{cases} \boldsymbol{Z_t} \left(\prod_{j=i+1}^{'} \boldsymbol{T}_j \right) & \text{si } \ i < t \\ \boldsymbol{Z}_t & \text{si } i = t \\ \boldsymbol{0} & \text{si } t < i \leq n \end{cases}$$

alors :

$$\underbrace{\boldsymbol{y_t}}_{\boldsymbol{p\times 1}} = \underbrace{\boldsymbol{X_t \alpha_0}}_{\boldsymbol{p\times 1}} + \underbrace{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{V'_{it} \eta_i^*}}_{\boldsymbol{p\times 1}} + \underbrace{\boldsymbol{\epsilon_t}}_{\boldsymbol{p\times 1}} \quad \boldsymbol{\varepsilon_t \sim \mathrm{N}\left(0, H_t\right)} \quad \boldsymbol{\eta_t^* \sim \mathrm{N}\left(0, Q_t^*\right)}, \quad \boldsymbol{t = 1, \ldots, n} \tag{4.2}$$

Nous obtenons ainsi une représentation en modèle à effets mixtes du modèle espace-état telle que :
y_t est un vecteur connu d'observations.
α_0 un vecteur inconnu d'effets fixes.
η_t^* un vecteur inconnu d'effets aléatoires.
ϵ_t un vecteur inconnu d'effets aléatoires.
X_t et V_{it} des matrices liant les observations y_t à α_0 et η_t respectivement.
L'utilisation d'une distribution de probabilité a priori pour l'initialisation est l'approche souvent utilisée par la plupart des analystes de séries chronologiques dans le cas où aucune information n'est fournie sur la valeur initiale α_0. Cependant, certains chercheurs trouvent que cette approche n'est pas satisfaisante. En effet, l'hypothèse d'une variance infinie n'est pas assurée puisque toutes les séries chronologiques observées ont des valeurs finies. Dans ce sens, une autre approche consiste à supposer que α_0 est une constante inconnue à estimer à partir des données via le maximum de vraisemblance. La forme la plus simple de cette idée est d'estimer α_0 par maximum de vraisemblance à partir de la première observation [48].

4.3 *"Local level model"*

Le *"local level model"* est un exemple simple des modèles espace-état. Il est formulé par le système d'équations suivant :

$$\begin{aligned} \boldsymbol{y_t} &= \boldsymbol{\mu_t + \varepsilon_t}, & \boldsymbol{\varepsilon_t} &\boldsymbol{\sim \text{i.i.d.} \mathcal{N}\left(0, U_\varepsilon\right)} \\ \boldsymbol{\mu_{t+1}} &= \boldsymbol{\mu_t + \xi_t}, & \boldsymbol{\xi_t} &\boldsymbol{\sim \text{i.i.d.} \mathcal{N}\left(0, W_\xi\right)} \end{aligned} \tag{4.3}$$

La représentation espace-état de ce modèle se déduit facilement et est donnée par :

$$\alpha_t = \mu_t, \quad \eta_t = \xi_t, \quad Z_t = T_t = R_t = 1, \quad Q_t = W_\xi, \quad H_t = U_\varepsilon$$

Notons que les propriétés dynamiques relatives à l'état du système au moment $t+1$ sont exprimés en fonction de l'état du système au moment t. Dans le cas où $\xi_t = 0$ pour $t = 1, \ldots, n$, le modèle est réduit à un modèle de régression linéaire qui peut être résolu analytiquement.
Le *"local level model"* peut être représenté sous forme d'un modèle à effets mixtes.
A partir de la formulation générale, cette représentation est donnée par :

$$X_t = 1 \text{ et } V_{it} = \begin{cases} 1 & \text{si} \quad i < t \\ 1 & \text{si} \quad i = t \\ 0 & \text{si} \quad t < i \leq n \end{cases}$$

Alors :

$$\boldsymbol{y_t = \alpha_0 + \sum_{i=1}^{n} V_{it}\eta_i^* + \epsilon_t \quad \varepsilon_t \sim \mathrm{N}\left(0, H_t\right) \quad \eta_t^* \sim \mathrm{N}\left(0, Q_t^*\right), \quad t = 1, \ldots, n.} \tag{4.4}$$

4.4 Les séries chronologiques structurelles

Une série chronologique est un ensemble d'observations $y_1, \ldots, y_n$ ordonnée dans le temps. Le modèle de base pour définir une série chronologique est le modèle additif donné par :

$$\boldsymbol{y_t = \mu_t + \gamma_t + \varepsilon_t, \quad t = 1, \ldots, n.} \tag{4.5}$$

μ_t est appelée la tendance (trend), γ_t est une composante périodique d'une période fixe appelée composante saisonnière et ε_t est appelée bruit ou résidu. Dans de nombreuses applications ces composantes se combinent de manière multiplicative :

$$\boldsymbol{y_t = \mu_t \gamma_t \varepsilon_t.} \tag{4.6}$$

Le modèle multiplicative peut se ramener au modèle additif en appliquant la fonction "log". Nous définissons μ_t et γ_t comme une marche aléatoire.
La marche aléatoire est une série α_t définie par $\alpha_{t+1} = \alpha_t + \eta_t$ où les η_t sont des variables aléatoires i.d.d avec une moyenne zero et une variance σ_η^2.
Les séries chronologiques structurelles sont des modèles où les composantes : tendancielle, saisonnière et résiduelle sont modélisées explicitement [49].

4.4.1 Composante tendancielle : *"local linear trend model"*

Le *"local level model"* (4.3), est une forme simple d'un modèle de série chronologique structurelle. En ajoutant un terme ν_t, qui est généré par une marche aléatoire (random walk), nous obtenons le modèle suivant :

$$\begin{aligned} \boldsymbol{y_t} &= \boldsymbol{\mu_t + \varepsilon_t,} & \boldsymbol{\varepsilon_t} &\boldsymbol{\sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right)} \\ \boldsymbol{\mu_{t+1}} &= \boldsymbol{\mu_t + \nu_t + \xi_t,} & \boldsymbol{\xi_t} &\boldsymbol{\sim \mathrm{N}\left(0, \sigma_\xi^2\right)} \\ \boldsymbol{\nu_{t+1}} &= \boldsymbol{\nu_t + \zeta_t,} & \boldsymbol{\zeta_t} &\boldsymbol{\sim \mathrm{N}\left(0, \sigma_\zeta^2\right)} \end{aligned} \tag{4.7}$$

Le modèle (4.7) est appelé *"local linear trend model"*. C'est un cas particulier des modèles espace-état. Sa représentation espace-état est donnée par :

$$\begin{aligned} y_t &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \varepsilon_t \\ \begin{pmatrix} \mu_{t+1} \\ \nu_{t+1} \end{pmatrix} &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \begin{pmatrix} \xi_t \\ \zeta_t \end{pmatrix} \end{aligned}$$

Le modèle (4.7) admet également une représentation modèle à effets mixtes.
En effet :

$$y_t = [1\ 0] \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \alpha_0 + \sum_{i=1}^{t-1} [1\ 0] \Big(\prod_{j=i+1}^{\prime} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \Big) \eta_i + [1\ 0]\eta_t + \epsilon_t, \quad t = 1 \cdots n$$

$$= [1\ 1]\alpha_0 + \sum_{i=1}^{t-1} [1\ 0] \Big(\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{t-i-1} \Big) \eta_i + [1\ 0]\eta_t + \epsilon_t, \quad t = 1 \cdots n$$

D'après la décomposition de Dunford :

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{t-i-1} = \begin{bmatrix} 1 & (t-i-1) \\ 0 & 1 \end{bmatrix}$$

Finalement :

$$y_t = X_t \alpha_0 + \sum_{i=1}^{n} V_{it} \eta_i + \epsilon_t, \quad \varepsilon_t \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right) \quad \eta_t \sim \mathrm{N}\left(0, \sigma_\xi^2\right), \quad t = 1, \ldots, n$$

avec :

$$X_t = [1\ 1]$$

et :

$$V_{it} = \begin{cases} [1 \quad (t-i-1)] & \text{si} \quad i < t \\ [1 \quad 0] & \text{si} \quad i = t \\ 0 & \text{si} \quad t < i \leq n \end{cases}$$

4.4.2 Composante saisonnière

La saisonnalité correspond a un phénomène qui se répète dans un intervalle de temps régulier (périodique), d'où le terme de variations saisonnières. Comme pour la composante tendancielle, nous pouvons supposer que la composante saisonnière change au fil du temps. Il existe de nombreuses raisons pour lesquelles des changements dans la configuration saisonnière peuvent avoir lieu [148].
Si la composante saisonnière est déterministe, les valeurs saisonnières variant du mois 1 à s peuvent être modélisées par des constantes γ_j^* $j = 1, ..., s$ tel que $\sum_{j=1}^{s^*} \gamma_j^* = 0$. Cette hypothèse garantie que dans ce cas, la composante saisonnière ne peut être confondu avec la composante tendancielle.
Dans le cas stochastique, la composante saisonnière est définie par le modèle suivant :

$$\boldsymbol{\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j} + \omega_t, \quad \omega_t \sim \mathrm{N}\left(0, \sigma_\omega^2\right).} \tag{4.8}$$

Nous pouvons aussi définir le modèle de saisonnalité stochastique sous la forme trigonométrique suivante [149, 150] :

$$\boldsymbol{\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{jt}.} \tag{4.9}$$

Tel que $\gamma_{j,t}$ est défini par :

$$\begin{bmatrix} \boldsymbol{\gamma_{j,t}} \\ \boldsymbol{\gamma_{j,t}^*} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\cos \lambda_j} & \boldsymbol{\sin \lambda_j} \\ \boldsymbol{-\sin \lambda_j} & \boldsymbol{\cos \lambda_j} \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma_{j,t-1}} \\ \boldsymbol{\gamma_{j,t-1}^*} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\omega_{j,t}} \\ \boldsymbol{\omega_{j,t}^*} \end{bmatrix}, \quad \boldsymbol{j = 1, \ldots, [s/2].} \tag{4.10}$$

avec $\lambda_j = 2\pi j/\mathrm{s}$, ω_t et ω_t^* deux bruit blanc mutuellement non corrélées avec une moyenne nulle et une variance commune σ_ω^2 pour $t = 1, \ldots, T$.
Avec $\omega_{jt} \sim \mathrm{N}(0, \sigma_\omega^2)$ et $\omega_{jt}^* \sim \mathrm{N}(0, \sigma_\omega^2)$ des variables indépendantes.

Si s est pair, $[s/2] = s/2$.
Si s est impair, $[s/2] = (s-1)/2$.
Quand s est pair, pour $j = s/2$ le système se réduit à

$$\gamma_{j,t} = \gamma_{j,t-1} \cos \lambda_j + \omega_{j,t}, \quad j = s/2$$

4.4.3 *"Local linear trend model"* avec composante saisonnière

Le *"local linear trend model"* avec composante saisonnière est défini par :

$$\begin{aligned} \boldsymbol{y_t} &= \boldsymbol{\mu_t} + \boldsymbol{\gamma_t} + \boldsymbol{\varepsilon_t}, \quad \boldsymbol{t = 1, \ldots, n} \\ \boldsymbol{\alpha_t} &= \begin{pmatrix} \boldsymbol{\mu_t} & \boldsymbol{\nu_t} & \boldsymbol{\gamma_t} & \boldsymbol{\gamma_{t-1}} & \cdots & \boldsymbol{\gamma_{t-s+2}} \end{pmatrix}' \end{aligned} \tag{4.11}$$

La formulation espace-état de ce modèle est définie par :

$$Z_t = \left(Z_{[\mu]}, Z_{[\gamma]}\right) \qquad T_t = \mathrm{diag}\left(T_{[\mu]}, T_{[\gamma]}\right)$$

$$R_t = \mathrm{diag}\left(R_{[\mu]}, R_{[\gamma]}\right) \qquad Q_t = \mathrm{diag}\left(Q_{[\mu]}, Q_{[\gamma]}\right)$$

avec :

$$Z_{[\gamma]} = (1, 0, ..., 0) \qquad Z_{[\mu]} = (1, 0, ..., 0)$$

$$T_{[\mu]} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \qquad T_{[\gamma]} = \begin{bmatrix} -1 & -1 & \ldots & -1 & -1 \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ 0 & 0 & & 1 & 0 \end{bmatrix}$$

$$R_{[\mu]} = I_2 \qquad R_{[\gamma]} = (1, 0, \cdots, 0)$$

$$Q_{[\mu]} = \begin{bmatrix} \sigma_\xi^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix} \qquad Q_{[\gamma]} = \sigma_w^2$$

Le *"local linear trend model"* avec composante saisonnière (4.11) admet également une représentation en modèle à effets mixtes.
Pour illustrer ce fait, et pour plus de simplicité, nous traitons l'exemple où $s = 4$.
Pour cet exemple, le modèle est défini par :

$$\begin{aligned} \boldsymbol{y_t} &= \boldsymbol{\mu_t} + \boldsymbol{\gamma_{1,t}} + \boldsymbol{\varepsilon_t}, & \boldsymbol{\varepsilon_t} &\sim \boldsymbol{\mathcal{N}}\left(\boldsymbol{0}, \boldsymbol{\sigma_\varepsilon^2}\right) \\ \boldsymbol{\mu_{t+1}} &= \boldsymbol{\mu_t} + \boldsymbol{v_t} + \boldsymbol{\xi_t}, & \boldsymbol{\xi_t} &\sim \boldsymbol{\mathcal{N}}\left(\boldsymbol{0}, \boldsymbol{\sigma_\xi^2}\right) \\ \boldsymbol{v_{t+1}} &= \boldsymbol{v_t} + \boldsymbol{\zeta_t}, & \boldsymbol{\zeta_t} &\sim \boldsymbol{\mathcal{N}}\left(\boldsymbol{0}, \boldsymbol{\sigma_\zeta^2}\right) \\ \boldsymbol{\gamma_{1,t+1}} &= -\boldsymbol{\gamma_{1,t}} - \boldsymbol{\gamma_{2,t}} - \boldsymbol{\gamma_{3,t}} + \boldsymbol{\omega_t}, & \boldsymbol{\omega_t} &\sim \boldsymbol{\mathcal{N}}\left(\boldsymbol{0}, \boldsymbol{\sigma_\omega^2}\right) \\ \boldsymbol{\gamma_{2,t+1}} &= \boldsymbol{\gamma_{1,t}} \\ \boldsymbol{\gamma_{3,t+1}} &= \boldsymbol{\gamma_{2,t}} \end{aligned} \tag{4.12}$$

La formulation espace-état est donnée par les composantes suivantes :

$$\alpha_t = \begin{pmatrix} \mu_t \\ v_t \\ \gamma_{1,t} \\ \gamma_{2,t} \\ \gamma_{3,t} \end{pmatrix}, T_t = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, Z_t = \begin{pmatrix} 1, 0, 1, 0, 0 \end{pmatrix}$$

$$Q_t = \begin{pmatrix} \sigma_\xi^2 & 0 & 0 \\ 0 & \sigma_\zeta^2 & 0 \\ 0 & 0 & \sigma_\omega^2 \end{pmatrix}, R_t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

La représentation modèle linéaire à effets mixtes équivalente est donnée par :

$$y_t = X_t\alpha_0 + \sum_{i=1}^{n} V_{it}\eta_i + \epsilon_t, \quad \varepsilon_t \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right) \quad \eta_t \sim \mathrm{N}\left(0, \sigma_\xi^2\right), \quad t = 1, \ldots, n$$

avec :

$$X_t = [1\ 1\ -1\ -1\ -1]$$

et :

$$V_{it} = \begin{cases} [1\ 0\ 1\ 0\ 0] \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}^{t-i-1} & \text{si} \quad i < t \\ [1\ 0\ 1\ 0\ 0] & \text{si} \quad i = t \\ 0 & \text{si} \quad t < i \leq n \end{cases}$$

Nous traitons un autre exemple pour le modèle (4.11) en considérant la formulation trigonométrique pour le cas impair, avec $s = 3$.
Dans cet exemple, $j = 1$ et $\lambda_1 = \frac{2\pi}{3}$:

$$Z_t = (1, 0, 1, 0) \qquad T_{[\gamma]} = [C_1]$$

$$C_1 = \begin{pmatrix} \frac{-1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{-1}{2} \end{pmatrix} \qquad T_t = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{-1}{2} & \frac{\sqrt{3}}{2} \\ 0 & 0 & -\frac{\sqrt{3}}{2} & \frac{-1}{2} \end{pmatrix} \qquad R_t = I_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$Q_t = \begin{pmatrix} \sigma_\xi^2 & 0 & 0 & 0 \\ 0 & \sigma_\zeta^2 & 0 & 0 \\ 0 & 0 & \sigma_\omega^2 & 0 \\ 0 & 0 & 0 & \sigma_\omega^2 \end{pmatrix}$$

La représentation en modèle linéaire mixte est donné par :

$$y_t = X_t\alpha_0 + \sum_{i=1}^{n} V_{it}\eta_i + \epsilon_t, \quad \varepsilon_t \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right) \quad \eta_t \sim \mathrm{N}\left(0, \sigma_\xi^2\right), \quad t = 1, \ldots, n$$

avec :

$$X_t = [1, 1, \frac{-1}{2}, \frac{\sqrt{3}}{2}]$$

et :

$$V_{it} = \begin{cases} [1\ 0\ 1\ 0] \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{-1}{2} & \frac{\sqrt{3}}{2} \\ 0 & 0 & -\frac{\sqrt{3}}{2} & \frac{-1}{2} \end{bmatrix}^{t-i-1} & \text{si} \quad i < t \\ [1\ 0\ 1\ 0] & \text{si} \quad i = t \\ 0 & \text{si} \quad t < i \leq n \end{cases}$$

4.4.4 Composante cyclique

Nous définissons la composante cyclique c_t par :

$$\boldsymbol{c_t = \tilde{c} \cos \lambda_c t + \tilde{c}^* \sin \lambda_c t.} \tag{4.13}$$

Avec λ_c la fréquence du cycle ; la période du cycle est $2\pi/\lambda_c$ qui est plus grande que la période saisonnière s. Comme pour la composante saisonnier, la composante cyclique peut varier de manière stochastique au fil du temps tel que :

$$\begin{aligned} c_{t+1} &= c_t \cos \lambda_c + c_t^* \sin \lambda_c + \tilde{\omega}_t \\ c_{t+1}^* &= -c_t \sin \lambda_c + c_t^* \cos \lambda_c + \tilde{\omega}_t^* \end{aligned}$$

Avec $\tilde{w}_t \sim \mathrm{N}(0, \sigma_{\tilde{w}}^2)$ et $\tilde{w}_t^* \sim \mathrm{N}(0, \sigma_{\tilde{w}}^2)$ des variables indépendantes. Les cycles de cette forme rentrent dans le cadre du modèle de séries chronologiques structurelles. La fréquence λ_c est un paramètre à estimer.
La représentation espace-état pour la composante cyclique est similaire à une composante saisonnière trigonométrique mais avec une fréquence λ_c.
La représentation espace-état de la composante cyclique est donnée par :

$$\begin{aligned} Z_{[c]} &= (1, 0), & T_{[c]} &= C_c \\ R_{[c]} &= I_2, & Q_{[c]} &= \sigma_{\tilde{\omega}}^2 I_2 \end{aligned}$$

La matrice C_c est définie par C_j avec $\lambda_j = \lambda_c$.
La représentation en modèle à effets mixtes équivalente à cette formulation espace-état est facilement déductible.

