I.S.F.A. **École Doctorale Sciences Économiques et de Gestion**

Mélanges de GLMs et nombre de composantes : application au risque de rachat en Assurance Vie

THÈSE

présentée et soutenue publiquement le 6/07/2012

pour l'obtention du

Doctorat de l'Université Claude Bernard Lyon I

(mathématiques appliquées)

par

Xavier MILHAUD

Composition du jury

Présidente :	Professeur Hansjoerg ALBRECHER
Rapporteurs :	Professeur Denys POMMERET Professeur Bernard GAREL
Examinateurs :	Professeur Stéphane LOISEL Professeure Véronique MAUME-DESCHAMPS Docteur Vincent LEPEZ

Laboratoire Science Actuarielle Financière — EA 2429

Remerciements

J'aimerais en tout premier lieu remercier mes tuteurs de thèse, Stéphane Loisel et Véronique Maume-Deschamps. Sans vous, cette enrichissante expérience n'aurait certainement pas eu le même goût, un goût de "reviens-y" (un peu comme le cassoulet). Votre ouverture et le fonctionnement que nous avons réussi à mettre en place ont été primordiaux pour le bon déroulement de cette thèse, qui aurait pu être bien plus compliquée au vu de la distance qui nous séparait et de nos emplois du temps. Je ne m'étends pas sur votre apport scientifique évident, mais merci du temps que vous m'avez consacré à discuter de sujets divers et variés. Votre sensibilité aux problématiques du monde de l'Assurance dans sa globalité a été un des facteurs clefs du succès de cette thèse, de même que votre désir de constamment m'aider à m'affirmer dans mes activités. Encore aujourd'hui dans ma nouvelle fonction (qui me convient à merveille !), vous avez été le point de départ par qui tout a commencé. Alors un grand merci pour tout, et surtout pour vos grandes qualités humaines ! Je remercie également Denys Pommeret et Bernard Garel d'avoir accepté de rapporter cette thèse, et je suis certain que leur expérience sur les différents sujets qui y sont abordés me permettra au travers de leurs remarques d'en améliorer encore le contenu. J'en profite pour souligner la gentillesse de Hansjoerg Albrecher dont le déplacement depuis l'étranger n'était pas tout à fait évident, mais qui me fait un immense plaisir en présidant ce jury. Par la même occasion, je tiens à remercier l'ensemble des personnes avec qui j'ai été amené à travailler durant ces trois dernières années : Marie-Pierre Gonon, Hai-Trung Pham, Vincent Lepez, Sylvain Coriat, ainsi que tout ceux que j'aurais oublié de citer. Leur oeil averti sur les problématiques de rachat ainsi que leur disponibilité pour aller toujours un peu plus loin dans la mise en commun de nos connaissances ont largement contribué au résultat final, que je suis fier de partager avec eux aujourd'hui.

Ensuite, j'aimerais souligner l'environnement exceptionnel que m'a offert le laboratoire SAF de l'ISFA. Si j'en suis là aujourd'hui, c'est parce qu'ils ont su me donner l'envie de m'investir dans ces sujets passionnants lorsque je doutais encore de mon orientation future. Cette équipe m'a transmis la flamme qui l'animait, mais a aussi su me faire partager l'ambiance unique qui y règne. Malheureusement, les contraintes parisiennes auront limité mes échanges avec les Lyonnais, mais j'espère vivement que ce ne soit que partie remise. J'attends d'ailleurs toujours d'être invité pour faire parti des fans de Didier ! Ou d'être recruté par l'OL, ce qui devrait être plus simple. Je remercie évidemment l'entreprise AXA de m'avoir soutenu moralement et financièrement dans ce projet, et de m'avoir permis de découvrir de nombreuses autres facettes du monde de l'Actuariat. Je pense aux multiples projets abordés dans l'équipe, et qui m'ont permis de développer une culture diversifiée de la modélisation des différents risques en Assurance Vie. Mon intégration au sein d'AXA Global Life s'est faite sans aucun accroc, et je garderai un souvenir ému de nombreuses personnes que j'y ai côtoyé. Merci aussi à toute l'équipe de l'ENSAE qui m'a accueilli les bras ouverts en Septembre dernier.