4.4.5 Variables explicatives et effets d'intervention

Les variables explicatives et les effets d'intervention rentrent dans le cadre de la classe des modèles structurels. Supposons que nous ayons k régresseurs $x_{1t}, \cdots, x_{kt}$ avec des coefficients de régression $\beta_1, \cdots, \beta_k$ qui sont constants dans le temps. Nous souhaitons mesurer la variation du niveau due à une intervention à l'instant τ. Nous définissons une variable d'intervention w_t comme suit :

$$\begin{aligned} w_t &= 0, & t < \tau \\ w_t &= 1, & t \geq \tau \end{aligned}$$

Considérons le modèle suivant [151] :

$$\boldsymbol{y_t = \mu_t + \gamma_t + c_t + \sum_{j=1}^{k} \beta_j x_{jt} + \delta w_t + \varepsilon_t, \quad t = 1, \ldots, n.} \tag{4.14}$$

δ mesure l'évolution du niveau de la série en τ en raison d'une intervention ponctuelle τ.
Le modèle (4.14) peut facilement être mis sous forme espace-état.
Par exemple, si $\gamma_t = c_t = \delta = 0$, $k = 1$ et μ_t est défini par (4.3), nous avons alors :

$$\alpha_t = (\mu_t \quad \beta_{1t})', \qquad Z_t = (1 \quad x_{1t})$$

$$T_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad R_t = (1 \;\; 0)' \qquad Q_t = (\sigma_\xi^2)$$

Le terme β_{1t} est constant tel que $\beta_{1,t+1} = \beta_{1t}$.
La formulation de ce modèle en modèle à effets mixtes est donnée par :

$$\begin{aligned} y_t &= [1 \;\; x_{1t}]\alpha_0 + \sum_{i=1}^{t-1} [1 \; x_{1t}] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{t-i-1} \eta_i + [1 \;\; x_{1t}]\eta_t + \epsilon_t \\ &= X_t\alpha_0 + \sum_{i=1}^{n} V_{it}\eta_i + \epsilon_t \end{aligned}$$

Avec :

$$X_t = [1 \;\; x_{1t}]$$

et :

$$V_{it} = \begin{cases} [1 \; x_{1t}] & \text{si} \quad i < t \\ [1 \; x_{1t}] & \text{si} \quad i = t \\ 0 & \text{si} \quad t < i \leq n \end{cases}$$

Dans le cas où nous considérons une variable de régression x_t et une variable d'intervention w_t :

$$\begin{aligned} \boldsymbol{y_t} &= \boldsymbol{\mu_t} + \boldsymbol{\beta_t x_t} + \boldsymbol{\lambda_t w_t} + \boldsymbol{\varepsilon_t}, & \boldsymbol{\varepsilon_t} &\sim \mathbf{NID}\left(\mathbf{0}, \boldsymbol{\sigma_\varepsilon^2}\right) \\ \boldsymbol{\mu_{t+1}} &= \boldsymbol{\mu_t} + \boldsymbol{\xi_t}, & \boldsymbol{\xi_t} &\sim \mathbf{NID}\left(\mathbf{0}, \boldsymbol{\sigma_\xi^2}\right) \\ \boldsymbol{\beta_{t+1}} &= \boldsymbol{\beta_t} \\ \boldsymbol{\lambda_{t+1}} &= \boldsymbol{\lambda_t} \end{aligned} \tag{4.15}$$

La représentation espace-état de (4.15) est donnée par :

$$\alpha_t = (\mu_t, \beta_t, \lambda_t)' \qquad \eta_t = \xi_t \qquad T_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$Z_t = (1 \; x_t \; w_t) \qquad H_t = \sigma_\epsilon^2$$

$$Q_t = \begin{bmatrix} \sigma_\xi^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad R_t = (1 \; 0 \; 0)$$

La représentation en modèle à effets mixtes équivalente est donnée par :

$$\begin{aligned} y_t &= [1 \; x_t \; w_t]\alpha_0 + \sum_{i=1}^{t-1} [1 \; x_t \; w_t] \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{t-i-1} \eta_i^* + [1 \; x_t \; w_t]\eta_t^* + \epsilon_t \\ &= X_t\alpha_0 + \sum_{i=1}^{n} V_{it}\eta_i^* + \epsilon_t \end{aligned}$$

avec :

$$X_t = [1 \quad x_t \quad w_t]$$

et :

$$V_{it} = \begin{cases} [1\ x_t\ w_t] & \text{si} \quad i < t \\ [1\ x_t\ w_t] & \text{si} \quad i = t \\ 0 & \text{si} \quad t < i \leq n \end{cases}$$

tel que $\beta = \beta_1 = \beta_t$ et $\lambda = \lambda_1 = \lambda_t$ pour $t = 1, \ldots, n$.
Il s'agit du modèle (4.3) avec une variable explicative continue x_t et une variable d'intervention w_t.

4.5 Matrice de transition avec paramètres

Dans cette section, nous montrons que la méthodologie proposé peut être étendue à quelques modèles espace-état générant une matrice de transition avec un ou deux paramètres. Avec une matrice générant plus de deux paramètres, la méthodologie proposée n'est plus généralisable. Il est ainsi recommandé d'utiliser l'approche espace-état.

4.5.1 Cas d'un seul paramètre : *"local linear trend model"* avec facteur d'amortissement

Le *"local linear trend model"* peut être modifié en rajoutant un facteur d'amortissement ρ [49] :

$$\begin{aligned} \boldsymbol{y_t} &= \boldsymbol{\mu_t} + \boldsymbol{\varepsilon_t}, & \boldsymbol{\varepsilon_t} &\sim \mathbf{N}\left(\mathbf{0}, \boldsymbol{\sigma_\varepsilon^2}\right) \\ \boldsymbol{\mu_t} &= \boldsymbol{\mu_{t-1}} + \boldsymbol{\nu_{t-1}} + \boldsymbol{\xi_t}, & \boldsymbol{\xi_t} &\sim \mathbf{N}\left(\mathbf{0}, \boldsymbol{\sigma_\xi^2}\right) \\ \boldsymbol{\nu_t} &= \boldsymbol{\rho\nu_{t-1}} + \boldsymbol{\zeta_t}, & \boldsymbol{\zeta_t} &\sim \mathbf{N}\left(\mathbf{0}, \boldsymbol{\sigma_\zeta^2}\right), \quad \mathbf{0 < \rho < 1} \end{aligned} \tag{4.16}$$

La représentation espace-état est donnée par :

$$\begin{aligned} y_t &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \varepsilon_t \\ \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} &= \begin{bmatrix} 1 & 1 \\ 0 & \rho \end{bmatrix} \begin{pmatrix} \mu_{t-1} \\ \nu_{t-1} \end{pmatrix} + \begin{pmatrix} \xi_t \\ \zeta_t \end{pmatrix} \end{aligned}$$

Ce modèle peut également être représenté comme un modèle à effets mixtes :

$$\begin{aligned} y_t &= [1\ 0] \begin{bmatrix} 1 & 1 \\ 0 & \rho \end{bmatrix} \alpha_0 + \sum_{i=1}^{t-1} [1\ 0] \Big(\prod_{j=i+1}^{'} \begin{bmatrix} 1 & 1 \\ 0 & \rho \end{bmatrix} \Big) \eta_i + [1\ 0]\eta_t + \epsilon_t, \quad t = 1 \cdots n \\ &= [1\ 1]\alpha_0 + \sum_{i=1}^{t-1} [1\ 0] \Big(\begin{bmatrix} 1 & 1 \\ 0 & \rho \end{bmatrix}^{t-i-1} \Big) \eta_i + [1\ 0]\eta_t + \epsilon_t, \quad t = 1 \cdots n \\ &= X_t \alpha_0 + \sum_{i=1}^{n} V_{it} \eta_i + \epsilon_t \end{aligned}$$

avec :

$$X_t = [1 \;\; 1]$$

et :

$$V_{it} = \begin{cases} [1\ 0] \begin{bmatrix} 1 & 1 \\ 0 & \rho \end{bmatrix}^{t-i-1} & \text{si} \quad i < t \\ [1\ 0] & \text{si} \;\; i = t \\ 0 & \text{si} \;\; t < i \leq n \end{cases}$$

4.5.2 Cas de deux paramètres :

Nous modifions le modèle (4.16) en rajoutant un autre facteur d'amortissement ϕ à μ_t :

$$\begin{aligned} \boldsymbol{y_t} &= \boldsymbol{\mu_t} + \boldsymbol{\varepsilon_t}, & \boldsymbol{\varepsilon_t} &\sim \mathbf{N}\left(\mathbf{0}, \boldsymbol{\sigma_\varepsilon^2}\right) \\ \boldsymbol{\mu_t} &= \boldsymbol{\phi\mu_{t-1}} + \boldsymbol{\nu_{t-1}} + \boldsymbol{\xi_t}, & \boldsymbol{\xi_t} &\sim \mathbf{N}\left(\mathbf{0}, \boldsymbol{\sigma_\xi^2}\right), \quad \mathbf{0 < \phi < 1} \\ \boldsymbol{\nu_t} &= \boldsymbol{\rho\nu_{t-1}} + \boldsymbol{\zeta_t}, & \boldsymbol{\zeta_t} &\sim \mathbf{N}\left(\mathbf{0}, \boldsymbol{\sigma_\zeta^2}\right), \quad \mathbf{0 < \rho < 1} \end{aligned} \tag{4.17}$$

La représentation espace-état est donnée par :

$$y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \varepsilon_t$$

$$\begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} = \begin{bmatrix} \phi & 1 \\ 0 & \rho \end{bmatrix} \begin{pmatrix} \mu_{t-1} \\ \nu_{t-1} \end{pmatrix} + \begin{pmatrix} \xi_t \\ \zeta_t \end{pmatrix}$$

Ce modèle peut également être représenté comme un modèle à effets mixtes :

$$\begin{aligned} y_t &= [1\ 0] \begin{bmatrix} \phi & 1 \\ 0 & \rho \end{bmatrix} \alpha_0 + \sum_{i=1}^{t-1} [1\ 0] \Big(\prod_{j=i+1}^{\prime} \begin{bmatrix} \phi & 1 \\ 0 & \rho \end{bmatrix} \Big) \eta_i + [1\ 0]\eta_t + \epsilon_t, \quad t = 1 \cdots n \\ &= [\phi\ 1]\alpha_0 + \sum_{i=1}^{t-1} [1\ 0] \begin{bmatrix} \phi & 1 \\ 0 & \rho \end{bmatrix}^{t-i-1} \eta_i + [1\ 0]\eta_t + \epsilon_t, \quad t = 1 \cdots n \\ &= X_t\alpha_0 + \sum_{i=1}^{n} V_{it}\eta_i + \epsilon_t \end{aligned}$$

avec :

$$X_t = [\phi \;\; 1]$$

et :

$$V_{it} = \begin{cases} [1\ 0] \begin{bmatrix} \phi & 1 \\ 0 & \rho \end{bmatrix}^{t-i-1} & \text{si} \quad i < t \\ [1\ 0] & \text{si} \;\; i = t \\ 0 & \text{si} \;\; t < i \leq n \end{cases}$$

Les modèles (4.16) et (4.17) ont montré que la méthodologie proposée reste valable pour des modèles générant des paramètres dans la matrice de transition. Néanmoins, cette méthodologie ne peux être généralisée pour des modèles avec des matrices de transitions plus complexes, c'est à dire avec plus de deux paramètres. Dans de tels cas, l'estimation de ces paramètres inconnus, figurant dans la matrice de transition, s'avère nécessaire.
Pour les cas simples, c'est à dire avec un ou deux paramètres dans la matrice de transition, nous suggérons d'utiliser la méthode du maximum de vraisemblance restreint ou profilé

[152]. En revanche, cette approche devient moins pratique lorsqu'il y a plusieurs paramètres (plus que deux paramètres) dans la matrice de transition T_t. Il est donc préférable d'utiliser les méthodes associées au modèle espace-état. Dans un cadre entièrement bayésien, l'estimation des modèles générant plusieurs paramètres dans la matrice de transition est relativement simple [153]. Le cas des modèles générant plusieurs paramètres dans la matrice de transition peut être considéré comme une limitation de la méthodologie proposée.

4.6 Validation par simulation

Afin d'illustrer la méthodologie proposée, nous avons simulé des données de série temporelle de taille 360. Ces données associent des observations rangées entre (6.86 et 10.95) à des marques temporelles successives équidistantes, c'est à dire séparées par la même durée (années).
Pour cet exemple, nous ajustons les données simulées via le *"local level model"* en utilisant les deux approches :
- Modèle espace-état basé sur le filtre de Kalman et l'algorithme EM [41, 42].
- Modèle à effets mixtes basé sur le maximum de vraisemblance restreint REML.

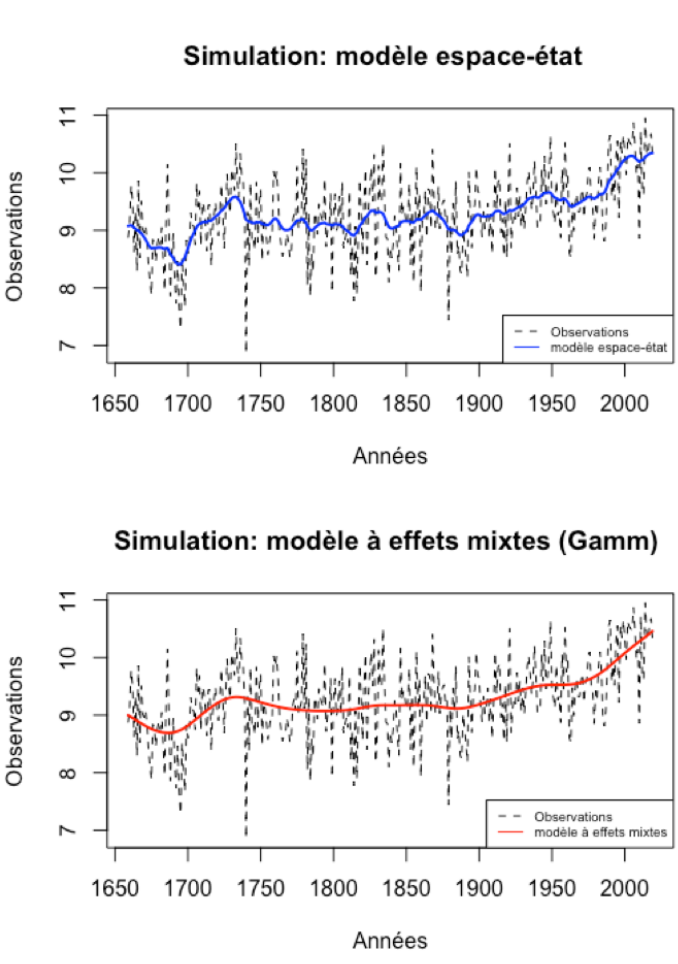

FIGURE 4.1 – Données simulées (observations annuelles de 1650 à 2019) : Modèle espace-état Vs Modèle à effets mixtes.

La courbe bleue figurant dans la Figure 4.1 représente l'estimation de l'état μ_t par le filtre de Kalman (lissage). La courbe rouge représente l'estimation par le maximum de vraisemblance restreint REML (BLUP) de l'état μ_t.
Nous remarquons que les deux estimations, via les deux approches sont quasi-équivalentes. En effet, pour l'estimation via la méthode de Kalman, propre au modèle espace-état, nous obtenons $\hat{\sigma_\epsilon} = 0.28$, le critère d'Akaike (AIC)= 641.07 et log-vraisemblance (LL)= -317.50. Concernant l'estimation via la méthode du maximum de vraisemblance restreint, propre au modèle à effets mixtes, nous obtenons $\hat{\sigma_\epsilon} = 0.32$, AIC = 656.85 et LL=-323.42.
Les résultats de cette simulation confirment que des techniques telles que le filtrage, la prévision et le lissage, basées sur le filtre de Kalman, sont équivalentes au BLUP dans un modèle mixte.
Cela revient au fait que les deux méthodes utilisent la même fonction de vraisemblance. La simulation a été conduite via le logiciel R. Ce modèle est implémenté avec une structure espace-état et modèle à effets mixtes.
Nous avons utilisé le package "mgcv" [154] et "nlme" [155] pour le modèle à effet mixtes. Concernant les modèles espace-état, plusieurs packages sont disponibles. Dans cette simulation, le package "MARSS" a été utilisé [156].

4.7 Application : Récupération fonctionnelle post AVC

Afin d'illustrer la méthodologie proposée par une application du monde réelle, nous avons utilisé des données du registre des accidents vasculaires cérébraux du sud de Londres (SLSR) [6].
Dans cet exemple, nous avons choisi d'observer la moyenne annuelle du score de Barthel de tous les patients enregistrés dans le (SLSR) jusqu'à 7 jours après l'admission, de 1995 à 2018 [9]. Le score de Barthel permet d'évaluer la progression de la récupération fonctionnelle avec des scores allant de 0 (totalement dépendant) à 20 (totalement indépendant). Cela nous permettra de décrire la variation du niveau du score de Barthel moyen annuel au cours des 23 années.
Cette application servira non seulement à la comparaison des deux approches de modélisation, mais aussi à l'évaluation et à la description de l'impact de la mise à jour des directives et conseils cliniques sur le diagnostic et la prise en charge des patients victime d'AVC ou d'accident ischémique transitoire (AIT), dans les 48 heures suivant l'apparition des symptômes [157, 158, 159, 160].
Nous proposons le *"local level model"* pour ajuster nos données. Les deux approches espace-état et modèle mixtes sont appliquées pour le même modèle et données.
Nous proposons deux exemples :
Dans le premier exemple, nous considérons tous les patients enregistrés dans le SLSR.
Dans le deuxième exemple, nous stratifions nos données par groupe en considérant les patients avec un age supérieur à 65 ans.