Enfin je dédie ce manuscript à ma famille et à Marion. Ma famille pour tout ce qu'elle a fait (et ce qu'elle continuera à faire !), et Marion pour son perpétuel soutien et sa compréhension de mon rythme pas toujours académique (surtout ces derniers temps !). Je crains que cette thèse ne m'ait bien trop fait "vieillir", mais j'espère avoir gardé un peu d'insouciance et par là-même satisfaire les mêmes propriétés que le bon vin. Je n'oublie pas mes amis qui m'ont permis de m'évader malgré nos très courts et peu fréquents séjours ensemble. Enfin il va être grand temps de se rattraper ; mais je sais que je peux compter sur eux de ce côté-là...

If you want to be happy...
... for an hour, take a nap
... for a day, go fishing
... for a month, get married
... for a year, get an inheritance
... for a lifetime, help someone.

Martin Seligman

Résumé

Segmentation et prévision du risque de rachat en Assurance Vie.

La question du rachat préoccupe les assureurs depuis longtemps notamment dans le contexte des contrats d'épargne en Assurance-Vie, pour lesquels des sommes colossales sont en jeu. L'émergence de la directive européenne Solvabilité II, qui préconise le développement de modèles internes (dont un module entier est dédié à la gestion des risques de comportement de rachat), vient renforcer la nécessité d'approfondir la connaissance et la compréhension de ce risque. C'est à ce titre que nous abordons dans cette thèse les problématiques de segmentation et de modélisation des rachats, avec pour objectif de mieux connaître et prendre en compte l'ensemble des facteurs-clefs qui jouent sur les décisions des assurés. L'hétérogénéité des comportements et leur corrélation ainsi que l'environnement auquel sont soumis les assurés sont autant de difficultés à traiter de manière spécifique afin d'effectuer des prévisions. Nous développons ainsi une méthodologie qui aboutit à des résultats très encourageants ; et qui a l'avantage d'être réplicable en l'adaptant aux spécificités de différentes lignes de produits. A travers cette modélisation, la sélection de modèle apparaît comme un point central. Nous le traitons en établissant les propriétés de convergence forte d'un nouvel estimateur, ainsi que la consistance d'un nouveau critère de sélection dans le cadre de mélanges de modèles linéaires généralisés.

Mots-clés: comportement de rachat, mélange, classification, GLM, sélection de modèle

Abstract

Segmenting and predicting surrender behaviours in Life insurance.

Insurers have been concerned about surrenders for a long time especially in Saving business, where huge sums are at stake. The emergence of the European directive Solvency II, which promotes the development of internal risk models (among which a complete unit is dedicated to surrender risk management), strengthens the necessity to deeply study and understand this risk. In this thesis we investigate the topics of segmenting and modeling surrenders in order to better know and take into account the main risk factors impacting policyholders' decisions. We find that several complex aspects must be specifically dealt with to predict surrenders, in particular the heterogeneity of behaviours and their correlations as well as the context faced by the insured. Combining them, we develop a methodology that seems to provide good results on given business lines, and that moreover can be adapted for other products with little effort. However the model selection step suffers from a lack of parsimoniousness: we suggest to use another criteria based on a new estimator, and prove its consistant properties in the framework of mixtures of generalized linear models.

Keywords: surrender behaviour, finite mixtures, classification, GLM, model selection

Table des matières

Conclusion et annexes

Introduction générale

Présentation de la thèse

Une partie de cette introduction est inspirée de l'article Milhaud et al. (2010), coécrit avec Stéphane Loisel et Marie-Pierre Gonon et paru en 2010 dans *Risques 83, p. 76-81.*

1 Contexte du sujet et motivations

Le contrat d'assurance vie est un accord entre une compagnie d'assurances qui prend l'engagement irrévocable de verser des prestations au bénéficiaire du contrat en fonction de la réalisation d'événements aléatoires viagers, en échange de quoi le souscripteur prend l'engagement **révocable** de verser des cotisations en fonction de la réalisation d'événements viagers. Le risque de rachat est omniprésent dans les problématiques de valorisation et de provisionnement des contrats d'épargne dans les sociétés d'assurance vie. Pour satisfaire par exemple à un besoin de liquidité immédiat, l'assuré peut à tout moment résilier son contrat et récupérer tout (rachat total) ou partie (rachat partiel) de son épargne capitalisée, éventuellement diminuée de pénalités prévues à cet effet et dépendantes des conditions fixées lors de la souscription. Une bonne compréhension du rachat, de ses facteurs explicatifs, et plus globalement du comportement des assurés permet d'adapter les clauses lors de la création de nouveaux produits (avec pour objectif la rétention de clients, le gain de parts de marché, ...) ; et de mettre en place de meilleures stratégies de gestion actif-passif.