4.7.1 Exemple 1 :

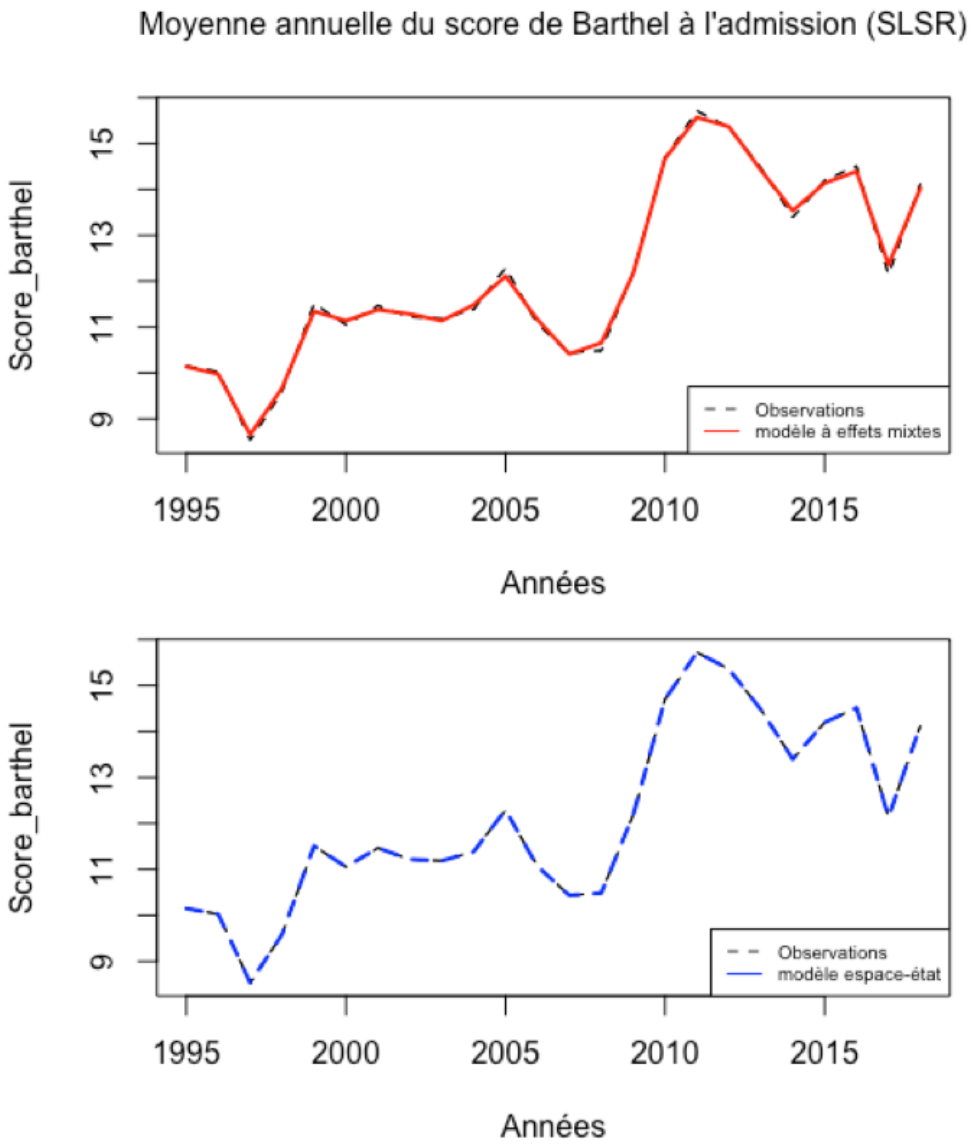

FIGURE 4.2 – Variation du score de Barthel moyen annuel jusqu'à 7 jours après l'admission (1995 -2019) : Modèle à effets mixtes Vs Approche espace-état.

La courbe bleue figurant dans la Figure 4.2 représente l'estimation de l'état μ_t par le filtre de Kalman (lissage). La courbe rouge représente l'estimation par le maximum de vraisemblance restreint REML (BLUP) de l'état μ_t.
Nous remarquons que les deux estimations, via les deux approches sont quasi-équivalentes, confirmant ainsi les résultats du cadre théorique proposé et la simulation conduite.
En effet, pour l'estimation via la méthode de Kalman, propre au modèle espace-état, nous obtenons $\hat{\sigma_\epsilon} = 0.04$, le critère d'Akaike (AIC)= 82.92 et log-vraisemblance (LL)=-37.86.
Concernant l'estimation via la méthode du maximum de vraisemblance restreint, propre au modèle à effets mixtes, nous obtenons $\hat{\sigma_\epsilon} = 0.08$, AIC = 80.90 et LL=-36.45.

4.7.2 Exemple 2

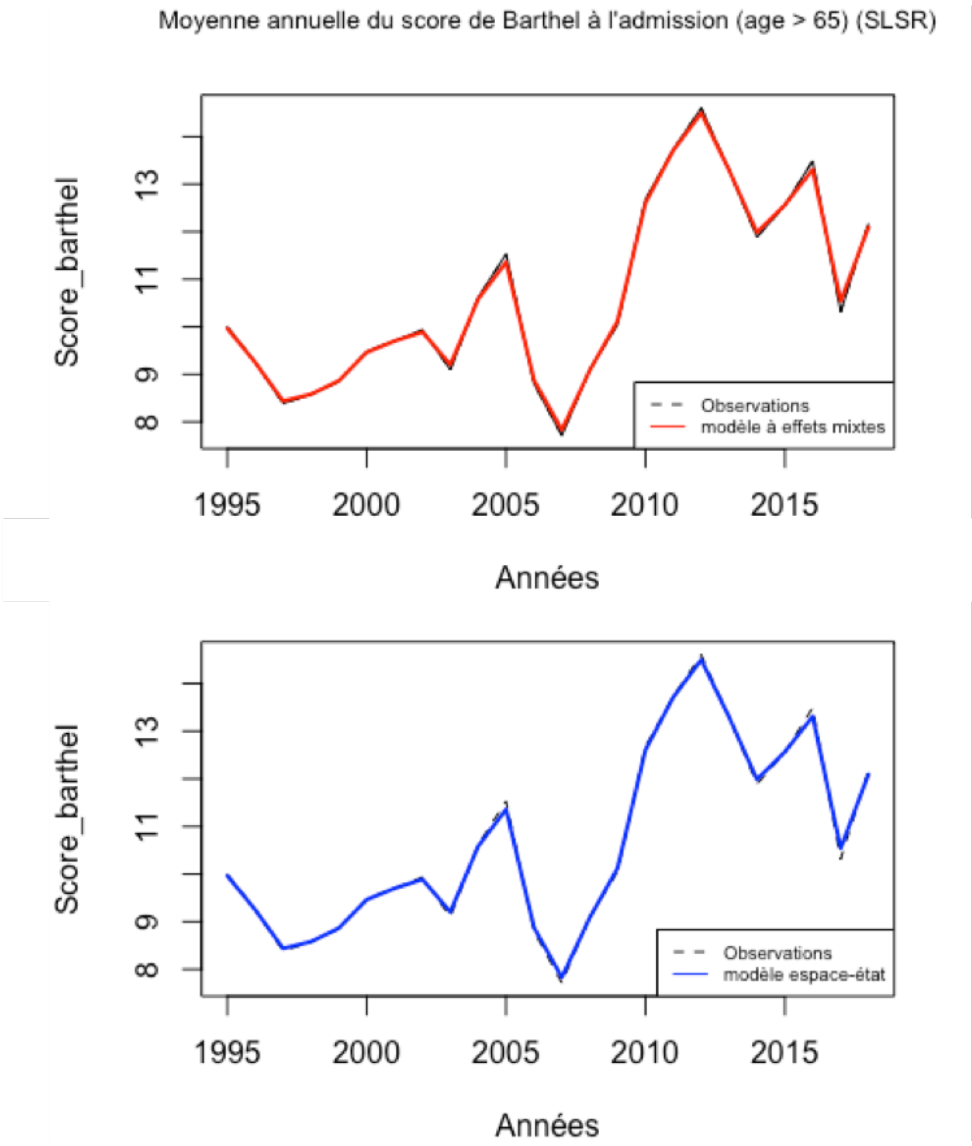

FIGURE 4.3 – Variation du score de Barthel moyen annuel jusqu'à 7 jours après l'admission (age>65) (1995 -2019) : Modèle à effets mixtes Vs Approche espace-état.

La figure 4.3 affirme que les deux estimations, via les deux approches sont quasi-équivalentes. En effet, pour l'estimation via la méthode de Kalman, relative au modèle espace-état, nous obtenons $\hat{\sigma_\epsilon} = 0.27$, AIC = 90.11 et log-vraisemblance (LL)=-41.45.
Concernant l'estimation via la méthode du maximum de vraisemblance restreint, propre au modèle à effets mixtes, nous obtenons $\hat{\sigma_\epsilon} = 0.27$, AIC = 88.39 et LL=-40.19.
Cette application confirme que la dynamique des effets des directives, souvent traitée par des techniques stochastiques, peut facilement être abordée par des modèles à effets mixtes.

4.7.3 Interpretation du modèle

Après avoir ajusté le modèle à nos données dans les deux exemples, nous pouvons commencer à l'utiliser et à l'interroger à diverses fins. Une question clé que nous pourrions poser au modèle est la suivante : Les scores de Barthel ont-ils augmenté ou diminué de manière statistiquement significative et comment peut on interpreter cliniquement ces

changements ?
Pour répondre à ces questions, et puisque les deux approches de modélisation sont équivalentes, nous considérons le modèle à effets mixtes. Une approche répondant à cette question consiste à calculer les dérivées premières de la tendance ajustée. Pour ce faire, nous pouvons utiliser la méthode des différences finies vu que la forme analytique de ces dérivées n'est pas explicite. Pour produire les dérivées via les différences finies, nous calculons les valeurs de la tendance ajustée sur une grille de points sur l'ensemble des données. D'après la figure 4.4, les dérivés suggèrent une période d'augmentation significative du score de Barthel de 2008 à 2011 avec un niveau de confiance de 99 %. Cette période est marquée par le changement dans les directives couvrant les interventions au stade aigu d'un AVC ou d'un accident ischémique transitoire (AIT) [158]. Ces directives offrent ainsi de meilleurs conseils cliniques sur le diagnostic et la prise en charge aiguë des AVC et des AIT dans les 48 heures suivant l'apparition des symptômes. Nous pouvons déduire que les recommandations et les changements apportés par ce (guidelines) [158] sont efficaces et ont permis une amélioration dans le protocole de gestion des patients, qui est expliqué par l'augmentation du score de Barthel.
Compte tenu de cet échantillon de données, et sachant que notre estimation de la tendance est sujette à l'incertitude, nous ne sommes pas en mesure de détecter d'autres périodes de changement significatif du score de Barthel autres que la période indiquée en bleu.

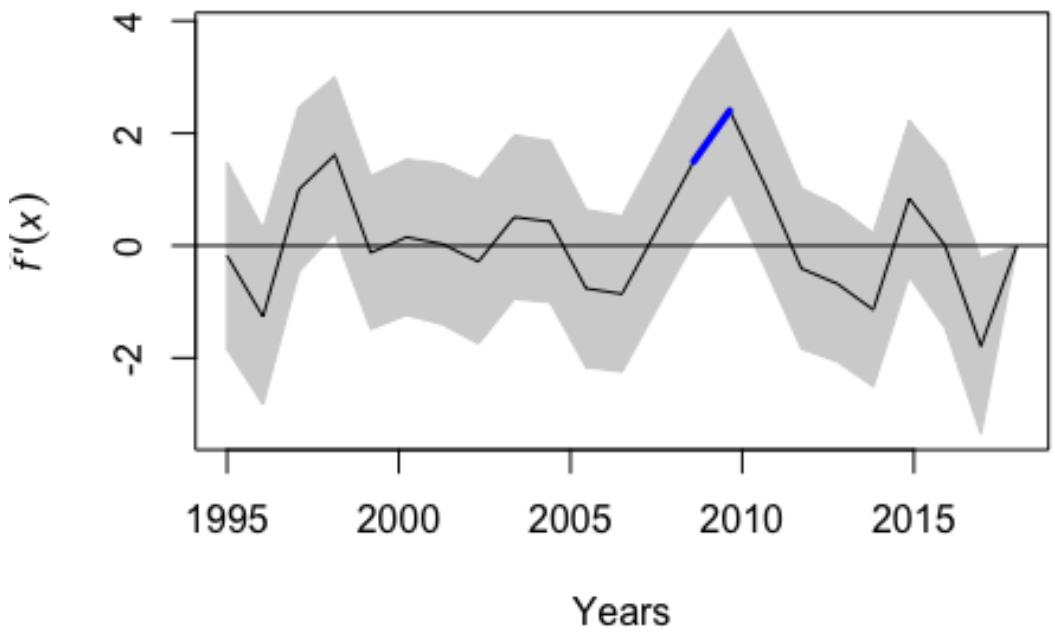

FIGURE 4.4 – Dérivées premières de la tendance ajustée avec un niveau de confiance de 99% (modèle à effets mixtes). Périodes de changement significatif (bleu).

4.8 Discussion

Dans ce chapitre, nous avons proposé une stratégie générale pour transformer un modèle espace-état en modèle à effets mixte.
Ce travail aborde pour la première fois, à notre connaissance, la méthodologie proposée d'une manière détaillée et explicite avec une application aux AVC.
Un intérêt particulier a été porté à la classe des séries temporelles structurelles. Nous avons montré que chaque modèle appartenant à cette classe, peut être décrit à la fois, par un modèle espace-état et un modèle à effets mixtes équivalent.
Nous avons également étudié le cas de modèles générant des paramètres dans la matrice de transition. Nous avons remarqué que l'approche proposée reste valable pour des matrices de transitions avec un ou deux paramètres. Dans ce cas, ces paramètres peuvent facilement être estimés par la méthode du maximum de vraisemblance restreint (REML) ou profilé. En raison de complexité du calcul, l'existence de plus de deux paramètres dans la matrice de transition est cependant considérée comme une limitation de cette méthodologie. L'approche proposée rend possible la modélisation des données complexes (cluster ou longitudinales) et des données avec plusieurs sources de variations [161]. En effet, de nombreuses séries chronologiques ont une structure complexe qui nécessite l'ajout d'effets fixes et aléatoires. Ces effets sont plus simple à considérer dans un modèle à effets mixtes que dans un modèle espace-état. De plus, le modèle à effets mixtes peut être vu comme une version régularisée du modèle linéaire généralisé (GLM) [44]. Il fusionne également les propriétés du (GLM) avec celle du modèle additif [47]. Ainsi, il devient facile de composer avec la saisonnalité de plusieurs périodes et de laisser l'analyste explorer des hypothèses sur les différentes composantes de la série chronologiques. Le modèle à effets mixtes est également un outil pratique pour traiter les données manquantes [162].
En citant les avantages que le modèle à effets mixtes procure, nous ne préconisons pas l'abandon de l'approche espace-état. En revanche, nous recommandons l'utilisation des modèles espace-état dans les cas où la méthodologie proposée n'est pas applicable c.-à-d. (matrice de transition avec plus de deux paramètres).
Un avantage conceptuel des modèles espace-état est la représentation explicite de la série chronologique comme un processus stochastique. Les procédures d'estimation des paramètres, dans un modèle espace-état (filtre de Kalman) sont très efficaces lorsque les séries temporelles sont d'une fréquence élevée. Cependant, il est souhaitable de convertir les modèles espace-état en modèle mixte lorsque les séries chronologiques sont de petite ou moyenne fréquence. Le coût et la complexité du calcul lié à la non utilisation du filtre de Kalman est supposé être meilleur.
D'un point de vue technique, les estimations résultantes des deux modèles sont équivalentes. Cela revient au fait que les deux approches sont basées sur la même fonction de vraisemblance dans les algorithmes et procédures d'estimations.
Via des simulations et des données réelles, utilisant le registre d'AVC du sud de Londres (SLSR), nous avons validé le cadre théorique proposé. Nous avons montré que les estimations basées sur le filtre de Kalman, spécifique aux modèles espace-état, sont quasi-équivalentes à la meilleure prédiction linéaire sans biais (BLUP), basée sur le RMLE dans un modèle mixte.
La méthodologie proposée a été appliquée pour la première fois sur des données d'AVC, et a permis de décrire la variation du score de Barthel moyen des patients victimes d'AVC enregistrés dans le SLSR (au cours des 23 années). Elle a également permis d'évaluer l'impact de la mise à jour des directives et conseils cliniques sur le diagnostic et la prise en charge des patients victime d'AVC, dans les 48 heures suivant l'apparition des symptômes.

Dans ce chapitre, nous nous sommes restreint au *"local level model"* pour illustrer et valider la méthodologie proposée. Comme perspectives, nous souhaitons élargir ce travail en appliquant les différents modèles étudiés théoriquement en détail dans ce chapitre, à d'autres sujets de santé publique, en l'occurrence l'AVC.
Par ce travail, nous offrons un cadre de modélisation optimal permettant l'utilisation des modèles à effets mixtes dans des situations souvent approchées par des modèles espace-état, mais ne nécessitant pas un tel cadre de modélisation complexe.

Conclusion générale et perspectives

Les résultats de ce travail de thèse se situent dans le cadre de la modélisation et l'analyse prédictive. Notre objectif a été de développer et d'affiner des techniques et modèles permettant de prédire avec précision des événements ou tendances futures, en particulier les risques et les conséquences des accidents vasculaires cérébraux (AVC). La thèse qui en résulte se compose de quatre chapitres.

Le chapitre 1 de cette thèse a fait l'objet d'une revue bibliographique. Nous avons tout d'abord passé en revue les différents modèle d'analyse prédictive et plus particulièrement, les modèles à effets mixtes (LMM). Les (LMM) ont été un outil central pour l'obtention des résultats clé de cette thèse. Ensuite, nous avons porté un grand intérêt aux méthodes de régularisation. Ces méthodes sont bien connues pour améliorer à la fois, la qualité de la prédiction et l'interprétablilité du modèle, surtout si la dimension des données dépasse largement la taille de l'échantillon. Finalement, nous avons clôturé ce chapitre par une réflexion et une synthèse globale sur les différentes méthodologies de construction de modèles de prédiction clinique.

Conscient des atouts de la régularisation et son rôle dans la modélisation prédictive, nous avons proposé dans le chapitre 2, une nouvelle méthode de régularisation pour le modèle de régression statistique en associant le cadre théorique des problèmes inverses avec celui des statistiques robustes. Nous avons montré à travers des simulations et des données réelles basées sur le registre des AVC du Sud de Londres (SLSR), que la méthode proposée est plus performante que d'autres méthodes (lasso, lasso adaptatif, ridge, elastic-net) sous différents scénarios et hypothèses à savoir, un grand conditionnement de la matrice de covariance associée aux données et une amplitude d'erreur variée. La méthode de régularisation proposée (Huber), a prouvé que l'approche problème inverse, couplé avec la méthodologie statistique robuste, a permis de caractériser une fonction de régularisation robuste et moins biaisée que d'autres méthodes existantes dans la littérature, et pourrait être utilisée avec assurance dans la pratique. En outre, elle offre une alternative améliorée à la méthode ridge, lasso, adaptive-lasso et l'elastic-net, qui se sont déjà révélées être des méthodes très performantes dans des études antérieures. Cette approche pourrait être étendue dans des travaux futurs, à la résolution des modèles à effets mixtes et à la réduction de dimension. Elle pourrait également être utilisée dans des problèmes d'apprentissage automatique statistique pour optimiser la fonction de perte [121]. Les résultats rapportés dans ce travail suggèrent que l'approche problème inverse pourrait être utile dans une grande variété de problèmes d'estimation statistique. Cette méthodologie n'est pas bien explorée dans la communauté statistique et une étude plus approfondie est cependant nécessaire. En recherche clinique, la méthodologie développée pourrait être appliquée comme un outil pour évaluer les effets bénéfiques des interventions fondées sur des données probantes et des structures de soins. En tant qu'outil de recherche, cela pourrait être utilisé pour tester de nouvelles interventions ou identifier des échantillons enrichis, ce qui réduirait le besoin d'essais randomisés contrôlés coûteux et souvent peu

pratiques. Cette stratégie d'enrichissement prédictive est importante pour la conception d'essais futurs, car elle permet le recrutement des patients les plus appropriés, permettant ainsi l'utilisation d'une population d'étude plus petite.

Dans le chapitre 3, nous avons développé une stratégie de modélisation permettant de prédire à long-terme (5 ans), les conséquences post-AVC telles que le déficit cognitif, la dépression ou la mortalité. Pour ce faire, nous avons étendu les travaux existants sur la méthodologie des courbes de récupération et ce d'un point de vue mathématique et applicatif. D'un point de vue mathématique, la régularisation est omniprésente. En effet, nous avons construit des « courbes de récupération régularisées » par le biais des modèles à effets mixtes au lieu des méthodes de régression classiques. Nous avons également évalué la performance de cette méthodologie en utilisant à la fois des métriques traditionnelles (discrimination et étalonnage) et non traditionnelles (utilité clinique). D'un point de vue applicatif, nous avons utilisé l'approche des « courbes de récupération régularisées » pour prédire à long-terme (5 ans), le déclin cognitif chez un individu après un AVC. Cette méthodologie est ainsi développée et évaluée, pour la première fois, pour des prévisions à long terme et d'autres conséquences d'AVC, en l'occurrence le déficit cognitif.
Les résultats obtenus par notre modèle, ont démontré que la méthodologie des courbes de récupération régularisées appliquées aux données d'AVC, s'est avérée être la méthode la plus flexible qui pourrait incorporer des informations de prédicteur longitudinal sans perte de qualité prédictive. Les mesures longitudinales sont plus appropriées que les mesures à des moments prédéfinis, et ajoutent une plus grande dimension aux prévisions. Les prédicteurs associés aux conséquences post-AVC sont généralement collectées régulièrement par des cliniciens. Par conséquent, la gestion et la prise en charge des patients survivant d'un AVC, peuvent être soutenus par des modèles de prédiction. Pour être utile et applicable à la pratique clinique, un modèle pronostique doit être non seulement validé, mais également facile à mettre en oeuvre (Wyatt et al., 1995). Autrement dit, il ne doit contenir que quelques variables facilement disponibles pour tous les patients. Les données devraient pouvoir être obtenues avec une grande fiabilité (Rowley et Fielding 1991). Dans cette perspective, les modèles de prédiction clinique doivent être affinés et améliorés. Les travaux futurs pourront envisager une évaluation plus approfondie du modèle présenté dans cette thèse, une étude de faisabilité pour amener ce modèle dans la pratique clinique et une application de la méthode des courbes de récupération régularisées à d'autres conséquences d'AVC.

Dans le chapitre 4, le profit tiré des modèles à effets mixtes (LMM) nous a poussé à comparer cette classe de modèle avec des modèles plus complexes et sophistiqués, en l'occurrence les modèles espace-état. Cette comparaison a permis de proposer un cadre de modélisation optimal permettant l'utilisation des modèles à effets mixtes dans des situations souvent approchées par des modèles espace-état, mais ne nécessitant pas un tel cadre de modélisation complexe. Cette méthodologie a été présentée pour la première fois d'une manière détaillée et explicite avec une application à l'AVC. Des motivations à la fois théoriques, numériques et pratiques ont conduit à la mise en oeuvre d'une stratégie générale pour transformer un modèle espace-état en un modèle mixte. Cette stratégie a ensuite été appliquée aux séries chronologiques structurelles, généralement approchées par un cadre espace-état. Nous avons illustré et validé le cadre théorique proposé via des simulations et des données réelles basées sur le registre d'AVC du Sud de Londres (SLSR). Les simulations et l'application aux données réelles, ont montré que les procédures d'estimation, spécifiques aux modèles espace-état (filtre de Kalman), sont équivalentes à celles d'un modèle à effets-mixtes (REML-BLUP). Il est cependant souhaitable de convertir les

modèles espace-état en modèle mixte lorsque les séries chronologiques sont de petite ou moyenne fréquence. Dans ce cas, le coût et la complexité du calcul lié à la non-utilisation du filtre de Kalman est supposé être meilleur. Néanmoins, les procédures d'estimation des paramètres, dans un modèle espace-état (filtre de Kalman) sont très efficaces lorsque les séries temporelles sont d'une fréquence élevée. Par ce travail, nous ne préconisons pas l'abandon de l'approche espace-état. En revanche, nous recommandons l'utilisation des modèles espace-état dans les cas où la méthodologie proposée n'est pas applicable c.-à-d., si la matrice de transition généré par le modèle est d'une structure complexe. Les chercheurs qui ont fréquemment besoin d'étudier des séries chronologiques avec des fréquences élevées devraient être encouragés à en savoir plus sur la méthodologie des séries chronologiques liées aux modèles d'espace d'état. Dans ce travail, nous nous sommes restreint au *"local level model"* au niveau des simulations et applications. Comme perspectives, nous souhaitons élargir ce travail en appliquant les différents modèles étudiés théoriquement à d'autres sujets de santé publique, en particulier l'AVC.

Bibliographie

[1] Antonio Carolei, Simona Sacco, Federica De Santis, and Carmine Marini. Epidemiology of stroke. *Clinical and Experimental hypertension*, 24(7-8) :479–483, 2002.

[2] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*, 58(1) :267–288, 1996.

[3] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 2006.

[4] Arthur E Hoerl and Robert W Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67, 1970.

[5] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society : series B (statistical methodology)*, 67(2) :301–320, 2005.

[6] Charles DA Wolfe, Siobhan L Crichton, Peter U Heuschmann, Christopher J McKevitt, Andre M Toschke, Andy P Grieve, and Anthony G Rudd. Estimates of outcomes up to ten years after stroke : analysis from the prospective south london stroke register. *PLoS Med*, 8(5) :e1001033, 2011.

[7] Kate Tilling, Jonathan AC Sterne, Anthony G Rudd, Thomas A Glass, Robert J Wityk, and Charles DA Wolfe. A new method for predicting recovery after stroke. *Stroke*, 32(12) :2867–2873, 2001.

[8] AM Toschke, K Tilling, AM Cox, AG Rudd, PU Heuschmann, and CDA Wolfe. Patient-specific recovery patterns over time measured by dependence in activities of daily living after stroke and post-stroke care : the south london stroke register (slsr). *European journal of neurology*, 17(2) :219–225, 2010.

[9] Abdel Douiri, Justin Grace, Shah-Jalal Sarker, Kate Tilling, Christopher McKevitt, Charles DA Wolfe, and Anthony G Rudd. Patient-specific prediction of functional recovery after stroke. *International Journal of Stroke*, 12(5) :539–548, 2017.

[10] Kathleen Tilling. *Statistical methods to study the incidence and outcome of stroke.* PhD thesis, King's College London (University of London), 2000.

[11] Rudolf Emil Kalman. When is a linear control system optimal ? 1964.

[12] George K Robinson et al. That blup is a good thing : the estimation of random effects. *Statistical science*, 6(1) :15–32, 1991.

[13] Youssef Hbid, Khaladi Mohamed, Charles DA Wolfe, and Abdel Douiri. Inverse problem approach to regularized regression models with application to predicting recovery after stroke. *Biometrical Journal*, 62(8) :1926–1938, 2020.

[14] Larry Wasserman. *All of statistics : a concise course in statistical inference*. Springer Science & Business Media, 2013.

[15] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. *Chapman & Hall*, 1989.

[16] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society : Series A (General)*, 135(3) :370–384, 1972.

[17] C Sidney Burrus, JA Barreto, and Ivan W Selesnick. Iterative reweighted least-squares design of fir filters. *IEEE Transactions on Signal Processing*, 42(11) :2926–2936, 1994.

[18] Edwin James George Pitman. Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the cambridge Philosophical society*, volume 32, pages 567–579. Cambridge University Press, 1936.

[19] Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3) :399–409, 1936.

[20] Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376) :817–823, 1981.

[21] Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2) :453–510, 1989.

[22] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

[23] Simon N Wood. *Generalized additive models : an introduction with R*. CRC press, 2017.

[24] Badi Baltagi. *Econometric analysis of panel data*. John Wiley & Sons, 2008.

[25] Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2) :399–433, 1919.

[26] H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3) :545–554, 1971.

[27] SR Searle, G Casella, and CE McCulloch. Variance components john wiley and sons. *Inc. New York*, 1992.

[28] Charles R Henderson. Estimation of genetic parameters. In *Biometrics*, volume 6, pages 186–187. International Biometric Soc 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210, 1950.

[29] Arthur S Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57(298) :369–375, 1962.

[30] Charles R Henderson, Oscar Kempthorne, Shayle R Searle, and CM Von Krosigk. The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15(2) :192–218, 1959.

[31] Herman O Hartley and Jon NK Rao. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1-2) :93–108, 1967.

[32] Richard Loree Anderson and Theodore Alfonso Bancroft. Statistical theory in research. Technical report, 1952.

[33] David A Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2) :383–385, 1974.

[34] Arunas Petras Verbyla. A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics*, 32(2) :227–230, 1990.

[35] Ole Barndorff-Nielsen. On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70(2) :343–365, 1983.

[36] Youngjo Lee, John A Nelder, and Yudi Pawitan. *Generalized linear models with random effects : unified analysis via H-likelihood*, volume 153. CRC Press, 2018.

[37] Arthur R Gilmour, Robin Thompson, and Brian R Cullis. Average information reml : an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, pages 1440–1450, 1995.

[38] DL Johnson and Robin Thompson. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *Journal of dairy science*, 78(2) :449–456, 1995.

[39] Ronald A Thisted. *Elements of statistical computing : Numerical computation*, volume 1. CRC Press, 1988.

[40] Nicholas T Longford. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74(4) :817–827, 1987.

[41] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) :1–22, 1977.

[42] Arthur P Dempster, Donald B Rubin, and Robert K Tsutakawa. Estimation in covariance components models. *Journal of the American Statistical Association*, 76(374) :341–353, 1981.

[43] Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421) :9–25, 1993.

[44] Walter W Stroup. *Generalized linear mixed models : modern concepts, methods and applications*. CRC press, 2012.

[45] Jiming Jiang. *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media, 2007.

[46] Yudi Pawitan. *In all likelihood : statistical modelling and inference using likelihood*. Oxford University Press, 2001.

[47] Xihong Lin and Daowen Zhang. Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the royal statistical society : Series b (statistical methodology)*, 61(2) :381–400, 1999.

[48] James Durbin and Siem Jan Koopman. *Time series analysis by state space methods*. Oxford university press, 2012.

[49] Andrew C Harvey. Forecasting, structural time series models and the kalman filter. 1990.

[50] Richard J Meinhold and Nozer D Singpurwalla. Understanding the kalman filter. *The American Statistician*, 37(2) :123–127, 1983.

[51] Ryan Tibshirani and Larry Wasserman. Sparsity and the lasso. *Statistical machine learning*, pages 1–15, 2015.

[52] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[53] Andrew W Moore. Cross-validation for detecting and preventing overfitting. *School of Computer Science Carneigie Mellon University*, 2001.

[54] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58 :267–288, 1994.

[55] Edward I George. The variable selection problem. *Journal of the American Statistical Association*, 95(452) :1304–1308, 2000.

[56] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1) :101, 2010.

[57] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity : the lasso and generalizations*. CRC press, 2015.

[58] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *Annals of statistics*, 32(2) :407–499, 2004.

[59] Berwin A Turlach. On algorithms for solving least squares problems under an l1 penalty or an l1 constraint. In *2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]*, pages 2572–2577. Citeseer, 2005.

[60] Michael R Osborne, Brett Presnell, and Berwin A Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3) :389–403, 2000.

[61] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482) :681–686, 2008.

[62] Yiyun Zhang, Runze Li, and Chih-Ling Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489) :312–323, 2010.

[63] Z John Daye and X Jessie Jeng. Shrinkage and model selection with correlated variables via weighted fusion. *Computational Statistics & Data Analysis*, 53(4) :1284–1298, 2009.

[64] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(1) :91–108, 2005.

[65] Mohamed Hebiri, Sara Van De Geer, et al. The smooth-lasso and other l1+ l2-penalized methods. *Electronic Journal of Statistics*, 5 :1184–1226, 2011.

[66] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1) :3–34, 2015.

[67] Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *Annals of applied statistics*, 1(2) :302–332, 2007.

[68] Anita J van der Kooij. Prediction accuracy and stability of regression with optimal scaling transformations. 2007.

[69] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49–67, 2006.

[70] LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2) :109–135, 1993.

[71] Wenjiang J Fu. Penalized regressions : the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3) :397–416, 1998.

[72] Howard D Bondell and Brian J Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1) :115–123, 2008.

[73] Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4) :373–384, 1995.

[74] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) the tripod statement. *Circulation*, 131(2) :211–219, 2015.

[75] Karel GM Moons, Patrick Royston, Yvonne Vergouwe, Diederick E Grobbee, and Douglas G Altman. Prognosis and prognostic research : what, why, and how ? *Bmj*, 338, 2009.

[76] Mark Woodward, Hugh Tunstall-Pedoe, and Sanne AE Peters. Graphics and statistics for cardiology : clinical prediction rules. *Heart*, 103(7) :538–545, 2017.

[77] Andrew J Vickers and Elena B Elkin. Decision curve analysis : a novel method for evaluating prediction models. *Medical Decision Making*, 26(6) :565–574, 2006.

[78] Mark Fitzgerald, Benjamin R Saville, and Roger J Lewis. Decision curve analysis. *Jama*, 313(4) :409–410, 2015.

[79] Graeme L Hickey, Stuart W Grant, Joel Dunning, and Matthias Siepe. Statistical primer : sample size and power calculations—why, when and how ? *European journal of cardio-thoracic surgery*, 54(1) :4–9, 2018.

[80] Mandip S Dhamoon, Yeseon Park Moon, Myunghee C Paik, Bernadette Boden-Albala, Tatjana Rundek, Ralph L Sacco, and Mitchell SV Elkind. Long-term functional recovery after first ischemic stroke : the northern manhattan study. *Stroke*, 40(8) :2805–2811, 2009.

[81] Mandip S Dhamoon, Leslie A McClure, Carole L White, Kamakshi Lakshminarayan, Oscar R Benavente, and Mitchell SV Elkind. Long-term disability after lacunar stroke : secondary prevention of small subcortical strokes. *Neurology*, 84(10) :1002–1008, 2015.

[82] Marion Fahey, Elise Crayton, Charles Wolfe, and Abdel Douiri. Clinical prediction models for mortality and functional outcome following ischemic stroke : a systematic review and meta-analysis. *PloS one*, 13(1) :e0185402, 2018.

[83] Sarah T Pendlebury and Peter M Rothwell. Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia : a systematic review and meta-analysis. *The Lancet Neurology*, 8(11) :1006–1018, 2009.

[84] Sheeba Rosewilliam, Carolyn Anne Roskell, and AD Pandyan. A systematic review and synthesis of the quantitative and qualitative evidence behind patient-centred goal setting in stroke rehabilitation. *Clinical rehabilitation*, 25(6) :501–514, 2011.

[85] Katharina Dworzynski, Gill Ritchie, Elisabetta Fenu, Keith MacDermott, and E Diane Playford. Rehabilitation after stroke : summary of nice guidance. *Bmj*, 346, 2013.

[86] Richard D Riley, Greta Ridley, Katrina Williams, Douglas G Altman, Jill Hayden, and HC De Vet. Prognosis research : toward evidence-based results and a cochrane methods group. *Journal of clinical epidemiology*, 60(8), 2007.

[87] Franco Angeleri, Vita Aurora Angeleri, Nicoletta Foschi, Salvatore Giaquinto, and Giuseppe Nolfe. The influence of depression, social activity, and family stress on functional outcome after stroke. *Stroke*, 24(10) :1478–1483, 1993.

[88] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. "mini-mental state" : a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3) :189–198, 1975.

[89] Mehool D Patel, Catherine Coshall, Anthony G Rudd, and Charles DA Wolfe. Cognitive impairment after stroke : clinical determinants and its associations with long-term stroke outcomes. *Journal of the American Geriatrics Society*, 50(4) :700–706, 2002.

[90] Rianne Oostenbrink, Karel GM Moons, Sacha E Bleeker, Henriëtte A Moll, and Diederick E Grobbee. Diagnostic research on routine care data : prospects and problems. *Journal of clinical epidemiology*, 56(6) :501–506, 2003.

[91] Thomas K Tatemichi, David W Desmond, Myunghee Paik, Miguel Figueroa, Toby I Gropen, Yaakov Stern, Mary Sano, R Remien, Janet BW Williams, Jay Preston Mohr, et al. Clinical determinants of dememtia related to stroke. *Annals of Neurology : Official Journal of the American Neurological Association and the Child Neurology Society*, 33(6) :568–575, 1993.

[92] Abdel Douiri, Anthony G. Rudd, and Charles D.A. Wolfe. Prevalence of poststroke cognitive impairment. *Stroke*, 2013.

[93] Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, and William B Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18) :1837–1847, 1998.

[94] Hans-Christoph Diener, Peter A Ringleb, and Pierre Savi. Clopidogrel for the secondary prevention of stroke. *Expert opinion on pharmacotherapy*, 6(5) :755–764, 2005.

[95] Walter N Kernan, Ralph I Horwitz, Lawrence M Brass, Catherine M Viscoli, and Kenneth JW Taylor. A prognostic system for transient ischemia or minor stroke. *Annals of internal medicine*, 114(7) :552–557, 1991.

[96] Walter N Kernan, Catherine M Viscoli, Lawrence M Brass, Robert W Makuch, Philip M Sarrel, Robin S Roberts, Michael Gent, Peter Rothwell, Ralph L Sacco, Ruei-Che Liu, et al. The stroke prognosis instrument ii (spi-ii) a clinical prediction instrument for patients with transient ischemia and nondisabling ischemic stroke. *Stroke*, 31(2) :456–462, 2000.

[97] Iris van Wijk, LJ Kappelle, Jan van Gijn, PJ Koudstaal, CL Franke, Marinus Vermeulen, Jan Willem Gorter, Ale Algra, LiLAC Study Group, et al. Long-term survival and vascular event risk after transient ischaemic attack or minor ischaemic stroke : a cohort study. *The Lancet*, 365(9477) :2098–2104, 2005.

[98] G Ntaios, M Faouzi, J Ferrari, W Lang, K Vemmos, and P Michel. An integer-based score to predict functional outcome in acute ischemic stroke : the astral score. *Neurology*, 78(24) :1916–1922, 2012.

[99] Kate Tilling, Jonathan AC Sterne, and Charles DA Wolfe. Estimation of the incidence of stroke using a capture-recapture model including covariates. *International journal of epidemiology*, 30(6) :1351–1359, 2001.

[100] Marieke Welten, Marlou LA de Kroon, Carry M Renders, Ewout W Steyerberg, Hein Raat, Jos WR Twisk, and Martijn W Heymans. Repeatedly measured predictors : a comparison of methods for prediction modeling. *Diagnostic and prognostic research*, 2(1) :1–10, 2018.

[101] John Aldrich et al. Fisher and regression. *Statistical Science*, 20(4) :401–417, 2005.

[102] Sir Austin Bradford Hill. The environment and disease : Association or causation ? *Proceedings of the Royal Society of Medicine*, 58(5) :295–300, 1965.

[103] Adrian Pagan and Aman Ullah. *Nonparametric Econometrics*. Themes in Modern Econometrics. *Cambridge University Press*, 1999.

[104] Lyle D Broemeling. *Bayesian analysis of linear models. New York : Marcel Dekker*, 1985.

[105] Peter BÂÂ¸hlmann and Sara van de Geer. *Statistics for High-Dimensional Data : Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2011.

[106] C Gauss, Friedrich, Davis, and H Charles. *Theory of the motion of the heavenly bodies moving about the sun in conic sections a translation of Gauss's "Theoria motus". Boston, Little, Brown and company*, 1857.