Deux questions sous-jacentes au rachat doivent être abordées : tout d'abord les conséquences financières d'un mauvais choix de modélisation pour les lois de rachat et ensuite son impact sur les garanties. Nous nous focalisons dans cette thèse sur le premier point, étroitement lié au contexte économique et financier et donc à la dynamique des taux d'intérêt. Selon la position de l'assureur (phase d'investissement ou de désinvestissement) et son anticipation du marché, un scénario haussier aussi bien que baissier des taux d'intérêts peut avoir des conséquences importantes au niveau de sa gestion actif-passif et de son stock de réserves. Ces conséquences peuvent même devenir critiques en cas de rachat massif (dans un scénario haussier) ou d'absence de rachat (scénario baissier), obligeant ainsi l'assureur à emprunter à prix fort ou à verser un taux garanti supérieur au rendement de ses propres actifs. Nous distinguons ainsi qu'un problème d'adéquation se pose pour l'assureur dans tous les cas de figure. Sans parler spécifiquement du risque de taux, le timing même du rachat est primordial : certains contrats sont tarifés en supposant un profil de rachat donné (lié au recouvrement des frais d'émission financés par l'assureur), ou permettent un arbitrage (non pas au sens financier du terme) sous forme de changement de support de l'investissement par l'assuré. Cette dernière possibilité ne sera pas étudiée dans cette thèse.

Habituellement, l'assureur fait l'hypothèse que son portefeuille d'assurés est composé de

personnes se comportant indépendamment les unes des autres, ce qui est relativement juste en régime de croisière. Néanmoins, un problème majeur se pose dans le cas d'une perturbation de l'équilibre économique et financier : cette hypothèse est alors clairement inadaptée et les comportements des assurés peuvent devenir fortement corrélés. La recherche académique s'est penchée sur le sujet et les travaux de Lee et al. (2008) apparaissent comme un premier essai de modélisation dynamique du comportement humain, de même que ceux de Fum et al. (2007), Kim et al. (2008) et Pan et al. (2006) qui étudient les réactions humaines en cas de panique. D'un point de vue plus quantitatif,McNeil et al. (2005) présentent certains outils théoriques servant à modéliser l'interaction et la corrélation entre plusieurs variables aléatoires.

Des problématiques telles que l'anti-sélection (Bluhm (1982)) et l'aléa moral ont aussi une importance toute particulière, notamment dans un contexte de contrat d'assurance vie prévoyance où la santé des assurés est la question centrale dans le processus de tarification et de gestion des risques. Il est par exemple interdit en France de racheter des rentes viagères pour éviter le phénomène d'anti-sélection. En épargne, la santé des marchés financiers a une incidence directe sur le comportement de rachat des assurés, créant une corrélation entre leurs décisions. Vandaele and Vanmaele (2008), Bacinello (2005), Kuen (2005) et Tsai et al. (2002) entre autres ont développé des méthodes de valorisation financière de l'option de rachat. Ces méthodes ont vocation à être améliorées pour une meilleure prise en compte de la modélisation comportementale et des réelles questions et besoins de l'assuré lors du rachat, notamment en ce qui concerne la rationalité de son choix.

Nous dressons dans la suite de cette introduction un bref panorama du rachat en France et évoquons certains aspects-clefs dans la compréhension de ce risque. Ensuite, nous abordons la question du risque de rachat et de son évaluation dans le contexte réglementaire de la directive européenne Solvabilité II.

L'assurance vie reste aujourd'hui le placement préféré des Français en offrant la meilleure combinaison risque-rendement-fiscalité : fin 2010, plus de 27 millions de contrats avaient déjà été ouverts et les prestations annuelles versées par les organismes d'assurance vie dépassaient 106 [1] milliards d'euros ! Les offres sont multiples et permettent de répondre aux besoins des épargnants, les principaux critères étant la liquidité des sommes investies, le rendement et la sécurité financière associée aux entreprises d'assurances. La liquidité se traduit principalement, pour les contrats d'épargne monosupports (un seul support en euros, à valeur plancher garantie) et multisupports, par la liberté de sortir du contrat sans perdre les avantages du produit. Pour cela, l'assuré devra connaître quelques bases de fiscalité applicables aux contrats d'assurance vie. Cette fiscalité évolue régulièrement mais continue d'offrir certains avantages, notamment pour les contrats d'une durée de détention supérieure à huit ans. Cette caractéristique est clairement identifiable dans les observations passées, avec un pic de rachat lors de la neuvième année de présence en portefeuille. L'assuré devra aussi porter attention aux pénalités qui pourraient lui être prélevées au moment de la sortie, qui, selon les codes des assurances, sont cependant limitées à 5% des intérêts de la somme épargnée sur dix ans. Enfin, il pourra étudier les options qui lui sont offertes dans le contrat pour bénéficier de cette liquidité sans pour autant clore son contrat d'assurance vie et ainsi ne pas perdre son antériorité fiscale, au travers notamment des possibilités de rachat partiel ou d'avance. On entend par rachat partiel la possibilité qui est offerte à un assuré de ne racheter qu'une partie de son épargne, et ainsi ne pas demander la fermeture de son contrat. L'avance quant à elle est assimilable à un prêt que consent l'assureur envers l'assuré, l'épargne de l'assuré constituant alors le colla-