[107] A. C. Aitken. Iv on least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55 :42–48, 1936.

[108] Baltagi. *Violations of the Classical Assumptions. Springer Berlin Heidelberg*, 2008.

[109] P.J Huber. Robust statistics. *John Wiley and Sons, New York*, 1981.

[110] C.R. Vogel. Computational methods for inverse problems. *SIAM*, 2002.

[111] J Hadamard. Sur les problemes aux derviees partielles et leur signification physique. *Univ Princeton Bull*, 1902.

[112] Stuart Geman. Stochastic relaxation methods for image restoration and expert systems. In *Maximum-Entropy and Bayesian Methods in Science and Engineering*, pages 265–311. Springer Netherlands, 1988.

[113] A Douiri, M Schweiger, J Riley, and S Arridge. Local diffusion regularization method for optical tomography reconstruction by using robust statistics. *Optics Letters*, 2005.

[114] Markus Unger, Thomas Pock, Manuel Werlberger, and Horst Bischof. *A Convex Approach for Variational Super-Resolution.* Springer Berlin Heidelberg, 2010.

[115] Per Christian Hansen. The l-curve and its use in the numerical treatment of inverse problems. *Computational Inverse Problems in Electrocardiology*, 2001.

[116] Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. Cvxr : An r package for disciplined convex optimization. 2017.

[117] NK Karmarkar. Some comments on the significance of the new polynomial–time algorithm for linear programming. *AT&T Bell Laboratories, Murray Hill, New Jersey*, 1984.

[118] Osman Güler. Barrier functions in interior point methods. *Mathematics of Operations Research*, 21(4) :860–885, 1996.

[119] D.A Belsley, E Kuh, and R E Welsch. *Regression diagnostics : identifying influential data and sources of collinearity. New York : Wiley*, 1980.

[120] W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman & Hall/CRC Interdisciplinary Statistics. *Taylor & Francis*, 1995.

[121] T.Poggio and L.Rosasco. *Course slides and videos from MIT 9.520 : Statistical Learning Theory and Applications.* 2016.

[122] Louise M Allan, Elise N Rowan, Michael J Firbank, Alan J Thomas, Stephen W Parry, Tuomo M Polvikoski, John T O'Brien, and Raj N Kalaria. Long term incidence of dementia, predictors of mortality and pathological diagnosis in older stroke survivors. *Brain*, 134(12) :3716–3727, 2011.

[123] PM Rothwell, AJ Coull, MF Giles, SC Howard, LE Silver, LM Bull, SA Gutnikov, P Edwards, D Mant, CM Sackley, et al. Oxford vascular study. change in stroke incidence, mortality, case-fatality, severity, and risk factors in oxfordshire, uk from 1981 to 2004. *Oxford vascular study). Lancet*, 363(9425) :1925–1933, 2004.

[124] PM Rothwell, AJ Coull, LE Silver, JF Fairhead, MF Giles, CE Lovelock, JNE Redgrave, LM Bull, SJV Welch, FC Cuthbertson, et al. Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (oxford vascular study). *The Lancet*, 366(9499) :1773–1783, 2005.

[125] L Fratiglioni, LJ Launer, K Andersen, MM Breteler, JR Copeland, JF Dartigues, A Lobo, J Martinez-Lage, H Soininen, and A Hofman. Incidence of dementia and major subtypes in europe : A collaborative study of population-based cohorts. neurologic diseases in the elderly research group. *Neurology*, 54(11 Suppl 5) :S10–5, 2000.

[126] Makoto Suzuki, Yuko Sugimura, Sumio Yamada, Yoshitsugu Omori, Masaaki Miyamoto, and Jun-ichi Yamamoto. Predicting recovery of cognitive function soon after stroke : differential modeling of logarithmic and linear regression. *PloS one*, 8(1) :e53488, 2013.

[127] Amy J Ross, Perminder S Sachdev, Wei Wen, Henry Brodaty, Amy Joscelyne, and Lisa M Lorentz. Prediction of cognitive decline after stroke using proton magnetic resonance spectroscopy. *Journal of the neurological sciences*, 251(1-2) :62–69, 2006.

[128] Monica Saini, Chuen S Tan, Saima Hilal, YanHong Dong, Eric Ting, Mohammad K Ikram, Vijay K Sharma, and Christopher Chen. Computer tomography for prediction of cognitive outcomes after ischemic cerebrovascular events. *Journal of Stroke and Cerebrovascular Diseases*, 23(7) :1921–1927, 2014.

[129] Eugene Yee Hing Tang, Louise Robinson, and Blossom Christa Maree Stephan. Risk prediction models for post-stroke dementia. *Geriatrics*, 2(3) :19, 2017.

[130] Eugene YH Tang, Christopher I Price, Louise Robinson, Catherine Exley, David W Desmond, Sebastian Köhler, Julie Staals, Bonnie Yin Ka Lam, Adrian Wong, Vincent Mok, et al. Assessing the predictive validity of simple dementia risk models in harmonized stroke cohorts. *Stroke*, 51(7) :2095–2102, 2020.

[131] Nagaendran Kandiah, Russell Jude Chander, Xuling Lin, Aloysius Ng, Yen Yeong Poh, Chin Yee Cheong, Alvin Rae Cenina, and Pryseley Nkouibert Assam. Cognitive impairment after mild stroke : Development and validation of the signal 2 risk score. *Journal of Alzheimer's Disease*, 49(4) :1169–1177, 2016.

[132] Russell J Chander, Bonnie YK Lam, Xuling Lin, Aloysius YT Ng, Adrian PL Wong, Vincent CT Mok, and Nagaendran Kandiah. Development and validation of a risk score (change) for cognitive impairment after ischemic stroke. *Scientific reports*, 7(1) :1–11, 2017.

[133] WN Venables and BD Ripley. Random and mixed effects. In *Modern applied statistics with S*, pages 271–300. Springer, 2002.

[134] HM Hodkinson. Evaluation of a mental test score for assessment of mental impairment in the elderly. *Age and ageing*, 1(4) :233–238, 1972.

[135] Sarah T Pendlebury, Fiona C Cuthbertson, Sarah JV Welch, Ziyah Mehta, and Peter M Rothwell. Underestimation of cognitive impairment by mini-mental state examination versus the montreal cognitive assessment in patients with transient ischemic attack and stroke : a population-based study. *Stroke*, 41(6) :1290–1293, 2010.

[136] Merrill F Elias, Lisa M Sullivan, Ralph B D'Agostino, Penelope K Elias, Alexa Beiser, Rhoda Au, Sudha Seshadri, Charles DeCarli, and Philip A Wolf. Framingham stroke risk profile and lowered cognitive performance. *Stroke*, 35(2) :404–409, 2004.

[137] Karolina Piotrowicz, Wojciech Romanik, Anna Skalska, Barbara Gryglewska, Katarzyna Szczerbińska, Jarosław Derejczyk, Roger M Krzyżewski, Tomasz Grodzicki, and Jerzy Gąsowski. The comparison of the 1972 hodkinson's abbreviated mental test score (amts) and its variants in screening for cognitive impairment. *Aging clinical and experimental research*, 31(4) :561–566, 2019.

[138] Siobhan Laura Crichton. *Methods for Handling Missing Data in a Population Based Cohort Study*. PhD thesis, King's College London, 2016.

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

[140] SUTTHICHAI JITAPUNKUL, ISWERI PILLAY, and SHAH EBRAHIM. The abbreviated mental test : its use and validity. *Age and ageing*, 20(5) :332–336, 1991.

[141] Jennifer R Harvan and Valerie T Cotter. An evaluation of dementia screening in the primary care setting. *Journal of the American Academy of Nurse Practitioners*, 18(8) :351–360, 2006.

[142] Robert Perneczky, Stefan Wagenpfeil, Katja Komossa, Timo Grimmer, Janine Diehl, and Alexander Kurz. Mapping scores onto stages : mini-mental state examination and clinical dementia rating. *The American journal of geriatric psychiatry*, 14(2) :139–144, 2006.

[143] John Bamford, P Sandercock, Martin Dennis, C Warlow, and JJTL Burn. Classification and natural history of clinically identifiable subtypes of cerebral infarction. *The Lancet*, 337(8756) :1521–1526, 1991.

[144] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca : a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4) :695–699, 2005.

[145] Robert H Shumway, David S Stoffer, and David S Stoffer. *Time series analysis and its applications*, volume 3. Springer, 2000.

[146] John V Tsimikas and Johannes Ledolter. Mixed model representation of state space models : New smoothing results and their application to reml estimation. *Statistica Sinica*, pages 973–991, 1997.

[147] Terrance P Callanan and David A Harville. Some new algorithms for computing restricted maximum likelihood estimates of variance components. *Journal of Statistical Computation and Simulation*, 38(1-4) :239–259, 1991.

[148] P Jeffrey Harrison and Colin F Stevens. Bayesian forecasting. *Journal of the Royal Statistical Society : Series B (Methodological)*, 38(3) :205–228, 1976.

[149] Peter Colin Young, Cho Nam Ng, Kevin Lane, and David Parker. Recursive forecasting, smoothing and seasonal adjustment of non-stationary environmental data. *Journal of Forecasting*, 10(1-2) :57–89, 1991.

[150] Tommaso Proietti. Comparing seasonal components for structural time series models. *International Journal of Forecasting*, 16(2) :247–260, 2000.

[151] Andrew C Harvey and James Durbin. The effects of seat belt legislation on british road casualties : A case study in structural time series modelling. *Journal of the Royal Statistical Society : Series A (General)*, 149(3) :187–210, 1986.

[152] Michael G Kenward and James H Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, pages 983–997, 1997.

[153] Mike West and Jeff Harrison. *Bayesian forecasting and dynamic models*. Springer Science & Business Media, 2006.

[154] Simon Wood. Mixed gam computation vehicle with gcv/aic/reml smoothness estimation and gamms by reml/pql. *R package version*, pages 1–8, 2018.

[155] José Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, Siem Heisterkamp, Bert Van Willigen, and R Maintainer. Package 'nlme'. *Linear and nonlinear mixed effects models, version*, 3(1), 2017.

[156] Elizabeth E Holmes, Eric J Ward, and Kellie Wills. Marss : Multivariate autoregressive state-space models for analyzing time-series data. *R journal*, 4(1), 2012.

[157] Jacqui Wise. New clinical guidelines for stroke published, 2000.

[158] National Collaborating Centre for Chronic Conditions (Great Britain). Stroke : national clinical guideline for diagnosis and initial management of acute stroke and transient ischaemic attack (tia). Royal College of Physicians, 2008.

[159] TG Robinson. National clinical guidelines for stroke, 2009.

[160] Intercollegiate Stroke Working Party et al. *National clinical guideline for stroke*, volume 20083. Citeseer, 2012.

[161] Youngjo Lee and John A Nelder. Hierarchical generalised linear models : a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4) :987–1006, 2001.

[162] Michael G Kenward and James Carpenter. Multiple imputation : current perspectives. *Statistical methods in medical research*, 16(3) :199–218, 2007.

Appendice 2

Codes Chapitre 2

2.1 Simulations : cas ($p \gg n$)

```
# High dimensional data simulation (p>>n) ------------------------------------

# Author: Youssef Hbid
# Email: youssef.hbid@etu.upmc.fr

#references:
#Fu, Anqi, Balasubramanian Narasimhan, and Stephen Boyd. 2017. "CVXR: An R Package

#Loading packages
-------------------------------------
library(CVXR)
library(devtools)
library(magrittr)
library(dplyr)
library(glmnet)
library(pracma)
library(testthat)

set.seed(12333) #for reproducibility

# Generate data
 -------------------------------------
n <- 200 # Number of observations
p <- 400 # Number of predictors included in model

# Beta of example 4 (in this example, set: n=200 and p=400 )
    beta_true <- rep(c(0.85), times = c(p))

# Beta of example 5 (in this example, set: n=30 and p=60 )
    # beta_true<- rep(c(3,0,9), times = c(20,20,20))

# Beta of example 6 (in this example, set: n=100 and p=40 )
    # beta_true<- rep(c(1.5,2.5), times = c(200,200))

# number of simulations
```

```
r <- 30

# Initialization in order to store Mse values and beta's estimation
Mse_huber = rep(0,r)
Mse_elastic = rep(0,r)
Mse_lasso = rep(0,r)
Mse_yadlasso = rep(0,r)
Mse_ridge = rep(0,r)
bhuber = zeros(r,p)
Belastic = zeros(r,p)
Blasso = zeros(r,p)
Bridge = zeros(r,p)
Badlasso = zeros(r,p)

# Generate lamba sequences
lambda_vals <- 10^seq(1, -1, by = -.1)

# Set the condition number of X
cond <- 100

for (j in 1:r){

  # Generate a random normal matrix with mean 0 and var 1
  X <- matrix(stats::rnorm(n*p, 0, 1), nrow = n, ncol = p)

  # SVD decomposition in order to introduce a condition number
  S <- svd(X)
  D <- diag(S$d)
  D[D!=0] <- linspace(cond,1,min(n,p))
  U <- S$u
  V <- S$v
  V <- t(V) #V'
  X <- U %*% D %*% V #  X = U D V'

  # Error amplitude
  sigma <- 3

  # Split data into train (2/3) and test (1/3) sets
  train_rows <- sample(1:n, .66*n)
  x.train <- X[train_rows, ]
  x.test <- X[-train_rows, ]
  # generate true model
  y.train <- x.train %*% beta_true + rnorm(nrow(x.train), mean = 0, sd = sigma)
  # test
  y.test <- x.test %*% beta_true + rnorm(nrow(x.test), mean = 0, sd = sigma)

  # Declare beta parameter of dimension p (CVX syntax)
  beta <- Variable(p)
```

```
# Regularization methods
-------------------------------------

# Change the following parameters:
# M (1 or 1.5 or 1.8) in Huber method;
#alphae (0.25 or 0.5 or 0.75) in Elastic-net method;
#tau (0.5 or 1 or 1.5) in Adaptive lasso method

# Huber method (proposed method)
-------------------------------------

      # Huber Parameter
      M <- 1.5;
      # looping over grid of lambda
      for (i in seq_along(lambda_vals)) {
        lambda <- lambda_vals[i]
        # Objective function
        objhub <- sum((y.train - x.train %*% beta)^2) / (2 * nrow(x.train))
        + lambda*sum((huber(beta, M)))

        # Minimization of the objective function
        prob1 <- Problem(Minimize(objhub))
   result1 <- solve(prob1, solver = "ECOS",
   verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
      }

      # Extracting beta huber
       bhuber[j,]<- drop(result1$getValue(beta))
       yhuber = x.test%*%bhuber[j,]
      # Huber MSE
       Mse_huber[j]<- mean((y.test - yhuber)^2)

# Elastic-net
-------------------------------------

      # Elastic-net parameter
       alpha <- 0.5
      # looping over grid of lambda
      for (i in seq_along(lambda_vals)) {
        lambda <- lambda_vals[i]
        # Objective function
        obj <- (sum((y.train- x.train %*% beta)^2) / (2 * nrow(x.train)) +
          lambda * ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))

       # Minimization of the objective function
       prob <- Problem(Minimize(obj))
       result2 <- solve(prob, solver = "ECOS",
       verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
     }
    # Extracting Beta elastic-net
```

```
      Belastic[j,]<- drop(result2$getValue(beta))
      yelastic = x.test%*%Belastic[j,]
    # Elastic MSE
      Mse_elastic[j] <- mean((y.test - yelastic)^2)

# lasso
-----------------------------------

    # Elastic-net parameter
      alpha <- 1
    # looping over grid of lambda
    for (i in seq_along(lambda_vals)) {
      lambda <- lambda_vals[i]
      # Objective function
      obj <- (sum((y.train- x.train %*% beta)^2) / (2 * nrow(x.train)) +
              lambda * ((1 - alpha) * sum_squares(beta) / 2
            + alpha * p_norm(beta, 1)))

      # Minimization of the objective function
      prob <- Problem(Minimize(obj))
      result3 <- solve(prob, solver = "ECOS",
      verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
    }
   # Extracting Beta lasso
     Blasso[j,] <- drop(result3$getValue(beta))
     ylasso = x.test%*%Blasso[j,]
   # lasso MSE
     Mse_lasso[j] <- mean((y.test - ylasso)^2)

# Ridge
-----------------------------------

    # Ridge parameter
    alpha <- 0
    # looping over grid of lambda
    for (i in seq_along(lambda_vals)) {
     lambda <- lambda_vals[i]
     # Objective function
     obj <- (sum((y.train - x.train %*% beta)^2) / (2 * nrow(x.train)) +
         lambda* ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))

     # Minimization of the objective function
     prob <- Problem(Minimize(obj))
     result4 <- solve(prob, solver = "ECOS",
     verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
   }
  # Extracting Beta ridge
     Bridge[j,] <- drop(result4$getValue(beta))
     yridge <- x.test%*%Bridge[j,]
  # Ridge MSE
```

```
     Mse_ridge[j] <- mean((y.test - yridge)^2)

# Adaptive lasso
-------------------------------------

    # Adaptive lasso parameter
      tau = 1
   # l2 weights as suggested by Zou in adaptive lasso paper
   #(high  dimensional data)
      ridge1_cv <- cv.glmnet(x = x.train,
      y = y.train,type.measure = "mse",nfold = 10, alpha = 0)
      Bweight <- as.numeric(coef(ridge1_cv,
      s = ridge1_cv$lambda.min))[-1]

    # Adaptive lasso weights
      w=abs(Bweight)^(-tau)

    # looping over grid of lambda
      for (i in seq_along(lambda_vals)) {
      lambda <- lambda_vals[i]
     # Objective function
       obj <- sum((y.train- x.train %*% beta)^2) / (2 * nrow(x.train)) +
       lambda* sum_entries(w*abs(beta))

     # Minimization of the objective function
       prob <- Problem(Minimize(obj))
       result6 <- solve(prob, solver = "ECOS",
       verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
     }
  # Extracting Beta ad-lasso

    Badlasso[j,]<- drop(result6$getValue(beta))
    yadlasso <- x.test%*%Badlasso[j,]
  # Mse ad-lasso
    Mse_yadlasso[j]<- mean((y.test - yadlasso)^2)

  }

# Average MSE results
-------------------------------------
Average_Mse_ridge1 <-mean(Mse_ridge)
Average_Mse_lasso1 <- mean(Mse_lasso)
Average_Mse_elastic1 <- mean(Mse_elastic)
Average_Mse_yadlasso1 <- mean(Mse_yadlasso)
Average_Mse_huber1 <- mean(Mse_huber)

# Average of Non-zero coefficient
-------------------------------------
bhuber <- round(bhuber,2)
nzhuber <- apply(bhuber, 1, nnz)
```

```
nzhuber <- mean(nzhuber)
Blasso <- round(Blasso,2)
nzlasso <- apply(Blasso, 1, nnz)
nzlasso < -mean(nzlasso)
Belastic <- round(Belastic,2)
nzelastic <- apply( Belastic, 1, nnz)
nzelastic <- mean(nzelastic)
Badlasso <- round(Badlasso,2)
nzadlasso <- apply(Badlasso, 1, nnz)
nzadlasso <- mean(nzadlasso)

# Standard error of MSE
-----------------------------------
sd_ridge <- sd(Mse_ridge)/sqrt(length(Mse_ridge))
sd_lasso <- sd(Mse_lasso)/sqrt(length(Mse_lasso))
sd_yadlasso <- sd(Mse_yadlasso)/sqrt(length(Mse_yadlasso))
sd_elastic <- sd(Mse_elastic)/sqrt(length(Mse_elastic))
sd_huber <- sd(Mse_huber)/sqrt(length(Mse_huber))

# Organizing results
 -----------------------------------

AverageMSE <- round(c(Average_Mse_ridge1,Average_Mse_lasso1,
 Average_Mse_yadlasso1,Average_Mse_elastic1,Average_Mse_huber1), 3)
sd <- round(c(sd_ridge, sd_lasso, sd_yadlasso, sd_elastic, sd_huber), 3)
Avnonzerocoeff <- c("All",nzlasso,nzadlasso, nzelastic, nzhuber)
Methods <- c("Ridge","lasso","Ad-lasso","Elastic-net","Huber")

# Print results
 -----------------------------------
# Tables figuring in Simulation section :
Example4-table4/Example5-table5/Example6-table6

#(load Example4-table4.RData) # output Example4-table4.csv
table4 <- data.frame(Methods,AverageMSE,sd,Avnonzerocoeff)
table4

#(load Example5-table5.RData) # output Example5-table5.csv
table5 <- data.frame(Methods,AverageMSE,sd,Avnonzerocoeff)
table5

#(load Example6-table6.RData) # output Example6-table6.csv
table6 <- data.frame(Methods,AverageMSE,sd,Avnonzerocoeff)
table6
```