1. Source : Fédération Française des Sociétés d'Assurances (FFSA)

téral. D'autres clients, souvent plus fortunés, sont également intéressés par le caractère non rachetable de leur contrat, mais cela est un autre sujet. Le phénomène de rachat est important pour les assureurs français. Meilleure en est leur connaissance, meilleure sera leur anticipation de gestion des flux entrants et sortants en termes de trésorerie. Ils pourront même parfois en profiter pour améliorer leur offre produit. Alors, quelles sont les variables explicatives du choix de l'assuré, comment pouvons-nous projeter les comportements humains, pouvons-nous répondre en moyenne ? Tout d'abord en termes de variables explicatives, on a observé pendant de nombreuses années l'influence directe de la fiscalité et des profils patrimoniaux des assurés. Ces observations avaient pour avantage qu'elles étaient connues parfaitement de l'assureur et pouvaient ainsi faire l'objet de modélisations relativement adaptées.

Puis, on a introduit des effets moins évidents tels que la tenue des marchés financiers. Lorsque les taux baissent, les assurés seraient-ils plus fidèles ? Lorsque les taux montent, auraient-ils tendance à sortir ? Rien n'est moins évident, d'autant que les données selon lesquelles les variations de taux pourraient influencer le comportement des assurés ne sont pas nombreuses. Même au plus haut de la crise financière que l'on vient de connaître, la fidélité des clients n'a pas été énormément entamée. Ce n'est pas tant que les assurés ne suivent pas l'actualité ou ne prennent pas le temps de modifier leur allocation d'épargne, mais plutôt qu'il leur est difficile d'évaluer à quel taux de marché il serait intéressant pour eux d'aller souscrire ailleurs en tenant compte de nouveaux frais d'entrée, d'une perte de l'antériorité fiscale, etc. Ces nombreux facteurs compliquent les décisions à prendre de manière rationnelle. En revanche, c'est lorsque l'on commence à s'intéresser au rendement relatif des contrats d'assurance vie entre eux que l'épargnant devient plus vigilant, voire plus susceptible. Et c'est ainsi que les assureurs se sont penchés sur cette question. Combien de clients vont racheter leur contrat lorsque je servirai un taux de rendement inférieur à celui de mon concurrent, et à partir de quel écart de taux commenceront-ils à réagir ? Là encore, les statistiques ne sont pas nombreuses. Quelques assureurs ont démarché leurs clients pour connaître leur sensibilité à un écart de taux, mais les résultats obtenus ne permettent pas de modéliser de façon fiable une courbe de comportement. D'autres facteurs, comme le relais d'information par la presse quant aux taux garantis, peuvent influencer les décisions des souscripteurs. A ce titre, de nombreux assureurs ont élu un produit vitrine sur lequel ils communiquent pour le marché. De là à penser que tous les assurés aient été aussi largement récompensés... Voyons maintenant comment Solvabilité II traite le risque de rachat.

Avant de parler du risque de rachat en tant que tel, évalué pour les besoins en solvabilité, rappelons que Solvabilité II réforme également les calculs de provisions. Dans un bilan au format économique, les provisions techniques seront estimées avec des hypothèses de type *best estimate*, quand elles sont aujourd'hui estimées avec des hypothèses prudentes et conservatrices définies par exemple par le Code des Assurances. Quelles sont les implications pour ce qui concerne les rachats ? Aujourd'hui, le taux de rachat estimé dans les provisions techniques d'épargne est de 100% à chaque instant. L'entreprise d'assurances se doit donc d'immobiliser en permanence, dans ses comptes, le montant de l'épargne de chaque client, comme si le rachat devait intervenir demain. Dans un bilan plus économique, les provisions incluront des probabilités de rachat estimées en best estimate, c'est-à-dire reflétant au mieux la probabilité de rachat observée par l'assureur. Ainsi, les hypothèses de rachat qui ne servaient hier que dans le cadre des calculs d'embedded value ou d'études ALM (gestion actif-passif) vont-elles être introduites dans la comptabilité des entreprises d'assurances.

Ces lois de probabilités vont également servir à simuler le comportement des assurés dans le cadre des stress tests qui interviennent dans le calcul du SCR (Solvency Capital Requirement).

Les assureurs sont encouragés à utiliser la meilleure connaissance possible qu'ils ont de leur portefeuille, pour modéliser les flux financiers de passif dans les scénarios de stress comme ceux des marchés financiers permettant d'estimer les SCR de taux ou d'actions. Les dernières pré-spécifications techniques de la cinquième étude quantitative d'impact (QIS 5) donnent des formules fermées (cf. TP 4.58) pour modéliser les rachats en fonction des garanties offertes, des taux servis, des conditions de marchés financiers... Ces différentes formules doivent encore être calibrées par les assureurs pour refléter au mieux leur portefeuille.

Enfin, dans la liste des risques nécessitant la mise en oeuvre d'un calcul de SCR se trouve à part entière le risque de rachat à l'intérieur du module de risque de souscription. Dans les dernières parutions des pré-spécifications techniques, les assureurs sont priés d'étudier l'impact d'une hausse constante du taux de rachat de 50 % (limité à un taux de rachat de 100 %), l'impact d'une baisse constante du taux de rachat de 50 % (limité à un taux de rachat diminué de 20 %) et l'impact d'un rachat massif de 30 % (choc absolu) de la population sous risque. L'impact le plus significatif sera retenu pour être intégré au risque de souscription vie selon la matrice de corrélation définie dans les textes (paramètre de pseudo-corrélation égal à 50 % avec le SCR du risque de dérive des frais). Pour plus de détails sur toutes ces questions, l'organisme européen de contrôle (EIOPA [2]) est également une source intéressante de données.

2 Revue bibliographique

Dans le monde académique, la modélisation des comportements de rachat a suscité un vif intérêt il y a une vingtaine d'années, avant de connaître un ralentissement. Historiquement, deux approches ont été privilégiées : l'hypothèse de la nécessité urgente de ressource pour l'assuré (Outreville (1990)) et l'hypothèse du taux d'intérêt (Pesando (1974) et Cummins (1975)). La première s'interprète facilement : admettons qu'un événement imprévu et coûteux se produise dans la vie d'un assuré (achat d'une voiture suite à un accident, achat d'un bien immobilier), le besoin d'argent pourrait le pousser à résilier son contrat d'assurance-vie afin de disposer des fonds nécessaires. L'hypothèse du taux d'intérêt est complètement différente : le principe est que si les taux d'intérêts du marché augmentent alors les taux de résiliation augmentent aussi, car des opportunités d'arbitrage apparaissent naturellement sur le marché. Ainsi, des contrats à niveau de prime et de garantie égales offrent de meilleurs rendements.

Renshaw and Haberman (1986) sont les premiers à s'intéresser à la modélisation du comportement des assurés de manière statistique : ils analysent les comportements de rachat d'Assurance Vie en Ecosse en 1976 et dégage quatre principaux facteurs de risque de rachat que sont la compagnie, le type de contrat, l'âge et l'ancienneté du contrat. Ils utilisent des modèles linéaires généralisés (GLM) avec des termes d'intéraction entre ces facteurs de risque afin de bien modéliser l'hétérogénéité du portefeuille et les effets de l'ancienneté du contrat. Kim (2005) tente d'utiliser la régression logistique afin d'expliquer les rachats individuels d'un portefeuille coréen, en considèrant diverses variables explicatives catégorielles ou continues telles que l'âge, le sexe ou même le taux de chômage. Cette approche constituera d'ailleurs la première modélisation étudiée dans ce mémoire au chapitre 1. Dans le même esprit, Cox and Lin (2006) utilisent un modèle Tobit et insistent sur l'importance de l'ancienneté du contrat comme facteur explicatif du rachat dans le cadre des taux de rachat de rentes. Un an auparavant, Kagraoka (2005) applique une loi de Poisson au cas de rachats de contrats d'assurance dommages au Japon. Pour capter la surdispersion des données, il réalise ensuite la même