2.2 Simulations : cas $(n > p)$

```
# (n>p case)
-----------------------------------
```

```
#set.seed(1317) # for reproducibility of example 1
#set.seed(112) # for reproducibility of example 2
#set.seed(01234) # for reproducibility of example 3

# Number of observations
n <- 20
# Number of predictors included in model
p <- 8
# number of simulations
r <- 50

# Choose beta_true (Examples 1-2-3)

# Examples 1 (set n=20, p=8, sigma=3)  Example1-table1 in Simulation section
  beta_true <- c(3, 1.5, 0, 0, 2,0,0,0)

# Example 2 (set n=20, p=8, sigma=3)  Example2-table 2 in Simulation section
  #beta_true <- rep(c(0.80), times = c(p))

# Example 3 (set n=100, p=40 and sigma=15) Example3-table 3 in Simulation section
 #beta_true <- rep(c(0,2,0,2), times = c(10,10,10,10))

# Initialization in order to store Mse values and beta's estimation
Mse_huber = rep(0,r)
Mse_elastic = rep(0,r)
Mse_lasso = rep(0,r)
Mse_yadlasso = rep(0,r)
Mse_ridge = rep(0,r)
bhuber = zeros(r,p)
Belastic = zeros(r,p)
Blasso = zeros(r,p)
Bridge = zeros(r,p)
Badlasso = zeros(r,p)

# Generate lambda sequences
lambda_vals <- 10^seq(1, -1, by = -.1)

# Set the condition number of X
cond <- 100

  for (j in 1:r) {

    # Generate a random normal matrix with mean 0 and var 1
    X <- matrix(stats::rnorm(n*p, 0, 1), nrow = n, ncol = p)

    # SVD decomposition in order to introduce a condition number
    S <- svd(X)
    D <- diag(S$d)
    D[D!=0] <- linspace(cond,1,min(n,p))
    U <- S$u
```

```
    V <- S$v
    V <- t(V) #V'
    X <- U %*% D %*% V #  X = U D V'

    # Error amplitude
    sigma <- 3

    # Split data into train (2/3) and test (1/3) sets
    train_rows <- sample(1:n, .66*n)
    x.train <- X[train_rows, ]
    x.test <- X[-train_rows, ]
    # generate true model
    y.train <- x.train %*% beta_true + rnorm(nrow(x.train), mean = 0, sd = sigma)
    # test y
    y.test <- x.test %*% beta_true + rnorm(nrow(x.test), mean = 0, sd = sigma)
    # Declare beta parameter of dimension p (CVX syntax)
    beta <- Variable(p)

# Regularization methods:
-------------------------------------
# Change the following parameters:
# M (1 or 1.5 or 1.8) in Huber method; alphae (0.25 or 0.5 or 0.75) in Elastic-net

# Huber method (proposed method)
-------------------------------------

    # Huber Parameter
      M <- 1.8;
    # looping over grid of lambda
      for (i in seq_along(lambda_vals)) {
         lambda <- lambda_vals[i]

         # Objective function
          objhub<- sum((y.train - x.train %*% beta)^2) / (2 * nrow(x.train))
           + lambda*sum((huber(beta, M)))

         # Minimization of the objective function

         prob1 <- Problem(Minimize(objhub))
         result1 <- solve(prob1, solver = "ECOS",
         verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
      }

      # Extracting beta huber
      bhuber[j,] <- drop(result1$getValue(beta))
      yhuber = x.test%*%bhuber[j,]
      # Huber MSE
      Mse_huber[j] <- mean((y.test - yhuber)^2)
```

```
# Elastic-net method
 ------------------------------------
      # Elastic-net parameter
        alphae <- 0.75
        for (i in seq_along(lambda_vals)) {
           lambda <- lambda_vals[i]

       # Objective function
           obj <- (sum((y.train- x.train %*% beta)^2) / (2 * nrow(x.train)) +
     lambda * ((1 - alphae) * sum_squares(beta) / 2 + alphae * p_norm(beta, 1)))

           # Minimization of the objective function
           prob <- Problem(Minimize(obj))
           result2 <- solve(prob, solver = "ECOS",
            verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
        }

    # Extracting Beta elastic
      Belastic[j,] <- drop(result2$getValue(beta))
      yelastic = x.test%*%Belastic[j,]
   # Elastic MSE
      Mse_elastic[j] <- mean((y.test - yelastic)^2)

# Adaptive lasso
-------------------------------------

      # Adaptive lasso parameter
      tau=1.5
      # compute Least square solution
      objective <- Minimize(sum((y.train - x.train %*% beta)^2))
      problem   <- Problem(objective)
      res <- solve(problem)
      Bols <- drop(res$getValue(beta))
      # Ad-lasso weights
      w = abs( Bols +1/sqrt(nrow(x.train)))^(-tau)

      # looping over grid of lambda
      for (i in seq_along(lambda_vals)) {
        lambda <- lambda_vals[i]

        # Objective function
        obj <- sum((y.train- x.train %*% beta)^2) / (2 * nrow(x.train)) +
          lambda* sum_entries(w*abs(beta))

        # Minimization of the objective function
        prob <- Problem(Minimize(obj))
        result6 <- solve(prob, solver = "ECOS",
        verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
      }
```

```
      # Extracting Beta ad-lasso
      Badlasso[j,] <- drop(result6$getValue(beta))
      yadlasso = x.test%*%Badlasso[j,]
      # Mse ad-lasso
      Mse_yadlasso[j] <- mean((y.test - yadlasso)^2)


# lasso
-------------------------------------

      alpha <-1
      # looping over grid of lambda
      for (i in seq_along(lambda_vals)) {
        lambda <- lambda_vals[i]
      # Objective function
         obj <- (sum((y.train- x.train %*% beta)^2) / (2 * nrow(x.train)) +
         lambda * ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))
      # Minimization of the objective function
        prob <- Problem(Minimize(obj))
        result3 <- solve(prob, solver = "ECOS",
        verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
     }

    # Extracting Beta lasso

    Blasso[j,] <- drop(result3$getValue(beta))
    ylasso = x.test%*%Blasso[j,]

    # lasso MSE
    Mse_lasso[j] <- mean((y.test - ylasso)^2)



# Ridge
-------------------------------------
   # Ridge parameter
    alpha <- 0

   # looping over grid of lambda
    for (i in seq_along(lambda_vals)) {
      lambda <- lambda_vals[i]

      # Objective function
      obj <- (sum((y.train - x.train %*% beta)^2) / (2 * nrow(x.train)) +
          lambda* ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))

      # Minimization of the objective function
      prob <- Problem(Minimize(obj))
      result4 <- solve(prob, solver = "ECOS",
```

```
      verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
    }

  # Extracting Beta ridge
   Bridge[j,] <- drop(result4$getValue(beta))
   yridge = x.test%*%Bridge[j,]
  # Ridge MSE
   Mse_ridge[j] <- mean((y.test - yridge)^2)

  }


# Calculating Average MSE results
-------------------------------------
Mse_ridge1 <- mean(Mse_ridge)
Mse_lasso1 <- mean(Mse_lasso)
Mse_elastic1 <- mean(Mse_elastic)
Mse_yadlasso1 <- mean(Mse_yadlasso)
Mse_huber1 <- mean(Mse_huber)

# Calculating Average Non-zero coefficients
-------------------------------------
 bhuber <- round(bhuber,1)
 nzhuber <- apply(bhuber, 1, nnz)
 nzhuber <- mean(nzhuber)
 Blasso <- round(Blasso,1)
 nzlasso <- apply(Blasso, 1, nnz)
 nzlasso <- mean(nzlasso)
 Belastic <- round(Belastic,1)
 nzelastic <- apply( Belastic, 1, nnz)
 nzelastic <- mean(nzelastic)
 Badlasso <- round(Badlasso,1)
 nzadlasso <- apply(Badlasso, 1, nnz)
 nzadlasso <- mean(nzadlasso)

# Calculating Standard error of MSE
-------------------------------------
 sd_ridge <- sd(Mse_ridge)/sqrt(length(Mse_ridge))
 sd_lasso <- sd(Mse_lasso)/sqrt(length(Mse_lasso))
 sd_yadlasso <- sd(Mse_yadlasso)/sqrt(length(Mse_yadlasso))
 sd_elastic <- sd(Mse_elastic)/sqrt(length(Mse_elastic))
 sd_huber <- sd(Mse_huber)/sqrt(length(Mse_huber))

# Organizing results
-------------------------------------

 AverageMSE1 <- round(c(Mse_ridge1, Mse_lasso1,
 Mse_yadlasso1, Mse_elastic1, Mse_huber1), 2)
 sd1 <- round(c(sd_ridge, sd_lasso, sd_yadlasso, sd_elastic, sd_huber), 2)
 Avnonzerocoeff1 <- c("All",nzlasso,nzadlasso, nzelastic, nzhuber)
```

```
 Parameters1 <- c("alpha=0","alpha=1","gamma"=tau,"alpha"=alphae,"sigma"= M)

 AverageMSE2 <- round(c(Mse_yadlasso1, Mse_elastic1, Mse_huber1), 2)
 sd2 <- round(c(sd_yadlasso, sd_elastic, sd_huber), 2)
 Avnonzerocoeff2 <- c(nzadlasso, nzelastic, nzhuber)
 Parameters2 <- c("gamma"=tau,"alpha"=alphae,"sigma"= M)

 AverageMSE3 <- round(c(Mse_yadlasso1, Mse_elastic1, Mse_huber1), 2)
 sd3 <- round(c(sd_yadlasso, sd_elastic, sd_huber), 2)
 Avnonzerocoeff3 <- c(nzadlasso, nzelastic, nzhuber)
 Parameters3 <- c("gamma"=tau,"alpha"=alphae,"sigma"= M)

 AverageMSE <- c(AverageMSE1,AverageMSE2,AverageMSE3)
 Parameters <- c(Parameters1,Parameters2,Parameters3)
 sd <- c(sd1,sd2,sd3)
 Avnonzerocoeff <- c(Avnonzerocoeff1,Avnonzerocoeff2,Avnonzerocoeff3)
 Methods <- c("Ridge","lasso","Ad-lasso","Elastic-net","Huber",
 "Ad-lasso","Elastic-net","Huber","Ad-lasso","Elastic-net","Huber")

# Print results
-------------------------------------
# Tables figuring in Simulation section:
(Example1-Table1; Example2-Table2; Example3-Table3)
-----------------------------------------------

 #(load Example1-table1.RData) # output Example1-table1.csv
 table1 <- data.frame(Methods,Parameters, AverageMSE,sd,Avnonzerocoeff)
 table1

 #(load Example2-table2.RData) # output Example2-table2.csv
 table2 <- data.frame(Methods,Parameters, AverageMSE,sd,Avnonzerocoeff)
 table2

 #(load Example3-table3.RData) # output Example3-table3.csv
 table3 <- data.frame(Methods,Parameters, AverageMSE,sd,Avnonzerocoeff)
 table3
```

2.3 Stroke data (12 weeks)

```
 # Stroke data (12 weeks)
 -------------------------------------

# DATA SHARING STATEMENT
 -------------------------------------
# The study and its consent procedure were approved by the ethics
#committees of Guy's and St Thomas' Hospital Trust,
# King's College Hospital, Queen's Square, and Westminster Hospital.
#Consent for data sharing was not obtained from study participants.
# The research team will consider reasonable requests for sharing of
```

```
# anonymised patient level data.

# For this purpose, we generate pseudo data that mimic the Real
#stroke data: We produce the pseudo data which are
# comparable to the original data in size and structure
#(distribution, mean and standard deviation).
# Pseudo data, compared to real stroke data
#(which we used in the paper) produces similar conclusions.

rm(list = ls())


# Load library
-------------------------------------
library(dplyr)
library(klaR)
library(nlme)
library(lme4)
library(MuMIn)
library(tidyr)
library(foreign)
library(tidyverse)
require(compiler)
require(parallel)
library(ggplot2)
library(CVXR)
library(devtools)
library(glmnet)
library(pracma)
library(testthat)
library(plotrix)

# for reproducibility
set.seed(33333)

# Load data
-------------------------------------

data <- read.delim("pseudo-data-12.txt")

# Recoding predictors

data$ethgrp <- as.factor(data$ethgrp)
data$sex <- as.factor(data$sex)
data$subtype <- as.factor(data$subtype)
data$batotw1 <- as.numeric(data$batotw1)
data$batotw12 <- as.numeric(data$batotw12)
data$age <- as.numeric(data$age)
data$nihtot <- as.numeric(data$nihtot)
data$glas_cs <- as.numeric(data$glas_cs)
```

```
# Defining X

X <- model.matrix(batotw12 ~ sex + ethgrp+glas_cs+
subtype+age+batotw1+nihtot+0, data = data)
X <- X[,-1]

# Defining Y
y <- matrix(data[, "batotw12"], ncol = 1)

# number of simulations

r=500

# Initialization in order to store Mse values and beta's estimation
-------------------------------------
p=ncol(X)
Mse_huber = rep(0,r)
Mse_elastic = rep(0,r)
Mse_lasso = rep(0,r)
Mse_yadlasso = rep(0,r)
Mse_ridge = rep(0,r)
bhuber=zeros(r,p)
Belastic=zeros(r,p)
Blasso=zeros(r,p)
Bridge=zeros(r,p)
Badlasso=zeros(r,p)

# Analysis
-------------------------------------

for(j in 1:r){

# Split data into train (2/3) and test (1/3) sets

  train_rows <- sample(nrow(X), .66*nrow(X))
  x.train <- X[train_rows, ]
  x.test <- X[-train_rows, ]
  y.train <- y[train_rows]
  y.test <- y[-train_rows]
  n = nrow(x.train)

# Lambda sequence
  lambda_vals <- 10^seq(1, -1, by = -.1)


# Huber method
 -------------------------------------

  beta <- Variable(p)
```

```
  M <- 1.5
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    objhub <- sum((y.train - x.train %*% beta)^2) / (2 *n)
     + lambda*sum((huber(beta, M)))
    prob1 <- Problem(Minimize(objhub))
    result1 <- solve(prob1, solver = "ECOS",
     verbose = TRUE, ABSTOL = 1e-3, RELTOL = 1e-2)
  }

  bhuber[j,] <- drop(result1$getValue(beta))
  yhuber = x.test%*%bhuber[j,]
  Mse_huber[j] <- mean((y.test - yhuber)^2)

# Elastic-net
-------------------------------------
  alpha <- 0.5
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    obj <- (sum((y.train- x.train %*% beta)^2) / (2 * n) +
         lambda * ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))
    prob <- Problem(Minimize(obj))
    result2 <- solve(prob, solver = "ECOS",
    verbose = TRUE, ABSTOL = 1e-3, RELTOL = 1e-2)
  }
  Belastic[j,] <- drop(result2$getValue(beta))
  yelastic = x.test%*%Belastic[j,]
  Mse_elastic[j] <- mean((y.test - yelastic)^2)

# lasso
-------------------------------------

  alpha <- 1
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    obj <- (sum((y.train- x.train %*% beta)^2) / (2 *n) +
       lambda * ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))
    prob <- Problem(Minimize(obj))
    result3 <- solve(prob, solver = "ECOS",
    verbose = TRUE, ABSTOL = 1e-3, RELTOL = 1e-2)
  }
  Blasso[j,] <- drop(result3$getValue(beta))
  ylasso = x.test%*%Blasso[j,]
  Mse_lasso[j] <- mean((y.test - ylasso)^2)

# Ridge
 -------------------------------------

  alpha <- 0
  for (i in seq_along(lambda_vals)) {
```

```
    lambda <- lambda_vals[i]
    obj <- (sum((y.train - x.train %*% beta)^2) / (2 * nrow(x.train)) +
         lambda* ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))
    prob <- Problem(Minimize(obj))
    result4 <- solve(prob, solver = "ECOS",
    verbose = TRUE)
  }
  Bridge[j,] <- drop(result4$getValue(beta))
  yridge = x.test%*%Bridge[j,]
  Mse_ridge[j] <- mean((y.test - yridge)^2)

# Adaptive lasso
 ------------------------------------

  # step1 compute LS solution

   objective <- Minimize(sum((y.train - x.train %*% beta)^2))
  problem   <- Problem(objective)
  res <- solve(problem)
  Bols <- drop(res$getValue(beta))
  # Adaptive lasso parameter

  tau=1

  # Adaptive weights
   ------------------------------------

  w=abs( Bols +1/sqrt(nrow(x.train)))^(-tau)
  for (i in seq_along(lambda_vals)) {
  lambda <- lambda_vals[i]
  obj <-sum((y.train- x.train %*% beta)^2) / (2 * nrow(x.train))
   + lambda* sum_entries(w*abs(beta))
    prob <- Problem(Minimize(obj))
    result6 <- solve(prob, solver = "ECOS",
     verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
  }
  Badlasso[j,] <- drop(result6$getValue(beta))
  yadlasso = x.test%*%Badlasso[j,]
  Mse_yadlasso[j] <- mean((y.test - yadlasso)^2)

}

# MSE results
 ------------------------------------

Mse_ridge1 <- mean(Mse_ridge)
Mse_yadlasso1 <- mean(Mse_yadlasso)
Mse_lasso1 <- mean(Mse_lasso)
Mse_elastic1 <- mean(Mse_elastic)
Mse_huber1 <- mean(Mse_huber)
```

```
# Parameters estimate
 ------------------------------------

bhuber <- round(bhuber,3)
Blasso <- round(Blasso,3)
Belastic <- round(Belastic,3)
Badlasso <- round(Badlasso,3)
beta_huber <- apply(bhuber, 2, mean)
beta_lasso <- apply(Blasso, 2, mean)
beta_elastic <- apply(Belastic, 2, mean)
beta_ridge <- apply(Bridge, 2, mean)
beta_adlasso <- apply(Badlasso, 2, mean)

# Standard error MSE
-------------------------------------

sd_ridge <- sd(Mse_ridge)/sqrt(length(Mse_ridge))
sd_lasso <- sd(Mse_lasso)/sqrt(length(Mse_lasso))
sd_yadlasso <- sd(Mse_yadlasso)/sqrt(length(Mse_yadlasso))
sd_elastic <- sd(Mse_elastic)/sqrt(length(Mse_elastic))
sd_huber <- sd(Mse_huber)/sqrt(length(Mse_huber))

# Standard error parameters
-------------------------------------

SdBridge <- std.error(Bridge)
SdBlasso <- std.error(Blasso)
SdBadlasso <- std.error(Badlasso)
SdBelastic <- std.error(Belastic)
SdBhuber <- std.error(bhuber)

# Print table of Stroke data (12 weeks)
 ------------------------------------
# Organizing results in table
a <- round(cbind(beta_ridge,SdBridge, beta_lasso,
 SdBlasso, beta_adlasso,SdBadlasso,beta_elastic,SdBelastic, beta_huber,SdBhuber),2)

  b <- round(cbind(Mse_ridge1,sd_ridge, Mse_lasso1, sd_lasso, Mse_yadlasso1,
  sd_yadlasso, Mse_elastic1, sd_elastic, Mse_huber1,sd_huber), 3)

table_stroke12 <- rbind(a,b)
rownames(table_stroke12) <- c("sex ","ethgroup1","ethgroup2","ethgroup3",
"ethgroup4","glas_cs","TACI","PACI","POCI","LACI","age","batotw1",
"nihtot","Average MSE (sd)")

# Print table_stroke12
 ------------------------------------
table_stroke12  #(load table_stroke12.RData) # output :table_stroke12.csv file
```