2. URL : https ://eiopa.europa.eu/

étude en utilisant une loi binomiale négative avec comme variables d'entrée le sexe, l'âge, la saisonnalité, l'ancienneté du contrat et le statut de travail. Dans ce contexte d'Assurance non-Vie, le mémoire de Dutang (2011) modélise les rachats par une approche élasticité prix, qui s'explique par la simple raison du terme du contrat (fixé à un an avec tacite reconduction). Plus généralement, des résultats pour la modélisation d'évènements rares peuvent être trouvés dans l'excellent papier de Atkins and Gallop (2007) ; qui présente plusieurs applications de modèles de régression associés à des données de comptage dont la loi de Poisson, la loi Binomiale Négative, et leurs extensions où le surplus de masse en 0 est modélisé ("zero-inflated"). Atkins and Gallop (2007) montrent que ces modèles sont adaptés lorsque les réponses présentent des distributions fortement asymétriques, et permettent d'éviter le biais introduit par les méthodes de régression par moindres carrés utilisées avec ce type de données (hypothèse de normalité des observations déraisonnable). En 2007, Fauvel and Le Pévédic (2007) rédigent un mémoire sur les rachats dans lequel l'ensemble des notions clefs sont abordées, avec une approche économiste via la théorie de l'espérance d'utilité couplée à des méthodes de finance quantitative. Le principal défaut selon nous de cette approche est qu'elle est basée sur la rationalité des assurés, une hypothèse relativement discutable.

Dans la littérature, d'autres auteurs comme Engle and Granger (1987) utilisent un critère de minimisation des erreurs par la théorie des moindres carrés ordinaires lors du calibrage d'un modèle cointégré. Ils proposent de modéliser les rachats grâce à un phénomène de cointégration entre les taux de rachat et certaines variables économiques, dans le but de séparer la dynamique court-terme (de ces taux) des relations long-terme potentielles avec le taux de chômage et les taux d'intérêts. Tsai et al. (2002) démontrent une relation d'équilibre long-terme entre le taux de rachat et le taux d'intérêt, leur but final étant d'estimer le provisionnement nécessaire en tenant compte d'un taux de mortalité stochastique, de taux d'intérêts et de rachats précoces. Ils étudient en plus les questions de gestion de risque et de solvabilité de l'assureur. Dans un autre registre, Albert et al. (1999) étudient la relation entre les taux de rachat et les taux de mortalité entre 1991 et 1992 aux Etats-Unis avec des données provenant de multiples compagnies d'assurance comme Allstate, Equitable, New York Life, Sun Life et bien d'autres. Les données sont différenciées par statut fumeur.

Enfin une longue série d'auteurs s'intéresse ensuite à l'évaluation financière de l'option de rachat contenue dans les contrats d'Assurance Vie. C'est clairement le domaine dans lequel la littérature est la plus abondante, avec notamment l'école italienne. Pour n'en citer que quelques uns, Bacinello (2005) propose un modèle basé sur l'approche de Cox-Ross-Rubinstein (CRR) et ses arbres binomiaux pour calculer la valeur de rachat de contrats à prime unique ou annuelle, avec une garantie plancher à maturité ou en cas de décès. Costabile et al. (2008) calculent le montant de primes périodiques associées à des contrats indexés sur les marchés financiers avec option de rachat et intérêts garantis, par le modèle CRR et un artifice de calcul (schéma avant-arrière couplé à une interpolation linéaire). Bacinello et al. (2008) se focalisent sur les rachats précoces et les considèrent comme des options américaines qu'ils valorisent grâce à un algorithme des moindres carrés Monte Carlo, à cause du fait que ces diverses options changent l'allure classique du payoff d'une option. Leur modèle prend en compte une mortalité stochastique et des sauts pour les indices financiers, le point fort de cet article est mis sur la performance de l'algorithme proposé. De Giovanni (2007) investigue plus particulièrement les différences dans la modélisation des comportements de rachat entre son approche, "l'espérance rationnelle", et l'approche communément utilisée sur la place : la théorie American Contingent Claim (ACC) basée sur le comportement optimal des assurés (d'un point de vue de l'exercice de leur option). Il s'appuie sur le fait bien connu que les agents ne sont ni rationnels ni opti-