2.4 Stroke data (26 weeks)

```
# Stroke data (26 weeks)
 ------------------------------------

# for reproducibility

set.seed(2222222)

# Load data
------------------------------------

data <- read.delim("pseudo-data-26.txt")

# Recoding predictors

data$ethgrp <- as.factor(data$ethgrp)
data$sex <- as.factor(data$sex)
data$subtype <- as.factor(data$subtype)
data$batotw1 <- as.numeric(data$batotw1)
data$batotw26 <- as.numeric(data$batotw26)
data$age <- as.numeric(data$age)
data$nihtot <- as.numeric(data$nihtot)
data$glas_cs <- as.numeric(data$glas_cs)

# Defining X
X <- model.matrix(batotw26 ~ sex + ethgrp+glas_cs+subtype+age+
batotw1+nihtot+0, data = data)
X <- X[,-1]

# Defining Y
y <- matrix(data[, "batotw26"], ncol = 1)

# number simumations for bootstrap
r <- 500

# Initialization in order to store Mse values and beta's estimation
------------------------------------
p = ncol(X)
Mse_huber = rep(0,r)
Mse_elastic = rep(0,r)
Mse_lasso = rep(0,r)
Mse_yadlasso = rep(0,r)
Mse_ridge = rep(0,r)
bhuber = zeros(r,p)
Belastic = zeros(r,p)
Blasso = zeros(r,p)
Bridge = zeros(r,p)
```

```
Badlasso = zeros(r,p)

# Analysis
-----------------------------------

for(j in 1:r){

  # Split data into train (2/3) and test (1/3) sets

  train_rows <- sample(nrow(X), .66*nrow(X))
  x.train <- X[train_rows, ]
  x.test <- X[-train_rows, ]
  y.train <- y[train_rows]
  y.test <- y[-train_rows]
  n = nrow(x.train)
  # Lambda sequence
  lambda_vals <- 10^seq(1, -1, by = -.1)


# Huber method
 -----------------------------------
  beta <- Variable(p)
  M <- 1.5
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    objhub <- sum((y.train - x.train %*% beta)^2) / (2 *n)
    + lambda*sum((huber(beta, M)))
    prob1 <- Problem(Minimize(objhub))
    result1 <- solve(prob1, solver = "ECOS",
    verbose = TRUE, ABSTOL = 1e-3, RELTOL = 1e-2)
  }

  bhuber[j,] <- drop(result1$getValue(beta))
  yhuber = x.test%*%bhuber[j,]
  Mse_huber[j] <- mean((y.test - yhuber)^2)


# Elastic-net
-----------------------------------
  alpha <- 0.5
  for (i in seq_along(lambda_vals)) {
  lambda <- lambda_vals[i]

  obj <- (sum((y.train- x.train %*% beta)^2) / (2 * n)
 + lambda * ((1 - alpha) * sum_squares(beta) / 2 +alpha * p_norm(beta, 1)))

  prob <- Problem(Minimize(obj))
  result2 <- solve(prob, solver = "ECOS",
    verbose = TRUE,  ABSTOL = 1e-3, RELTOL = 1e-2)
  }
```

```
  Belastic[j,] <- drop(result2$getValue(beta))
  yelastic = x.test%*%Belastic[j,]
  Mse_elastic[j] <- mean((y.test - yelastic)^2)

# lasso
------------------------------------

  alpha <- 1
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    obj <- (sum((y.train- x.train %*% beta)^2) / (2 *n)
    + lambda * ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))
    prob <- Problem(Minimize(obj))
    result3 <- solve(prob, solver = "ECOS",
     verbose = TRUE, ABSTOL = 1e-3, RELTOL = 1e-2)
  }
  Blasso[j,] <- drop(result3$getValue(beta))
  ylasso = x.test%*%Blasso[j,]
  Mse_lasso[j] <- mean((y.test - ylasso)^2)

# Ridge
------------------------------------
  alpha <- 0
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    obj <- (sum((y.train - x.train %*% beta)^2) / (2 * nrow(x.train))
    +  lambda* ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))
    prob <- Problem(Minimize(obj))
    result4 <- solve(prob, solver = "ECOS", verbose = TRUE)
  }
  Bridge[j,] <- drop(result4$getValue(beta))
  yridge = x.test%*%Bridge[j,]
  Mse_ridge[j] <- mean((y.test - yridge)^2)


# Adaptive lasso
------------------------------------
   # step1 compute LS solution
  objective <- Minimize(sum((y.train - x.train %*% beta)^2))
  problem   <- Problem(objective)
  res <- solve(problem)
  Bols <- drop(res$getValue(beta))
  # Adaptive lasso parameter
  tau = 1
  # Adaptive weights
  w = abs( Bols +1/sqrt(nrow(x.train)))^(-tau)
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    obj <- sum((y.train- x.train %*% beta)^2) / (2 * nrow(x.train))
    + lambda* sum_entries(w*abs(beta))
```

```
    prob <- Problem(Minimize(obj))
    result6 <- solve(prob, solver = "ECOS",
    verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
  }
  Badlasso[j,] <- drop(result6$getValue(beta))
  yadlasso = x.test%*%Badlasso[j,]
  Mse_yadlasso[j] <- mean((y.test - yadlasso)^2)
}

# MSE results
------------------------------------
Mse_ridge1 <- mean(Mse_ridge)
Mse_yadlasso1 <- mean(Mse_yadlasso)
Mse_lasso1 <- mean(Mse_lasso)
Mse_elastic1 <- mean(Mse_elastic)
Mse_huber1 <- mean(Mse_huber)


# Parameters estimate
------------------------------------

bhuber <- round(bhuber,3)
Blasso <- round(Blasso,3)
Belastic <- round(Belastic,3)
Badlasso <- round(Badlasso,3)
beta_huber <- apply(bhuber, 2, mean)
beta_lasso <- apply(Blasso, 2, mean)
beta_elastic <- apply(Belastic, 2, mean)
beta_ridge <- apply(Bridge, 2, mean)
beta_adlasso <- apply(Badlasso, 2, mean)

# Standard error MSE
------------------------------------
sd_ridge <- sd(Mse_ridge)/sqrt(length(Mse_ridge))
sd_lasso <- sd(Mse_lasso)/sqrt(length(Mse_lasso))
sd_yadlasso <- sd(Mse_yadlasso)/sqrt(length(Mse_yadlasso))
sd_elastic <- sd(Mse_elastic)/sqrt(length(Mse_elastic))
sd_huber <- sd(Mse_huber)/sqrt(length(Mse_huber))

# Standard error parameters
------------------------------------
SdBridge <- std.error(Bridge)
SdBlasso <- std.error(Blasso)
SdBadlasso <- std.error(Badlasso)
SdBelastic <- std.error(Belastic)
SdBhuber <- std.error(bhuber)

# Table of Stroke data (26 weeks)
------------------------------------
# Organizing results in table
```

```
a <- round(cbind(beta_ridge,SdBridge, beta_lasso,
SdBlasso, beta_adlasso,SdBadlasso,
beta_elastic,SdBelastic, beta_huber,SdBhuber), 2)
b <- round(cbind(Mse_ridge1,sd_ridge,
Mse_lasso1,sd_lasso, Mse_yadlasso1,sd_yadlasso,
 Mse_elastic1, sd_elastic, Mse_huber1,sd_huber), 3)
table_stroke26 <- rbind(a,b)
rownames(table_stroke26) < c("sex","ethgroup1","ethgroup2","ethgroup3 ","ethgroup4
"TACI ","PACI ","POCI","LACI ","age ","batotw1",
"nihtot","Average MSE (sd)")

# Print table_stroke26
-------------------------------------
table_stroke26  #(load table_stroke26.RData) # output: table_stroke26.csv file
```

2.5 Stroke data (52 weeks)

```
 # Stroke data (52 weeks)
 -------------------------------------
# Load data
-------------------------------------
data <- read.delim("pseudo-data-52.txt")

# Recoding predictors

data$ethgrp <- as.factor(data$ethgrp)
data$sex <- as.factor(data$sex)
data$subtype <- as.factor(data$subtype)
data$batotw1 <- as.numeric(data$batotw1)
data$batotw52 <- as.numeric(data$batotw52)
data$age <- as.numeric(data$age)
data$nihtot <- as.numeric(data$nihtot)
data$glas_cs <- as.numeric(data$glas_cs)

# Defining X

X <- model.matrix(batotw52 ~ sex + ethgrp+glas_cs+subtype+age+batotw1+
nihtot+0, data = data)
X <- X[,-1]

# Defining Y
y <- matrix(data[, "batotw52"], ncol = 1)

# number simulations  for bootstrap

r <- 500

# Initialization in order to store Mse values and beta's estimation ---------------

p = ncol(X)
```

```
Mse_huber = rep(0,r)
Mse_elastic = rep(0,r)
Mse_lasso = rep(0,r)
Mse_yadlasso = rep(0,r)
Mse_ridge = rep(0,r)
bhuber = zeros(r,p)
Belastic = zeros(r,p)
Blasso = zeros(r,p)
Bridge = zeros(r,p)
Badlasso = zeros(r,p)

# Analysis
-------------------------------------

for(j in 1:r){

  # Split data into train (2/3) and test (1/3) sets
  train_rows <- sample(nrow(X), .66*nrow(X))
  x.train <- X[train_rows, ]
  x.test <- X[-train_rows, ]
  y.train <- y[train_rows]
  y.test <- y[-train_rows]
  n = nrow(x.train)
  # Lambda sequence
  lambda_vals <- 10^seq(1, -1, by = -.1)


# Huber method
-------------------------------------

  beta <- Variable(p)
  M <- 1.5
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    objhub <- sum((y.train - x.train %*% beta)^2) / (2 *n)
    + lambda*sum((huber(beta, M)))
    prob1 <- Problem(Minimize(objhub))
    result1 <- solve(prob1, solver = "ECOS",
    verbose = TRUE, ABSTOL = 1e-3, RELTOL = 1e-2)
  }

  bhuber[j,] <- drop(result1$getValue(beta))
  yhuber = x.test%*%bhuber[j,]
  Mse_huber[j] <- mean((y.test - yhuber)^2)


# Elastic-net
-------------------------------------

  alpha <- 0.5
```

```
  for (i in seq_along(lambda_vals)) {
      lambda <- lambda_vals[i]
      obj <- (sum((y.train- x.train %*% beta)^2) / (2 * n)
      + lambda * ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))
      prob <- Problem(Minimize(obj))
      result2 <- solve(prob, solver = "ECOS",
      verbose = TRUE, ABSTOL = 1e-3, RELTOL = 1e-2)
     }
  Belastic[j,] <- drop(result2$getValue(beta))
  yelastic = x.test%*%Belastic[j,]
  Mse_elastic[j] <- mean((y.test - yelastic)^2)

# lasso
-------------------------------------

  alpha <- 1
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    obj <- (sum((y.train- x.train %*% beta)^2) / (2 *n)
    +  lambda * ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))
    prob <- Problem(Minimize(obj))
    result3 <- solve(prob, solver = "ECOS",
    verbose = TRUE, ABSTOL = 1e-3, RELTOL = 1e-2)
  }
  Blasso[j,] <- drop(result3$getValue(beta))
  ylasso = x.test%*%Blasso[j,]
  Mse_lasso[j] <- mean((y.test - ylasso)^2)

# Ridge
-------------------------------------

  alpha <- 0
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    obj <- (sum((y.train - x.train %*% beta)^2) / (2 * nrow(x.train)) +
    lambda* ((1 - alpha) * sum_squares(beta) / 2 + alpha * p_norm(beta, 1)))
    prob <- Problem(Minimize(obj))
    result4 <- solve(prob, solver = "ECOS", verbose = TRUE)
  }
  Bridge[j,] <- drop(result4$getValue(beta))
  yridge = x.test%*%Bridge[j,]
  Mse_ridge[j] <- mean((y.test - yridge)^2)

# Adaptive lasso
-------------------------------------

  # step1 compute LS solution
  objective <- Minimize(sum((y.train - x.train %*% beta)^2))
  problem   <- Problem(objective)
  res <- solve(problem)
```

```
  Bols <- drop(res$getValue(beta))
  # Adaptive lasso parameter
  tau = 1
  # Adaptive weights
  w = abs( Bols +1/sqrt(nrow(x.train)))^(-tau)
  for (i in seq_along(lambda_vals)) {
    lambda <- lambda_vals[i]
    obj <- sum((y.train- x.train %*% beta)^2) / (2 * nrow(x.train))
    + lambda* sum_entries(w*abs(beta))
    prob <- Problem(Minimize(obj))
    result6 <- solve(prob, solver = "ECOS",
     verbose = TRUE, ABSTOL = 1e-12, RELTOL = 1e-10)
  }
  Badlasso[j,] <- drop(result6$getValue(beta))
  yadlasso = x.test%*%Badlasso[j,]
  Mse_yadlasso[j] <- mean((y.test - yadlasso)^2)
}

# MSE results
-------------------------------------
Mse_ridge1 <- mean(Mse_ridge)
Mse_yadlasso1 <- mean(Mse_yadlasso)
Mse_lasso1 <- mean(Mse_lasso)
Mse_elastic1 <- mean(Mse_elastic)
Mse_huber1 <- mean(Mse_huber)

# Parameters estimate
-------------------------------------

bhuber <- round(bhuber,3)
Blasso <- round(Blasso,3)
Belastic <- round(Belastic,3)
Badlasso <- round(Badlasso,3)
beta_huber <- apply(bhuber, 2, mean)
beta_lasso <- apply(Blasso, 2, mean)
beta_elastic <- apply(Belastic, 2, mean)
beta_ridge <- apply(Bridge, 2, mean)
beta_adlasso <- apply(Badlasso, 2, mean)

# Standard error MSE
 -------------------------------------
sd_ridge <- sd(Mse_ridge)/sqrt(length(Mse_ridge))
sd_lasso <- sd(Mse_lasso)/sqrt(length(Mse_lasso))
sd_yadlasso <- sd(Mse_yadlasso)/sqrt(length(Mse_yadlasso))
sd_elastic <- sd(Mse_elastic)/sqrt(length(Mse_elastic))
sd_huber <- sd(Mse_huber)/sqrt(length(Mse_huber))

# Standard error parameters
-------------------------------------
SdBridge <- std.error(Bridge)
```

```
SdBlasso <- std.error(Blasso)
SdBadlasso <- std.error(Badlasso)
SdBelastic <- std.error(Belastic)
SdBhuber <- std.error(bhuber)

# Table of Stroke data (52 weeks)
-----------------------------------
# Organizing results in table
a <- round(cbind(beta_ridge,SdBridge,
beta_lasso,SdBlasso, beta_adlasso,SdBadlasso,beta_elastic,
SdBelastic, beta_huber,SdBhuber), 2)
b <- round(cbind(Mse_ridge1,sd_ridge,
Mse_lasso1,sd_lasso, Mse_yadlasso1,sd_yadlasso,
 Mse_elastic1, sd_elastic, Mse_huber1,sd_huber), 3)
table_stroke52 <- rbind(a,b)
rownames(table_stroke52) <- c("sex","ethgroup1","ethgroup2",
"ethgroup3 ","ethgroup4","glas_cs","TACI","PACI ",
"POCI ","LACI ","age","batotw1 ","nihtot","Average MSE (sd)")

# Print table_stroke52
-----------------------------------
table_stroke52  #(load table_stroke52.RData) # output: table_stroke52.csv file
```

Appendice 3

Codes Chapitre 3

3.1 Construction du modèle | Figure 3.1

```
#Load packages
library(tidyverse)
library(mice)
library(lattice)
library(ggplot2)
library(lme4)
require(caTools)
library(scales)
library(caret)
library(klaR)
library(nlme)
library(boot)
library(MuMIn)
library(tidyr)
library(dplyr)
require(lazyeval)
library(foreign)
require(compiler)
require(parallel)
library(magrittr)
require("splines")

## load data
-------------------------------------
data <- read.csv("~/2018_data_final.csv")

## Socio-demographic factors
-------------------------------------
data$ethcat<-as.factor(data$ethcat)
data$sex<-as.factor(data$sex)
data$age<-as.numeric(data$age)

# Time
-------------------------------------
data$ time<-as.numeric(data$time)
```

```
# Observation
-------------------------------------
data$mtotfollowup<-as.numeric(data$mtotfollowup)

# Stroke subtype: Structural/modified
-------------------------------------
data$subtype<-as.factor(data$subtype)  # stroke subtype
data$ dysphas<-as.factor(data$dysphas) # dysphasia
data$ rfptia<-as.factor(data$ rfptia) # TIA
data$ rfpaf<-as.factor(data$ rfpaf) # Atrial fibrillation
data$ rfpihd<-as.factor(data$ rfpihd) # Ischaemic Heart Disease
data$ rfphyp<-as.factor(data$rfphyp) # phypertension
data$ glas_cs<-as.numeric(data$ glas_cs) # Glasgow-coma scale
data$incont<-as.factor(data$incont) # Incontinence
data$rfsmok<-as.factor(data$rfsmok)   # smoking
data$SUBTYPE[ data$subtype==4]<- "LACI"
data$SUBTYPE[ data$subtype==2]<- "PACI"
data$SUBTYPE[ data$subtype==3]<- "POCI"
data$SUBTYPE[ data$subtype==1]<- "TACI"
data$laci<-as.factor(data$subtype == 4)
data$paci<- as.factor(data$subtype == 2)
data$poci<- as.factor(data$subtype == 3)
data$taci<- as.factor(data$subtype == 1)
data <- data[complete.cases(data), ]
left_stroke <- (data$nih12l== 1 & data$nih12l<5) | (data$nih5l==1
& data$nih5l<5) # defining left-stroke
data$left_stroke <- as.integer(as.logical(left_stroke)) # left-stroke

# Development cohort
-------------------------------------
data.dev <- data[ which(data$strk_y<=2010), ]

# Validation cohort
-------------------------------------
data.val <- data[ which(data$strk_y>2010 & data$strk_y<=2018), ]

# Mixed Model
-------------------------------------
model <- lmer(mtotfollowup~sqrt(time)*(sex+age+ethcat+dysphas+rfpdia
+left_stroke+batot+glas_cs+subtype+mtot)+
(id|time)+(bs(time)|id),data=data.dev)

# Model performance
-------------------------------------
model_performance(model)
summary(model)

# Prediction on validation data
-------------------------------------
predictedscore<-predict(model,data.val, allow.new.levels=TRUE)
```

```
predictedscore<-as.numeric(predictedscore)
predictedscore[predictedscore<0] <- 0
predictedscore[predictedscore>10] <- 10

# Plot observed VS predicted values (continous score)
-------------------------------------
ggplot(data=data.val, aes(x=time))+ xlim(0,5)+ ylim(0,10)
 + xlab("Time since stroke (years)")+ ylab("MMSE/AMT")
+ theme_classic()+stat_summary(fun.data="mean_cl_boot",
geom = "smooth",colour="black", size=0.5,
aes(y=predictedscore))+stat_summary(fun.data="mean_cl_boot",
 geom = "smooth", colour="black",
 linetype = "dashed",size=0.5,  aes(y=mtotfollowup))

data.val$predictedscore<-predictedscore
```