maux, et donne des résultats sur la différence d'impact des taux d'intérêts et de l'élasticité prix en utilisant l'approximation quadratique d'une fonction modélisant le comportement. Kuen (2005) utilise l'approche ACC et le mouvement brownien pour décomposer les contrats avec participation au bénéfice en trois différentes options sous-jacentes : une obligation, une option de rachat et une option bonus. Il valorise ces contrats et arrive à la conclusion que la valeur de ce type de contrat est fortement sensible à l'option de bonus. Shen and Xu (2005) quantifient l'impact de l'option de rachat anticipé sur la valorisation de contrats en unités de compte (UC) à taux garanti, en utilisant le mouvement brownien géométrique et des équations différentielles partielles. Vandaele and Vanmaele (2008) explicitent la stratégie de couverture d'un portefeuille de contrats en UC comprenant une option de rachat en incorporant des hypothèses intéressantes : les temps de paiement et le temps de rachat ne sont clairement pas indépendants du marché financier (un processus de Lévy modélise ce marché). Toutefois cet article n'est pas facilement abordable à cause de sa technicité. Récemment, Nordahl (2008) a écrit un article intéressant sur la valorisation de l'option de rachat pour des contrats retraite. Il se base sur l'approche Longstaff-Schwartz et sur des simulations de Monte Carlo, en considérant le fait que l'option de rachat est comparable à une option américaine avec un strike stochastique. Torsten (2009) se place lui dans le cadre de contrats avec participation aux bénéfices de la compagnie, et souligne le fait qu'en général le problème de couverture et celui de valorisation ne peuvent être séparés. En effet le portefeuille d'investissement de l'assureur sert souvent de sous-jacent à la couverture, ce qui amène des difficultés supplémentaires.

Enfin, la plupart des articles cités concerne la valorisation de ladite option de rachat, mais rares sont ceux dont l'objectif est la modélisation du comportement de rachat (notre intérêt). Par conséquent nous ne développerons pas de méthodes financières dans notre étude, mais cette revue est l'occasion de se rendre compte du clivage entre les deux principales écoles qui étudient les rachats : la première, composée essentiellement de chercheurs académiques, s'attache à l'étude des aspects conjoncturels tandis que la seconde, menée par des professionnels, se focalise davantage sur les aspects structurels. Pourtant ces deux aspects vont de pair et semblent aussi importants l'un que l'autre. Pour une plus grande exhaustivité sur le sujet, une très vaste littérature à considérer dans le cadre des rachats conjoncturels est d'ailleurs accessible via la modélisation des rachats de crédits en finance (Stanton (1995) ou encore Hin and Huiyong (2006)). A ce propos, Hin and Huiyong (2006) abordent un nouvel aspect de modélisation pour les comportements de rachat de crédits : l'utilisation de modèle de survie. Ils étudient mensuellement les rachats à Shangai entre 1999 et 2003 par l'usage du modèle à hasards proportionnels (de type modèle de Cox, Cox (1972)) dans le but de comprendre le fonctionnement du marché résidentiel chinois, avec des variables telles que le revenu des emprunteurs, le PIB. Ce papier détermine des facteurs de risque tout en développant l'aspect catégorisation des données d'entrée, mais ne traite malheureusement pas de la question des prévisions.

Pour clôturer ces quelques lignes, il est important de préciser que cette thèse traite de la modélisation des **comportements** de rachat, bien que beaucoup de professionnels abordent le problème des rachats sous l'angle des montants rachetés plutôt que des décisions d'assurés. Ce regard est complètement légitime dans la mesure où le rachat d'un assuré très riche a évidemment plus d'impact que celui d'un assuré qui ne retirerait que peu d'argent. En termes de modélisation, les outils diffèrent suivant le phénomène étudié (lois de probabilité continues dans le cas des montants) mais les principaux facteurs d'influence sont identiques, d'où une transposition possible de notre travail à cette vision alternative.

3 Intuitions sur les facteurs de risque et pratiques de gestion

A première vue, il existe bien des facteurs pouvant affecter les comportements de rachat. Globalement ces facteurs de risque peuvent se résumer en deux grandes catégories, les effets structurels et les effets conjoncturels. Il est anecdotique de souligner que la liste des éléments donnée ci-dessous influençant les rachats est incomplète, puisque tout un chacun peut avoir ses propres motivations n'entrant pas dans ce cadre bien posé. De plus, les informations privées quant à l'anticipation des taux de rachat par les compagnies d'assurance ne sont généralement pas disponibles car il s'agit d'une information stratégique. Cependant, nous distinguons :