3.2 Analyse par sous groupes | Figure 3.3

```
# plot by GCS
-------------------------------------
data.val$GCS[data.val$glas_cs==15] <- "15"
data.val$GCS[data.val$glas_cs<15] <- "< 15"
ggplot(data.val,aes(x =time, y =predictedscore,
col = factor(GCS)))+xlim(0,5)+ylim(7,10)+
xlab("Time since stroke(years)")+ylab("MMSE/AMT")
+theme_classic()+(aes(linetype=GCS))
+stat_smooth(method = "loess", size = 0.5,colour="black",se = FALSE
+theme(legend.position="bottom")

# plot by age
-------------------------------------
data.val$AGE[data.val$age<=64] <- "0-64"
data.val$AGE[data.val$age >64 & data.val$age<75] <- "65-74"
data.val$AGE[data.val$age >75 & data.val$age<85] <- "75-84"
data.val$AGE[data.val$age>=85] <- "85+"
data.val <- data.val[complete.cases(data.val), ]
ggplot(data.val,aes(x =time, y =predictedscore, col = factor(AGE))
+xlim(0,5)+ylim(7,10)+xlab("Time since stroke (years)")
+ ylab("MMSE/AMT")+theme_classic()+(aes( linetype= AGE))
+stat_smooth(method = "loess", size = 0.5,colour="black",
se = FALSE )+theme(legend.position="bottom")


# Plot by Subtypye
-------------------------------------------------------------
data.val$SUBTYPE[ data.val$subtype==4]<- "LACI"
data.val$SUBTYPE[ data.val$subtype==2]<- "PACI"
data.val$SUBTYPE[ data.val$subtype==3]<- "POCI"
data.val$SUBTYPE[ data.val$subtype==1]<- "TACI"
data.val <- data.val[complete.cases(data.val), ]
```

```
ggplot(data.val,aes(x =time, y =predictedscore,
col = factor(SUBTYPE)))+xlim(0,5)+ylim(7,10)
+xlab("Time since stroke (years)")+ ylab("MMSE/AMT")+
  theme_classic()+(aes(linetype=SUBTYPE))
  +stat_smooth(method = "loess", size = 0.5,colour="black",
  se = FALSE )+theme(legend.position="bottom")


# Plot by left_stroke
-----------------------------------------------------
data.val$LEFT_STROKE[ data.val$left_stroke==1]<- "No left stroke"
data.val$LEFT_STROKE[ data.val$left_stroke==0]<- "left stroke"
ggplot(data.val,aes(x =time, y =predictedscore,
col = factor(LEFT_STROKE)))+xlim(0,5)+ylim(5,10)
+xlab("Time since stroke (years)")+ ylab("MMSE/AMT")+
theme_classic()+(aes(linetype=LEFT_STROKE))+
 stat_smooth(method = "loess", size = 0.5,colour="black",
 se=FALSE )+theme(legend.position="bottom")
```

3.3 Performance du modèle à différents seuils (Figure 3.4)

```
#Load packages
library(pROC)
library(ROCR)
library(dplyr)
library(epiR)
library(DescTools)
library(ResourceSelection)

# Filter data (3months)
 ------------------------------------
dat.sub3m<-filter(data.val,time==0.25)
#set Cutoff
dat.sub3m$catcognition4<-as.factor(ifelse(dat.sub3m$mtotfollowup>4,1,0))
dat.sub3m$catcognition8<-as.factor(ifelse(dat.sub3m$mtotfollowup>8,1,0))
# get predicted values
model1 <- glm(catcognition4 ~ predictedscore,data=dat.sub3m,family=binomial)
model2 <- glm(catcognition8 ~ predictedscore,data=dat.sub3m,family=binomial)
p1<-predict(model1,type="response")
as.numeric(p1)
p2<-predict(model2,type="response")
as.numeric(p2)
## ROC curve
roccurve1 <- roc(dat.sub3m$catcognition4 ~ p1)
roccurve2 <- roc(dat.sub3m$catcognition8 ~ p2)
plot(roccurve1)
plot(roccurve2)
```

```
## AUC
auc_5<-auc(roccurve1)
auc_8<-auc(roccurve2)
ci.auc(dat.sub3m$catcognition4, p1)
ci.auc(dat.sub3m$catcognition8, p2)
## Performance
tab1<-table(dat.sub3m$catcognition4, p1 > 0.5)
tab2<-table(dat.sub3m$catcognition8, p2 > 0.5)
epi.tests(tab1, conf.level = 0.95)
epi.tests(tab2, conf.level = 0.95)
auc_5
auc_8
## Brier score
BrierScore(model1)
BrierScore(model2)

# Filter data (1 year)
-------------------------------------
dat.sub1y<-filter(data.val,time==1)
#set Cutoff
dat.sub1y$catcognition4<-as.factor(ifelse(dat.sub1y$mtotfollowup>4,1,0))
dat.sub1y$catcognition8<-as.factor(ifelse(dat.sub1y$mtotfollowup>8,1,0))
# get predicted values
model1 <- glm(catcognition4 ~ predictedscore,data=dat.sub1y,family=binomial)
model2 <- glm(catcognition8 ~ predictedscore,data=dat.sub1y,family=binomial)
p1<-predict(model1,type="response")
as.numeric(p1)
p2<-predict(model2,type="response")
as.numeric(p2)
## ROC curve
roccurve1 <- roc(dat.sub1y$catcognition4 ~ p1)
roccurve2 <- roc(dat.sub1y$catcognition8 ~ p2)
plot(roccurve1)
plot(roccurve2)
## AUC
auc1<-auc(roccurve1)
auc2<-auc(roccurve2)
ci.auc(dat.sub1y$catcognition4, p1)
ci.auc(dat.sub1y$catcognition8, p2)
## Performance
tab1<-table(dat.sub1y$catcognition4, p1 > 0.5)
tab2<-table(dat.sub1y$catcognition8, p2 > 0.5)
epi.tests(tab1, conf.level = 0.95)
epi.tests(tab2, conf.level = 0.95)
auc1
auc2
## Brier score
BrierScore(model1)
BrierScore(model2)
```

```
# Filter data (5 year)
-------------------------------------------------------------
dat.sub5y<-filter(data.val,time==5)
#set Cutoff
dat.sub5y$catcognition4<-as.factor(ifelse(dat.sub1y$mtotfollowup>4,1,0))
dat.sub5y$catcognition8<-as.factor(ifelse(dat.sub1y$mtotfollowup>8,1,0))
# get predicted values
model1 <- glm(catcognition4 ~ predictedscore,data=dat.sub5y,family=binomial)
model2 <- glm(catcognition8 ~ predictedscore,data=dat.sub5y,family=binomial)
p1<-predict(model1,type="response")
as.numeric(p1)
p2<-predict(model2,type="response")
as.numeric(p2)
## ROC curve
roccurve1 <- roc(dat.sub5y$catcognition4 ~ p1)
roccurve2 <- roc(dat.sub5y$catcognition8 ~ p2)
plot(roccurve1)
plot(roccurve2)
## AUC
auc1<-auc(roccurve1)
auc2<-auc(roccurve2)
ci.auc(dat.sub5y$catcognition4, p1)
ci.auc(dat.sub5y$catcognition8, p2)
## Performance
tab1<-table(dat.sub5y$catcognition4, p1 > 0.5)
tab2<-table(dat.sub5y$catcognition8, p2 > 0.5)
epi.tests(tab1, conf.level = 0.95)
epi.tests(tab2, conf.level = 0.95)
auc1
auc2
## Brier score
BrierScore(model1)
BrierScore(model2)
```

3.4 Analyse de la courbe de décision (DCA) (Figure 3.5)

```
# Decision curve analysis  (DCA)
# load packages
library(MASS)
library(rmda)
#------------------
dat.sub3m<-filter(data.val,time==0.25)
dat.sub1y<-filter(data.val,time==1)
dat.sub5y<-filter(data.val,time==5)
dat.sub3m$catcognition8<-ifelse(dat.sub3m$mtotfollowup>8,1,0)
dat.sub1y$catcognition8<-ifelse(dat.sub1y$mtotfollowup>8,1,0)
dat.sub5y$catcognition8<-ifelse(dat.sub5y$mtotfollowup>8,1,0)
```

```
# At 3 months & cutoff = 8
------------------
full.model <- decision_curve(catcognition8~predictedscore,
data = dat.sub3m, thresholds = seq(0, 1, by = .05),bootstraps = 30)
plot_decision_curve(full.model,confidence.intervals = FALSE,
legend.position = "none",cost.benefit.axis = FALSE)
summary(full.model, measure = "NB")

# At 1 year & cutoff = 8
------------------
dat.sub1y$catcognition8<-ifelse(dat.sub1y$mtotfollowup>8,1,0)
full.model <- decision_curve(catcognition8~predictedscore,
data =dat.sub1y, fitted.risk = F, thresholds = seq(0, 1, by = .05),
bootstraps = 30)
plot_decision_curve(full.model,confidence.intervals = FALSE,
legend.position = "none",cost.benefit.axis = FALSE)
summary(full.model, measure = "NB")

# 5 year & cutoff = 8 ------------------
dat.sub5y$catcognition8<-ifelse(dat.sub5y$mtotfollowup>8,1,0)
full.model <- decision_curve(catcognition8~predictedscore,
data = dat.sub5y,fitted.risk = F, thresholds = seq(0, 1, by = .05),
bootstraps = 30)
plot_decision_curve(full.model, confidence.intervals = FALSE,
legend.position = "none",cost.benefit.axis = FALSE)
```

Appendice 4

Codes Chapitre 4

4.1 Simulations : Figure 4.1 | Figure 4.2

```
# load simulated data
-----------------------------------

simdata <- read.table(simdata, sep = "", skip = 6, header = TRUE,
fill = TRUE, na.string = c(-99.99, -99.9))
names(simdata) <- c(month.abb, "Annual")
simdata <- simdata[-nrow(simdata), ]
 rn <- as.numeric(rownames(simdata))
Years <- rn[1]:rn[length(rn)]
simdata <- data.frame(Observations = simdata[, ncol(simdata)],Year = Years)
simdata <- simdata[, -ncol(simdata)]
## stack the data
 simdata <- stack(simdata)[,2:1]
 names(simdata) <- c("Month","Observations")
simdata <- transform(simdata, Year = (Year <- rep(Years, times =
12)),nMonth = rep(1:12, each = length(Years)),
Date = as.Date(paste(Year, Month, "15", sep = "-"),
format = "%Y-%b-%d"))
 ## sort into temporal order
 simdata <- simdata[with(simdata, order(Date)), ]
 ## Add in a Time variable
simdata <- transform(simdata, Time = as.numeric(Date) / 1000)
-----------------------------------------------------------------
# State-space model (local level model)

# METHODE 1 (StructTS function)
------------
ts<-ts(data = simdata$Observations, start = 1659, end =2019,
frequency = 1)
plot(ts, xlab = expression(Années), ylab = expression(Observations),
type = "l", main = "Simulation: modèle espace-état ",
lty = "dashed")
fit.ts <- StructTS(ts,type = "level")
lines(tsSmooth(fit.ts), lty = "solid",col = "blue", lwd = 2)
legend("bottomright", legend = c("Observations", paste("modèle
```

```
espace-état")), col = c("black", "blue"), lty = c("dashed",
rep("solid",3)),cex=0.6)

# METHODE 2 (MARSS package)
mod2 = list(B = matrix(1), U = matrix(0), Q = matrix("q"),
            Z = matrix(1), A = matrix(0), R = matrix("r"), x0 = matrix("mu"),
            tinitx = 0)
A <- "zero"
U <- "zero"
kem.2 <- MARSS(ts, model = mod2)
c(coef(kem.2, type = "vector"), LL = kem.2$logLik, AICc = kem.2$AICc)
plot(kem.2)

## Mixed model (local level model)
----------------------------
plot(Observations ~ Year, data = simdata, xlab = expression(Années),
ylab = expression(Observations), type = "l",lty = "dashed" ,
main ="Simulation: modèle à effets mixtes (Gamm) ")
m1 <- gamm(Observations ~ s(Year, k = 361), data = simdata,
correlation = corARMA(form = ~ 1|Year, p = 1))
lines(fitted(m1$lme) ~ simdata$Year, lty = "solid",
col = "red", lwd= 2)
legend("bottomright", legend = c("Observations", paste("modèle à
effets mixtes")), col = c("black", "red"), lty = c("dashed",
rep("solid",3)),cex=0.6)
```

4.2 Application

4.2.1 Figure 4.3 | Figure 4.4

```
#Load data
data<-read.csv("~/data_long_mtot.csv")
TSdata <- read.csv("~/Desktop/cog-youssef/Execl-Manip/TSdata.csv", sep=";")

## Missing data
TSdata_n <- mice(TSdata,m=2,seed=500)
TSdata_n  <- complete(TSdata_n ,2)
write.csv(TSdata_n , file="TSdata_n.csv")
TSdata_n <- read.csv("~/TSdata_n.csv")
TSdata_mean<-aggregate(TSdata_n[, 4:5], list(TSdata_n$adm_y), mean)
colnames(TSdata_mean)[1] <- "Years"
TSdata_mean <-TSdata_mean[!TSdata_mean$Years == 2019, ]

##Mixed model
gamm <- gamm(batot_7~ s(Years,k=21,bs="fs",m=1), data=TSdata_mean,
method = "REML",correlation =corARMA(form = ~ 1|Years, p = 1))

plot(TSdata_mean$batot_7 ~ TSdata_mean$Years,lty = "dashed",
type ="l", xlab = expression(Années), ylab =
```

```
expression(Score_barthel),main="Moyenne annuelle du score de Barthel
à l'admission (SLSR)",cex.main=1,font.main=1)
lines(TSdata_mean$batot_7~ TSdata_mean$Years, lty = "dashed",
lwd =1)
lines(fitted(gamm$lme) ~ TSdata_mean$Years, lty = "solid",
col ="red", lwd = 2)
legend("bottomright", legend = c("Observations", paste("modèle à
effets mixtes")), col = c("black", "red"), lty = c("dashed",
rep("solid",3)),cex=0.6)
gamm$lme$logLik
AIC(gamm$lme)
gamm$lme$modelStruct

## State space  model

##METHODE 1 (StructTS)
ts<-ts(data = TSdata_mean$batot_7, start = 1995, end =2018
frequency = 1)
plot(ts,lty = "dashed", type = "l", xlab = expression(Années),
 ylab= expression(Score_barthel),main="Moyenne annuelle du score de
Barthel à l'admission (SLSR)",cex.main=1,font.main=1)
lines(TSdata_mean$batot_7~ TSdata_mean$Years, lty = "dashed",
lwd =1)
fit.ts <- StructTS(ts,type = "level")
lines(fitted(fit.ts), lty = "dashed",col = "blue", lwd = 2)
legend("bottomright", legend = c("Observations", paste("modèle
espace-état")), col = c("black", "blue"), lty = c("dashed",
rep("solid",3)),cex=0.6)


##METHODE 2 (MARSS)
mod2 = list(B = matrix(1), U = matrix(0), Q = matrix("q"),
Z = matrix(1), A = matrix(0), R = matrix("r"), x0
 = matrix("mu"), tinitx = 0)
A <- "zero"
U <- "zero"
kem.2 <- MARSS(ts, model = mod2)
c(coef(kem.2, type = "vector"), LL = kem.2$logLik,
 AICc = kem.2$AICc)
plot(kem.2)
```

4.2.2 Barthel index (age $>$ 65) : Figure 4.5|Figure 4.6|Figure 4.7

```
# GAMM for newdata (age > 65)
gamm <- gamm(batot_7~ s(Years,k=23,bs="fs",m=1), data=newdata,
method = "REML",correlation =corARMA(form = ~ 1|Years, p = 1))
plot(newdata$batot_7 ~ newdata$Years,type = "l",
xlab = expression(Années),
 ylab = expression(Score_barthel),main="Moyenne annuelle du score
 de Barthel à l'admission (age > 65(SLSR)",cex.main=0.8,font.main=1)
```

```
lines(newdata$batot_7~ newdata$Years, lty = "dashed", lwd = 1)
lines(fitted(gamm$lme) ~ TSdata_mean$Years, lty = "solid",
col = "red", lwd = 2)
legend("bottomright", legend = c("Observations", paste("modèle à
effets mixtes")), col = c("black", "red"),
lty = c("dashed", rep("solid",3)),cex=0.6)
gamm$lme$logLik
AIC(gamm$lme)
-----------------
# Time series: METHODE 1: StructTS (newdata)(age > 65)
ts<-ts(data = newdata$batot_7, start = 1995, end =2018,
frequency = 1)
plot(ts,lty = "dashed", type = "l", xlab = expression(Années),
ylab= expression(Score_barthel),main="Moyenne annuelle du score
deBarthel à l'admission (age > 65) (SLSR)",cex.main=0.8,font.main=1)
fit.ts <- StructTS(ts,type = "level")
#lines(fitted(fit.ts), lty = "solid",col = "red", lwd = 2)
lines(tsSmooth(fit.ts), lty = "solid",col = "blue", lwd = 2)
legend("bottomright", legend = c("Observations", paste("modèle
espace-état")), col = c("black", "blue"),
lty = c("dashed", rep("solid",3)),cex=0.6)
-----------------
# Time series: METHODE 2 MARSS (newdata)
ts<-ts(data = newdata$batot_7, start = 1995, end =2018,
frequency = 1)
mod2 = list(B = matrix(1), U = matrix(0), Q = matrix("q"),
Z = matrix(1), A = matrix(0), R = matrix("r"), x0 = matrix("mu"),
tinitx = 0)
A <- "zero"
U <- "zero"
kem.2 <- MARSS(ts, model = mod2)
c(coef(kem.2, type = "vector"), LL = kem.2$logLik,
AICc = kem.2$AICc)
plot(kem.2)
-----------------
derivgam<- Deriv(gamm$gam, n = 24)
plot(derivgam, sizer = TRUE, alpha = 0.01, xlab = Années )
```