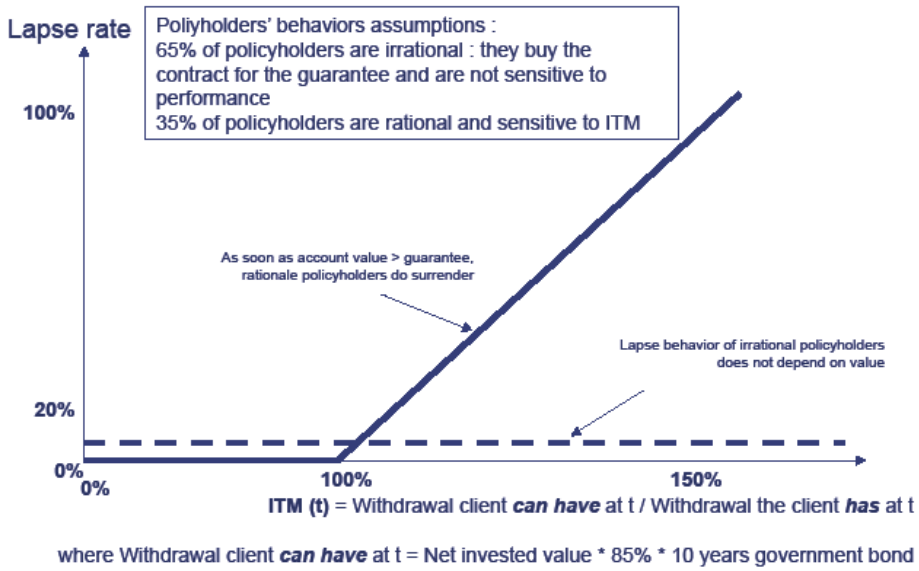
- parmi les facteurs *structurels* :
  - la **ligne d'affaire** et le **type de contrat** correspondant : épargne avec pure épargne, contrat mixte, unités de compte ; et prévoyance avec santé, maladies redoutées, incapacité-invalidité, décès ;
  - les **caractéristiques des contrats** : la participation aux bénéfices ; l'ancienneté du contrat ; le montant, le nivellement et la fréquence de la prime, le commissionnement, le réseau de distribution ;
  - **les caractéristiques des assurés** : le sexe, la catégorie socio-professionnelle, le lieu de résidence, le rapport prime sur salaire ou plus globalement la richesse de l'assuré, le statut fumeur, le statut marital, l'âge ;
- parmi les facteurs *conjoncturels* :
  - le changement de **législation** (voir annexe A.2) ;
  - le spread de taux de rendement avec la **concurrence** ;
  - **l'image** et le **rating** de la compagnie (voir annexe A.1) ;
  - l'évolution de l'offre (lancement de produits) et les **stratégies de vente** ;
  - les changements démographiques ;
  - l'évolution des taux d'intérêts, le taux de chômage, l'inflation, la croissance, le PIB. En fait cela représente tout ce qui est lié de près ou de loin au **contexte économique**.

Plusieurs organismes ont réalisé des études empiriques sur les déclencheurs de rachat ; dont les principaux sont la Society Of Actuaries (SOA) aux Etats-Unis, la Fédération Française des Sociétés d'Assurance (FFSA) en France, et la Fellow Institute of Actuaries (FIA) au UK. Toutefois il faut veiller à garder un certain recul par rapport à ces résultats pour diverses raisons, notamment le fait que les compagnies d'assurance y participant ne sont pas forcément représentatives de l'industrie (ou du marché dans lequel nous sommes positionnés). De plus, les données sont souvent imprécises et relativement peu fiables, car agrégées pour des raisons de confidentialité.

Ce qu'il faut retenir, c'est que la plupart des compagnies d'assurance "modélisent" aujourd'hui le taux de rachat à l'aide de tableaux à double (voire multiples) entrée(s) pour le risque de base (les entrées étant les facteurs de risque), auxquels elles appliquent une fonction de rachat dynamique permettant d'ajuster ce risque de base en fonction des conditions de marché (concurrentiel, financier). En fonction des acteurs, la fonction dynamique prend usuellement une des quatre formes présentées ci-dessous, dont les deux premières sont basées sur des méthodes financières de valorisation de l'option de rachat.

Fonction linéaire Cette fonction dynamique se base sur l'évaluation financière de l'option de rachat et ajuste le taux de rachat de base en fonction de la valeur de cette option. Pour cela, le but est de déterminer si l'option est dans la monnaie ou non : dès que l'option est dans la monnaie, on augmente linéairement (figure 1) le taux de rachat pour la part des assurés qui

FIGURE 1 – Exemple de fonction dynamique des rachats de forme linéaire.



est estimée sensible au marché.

Fonction en escalier Lorsque la valeur de l'option de rachat est élevée, l'assuré est supposé avoir tout intérêt à la conserver et le niveau de rachat est bas. Le raisonnement inverse est valable et suppose la définition d'un facteur multiplicatif (ici 10 dans la figure 2) qui provoque une forme en escalier et donc un changement brutal du taux.

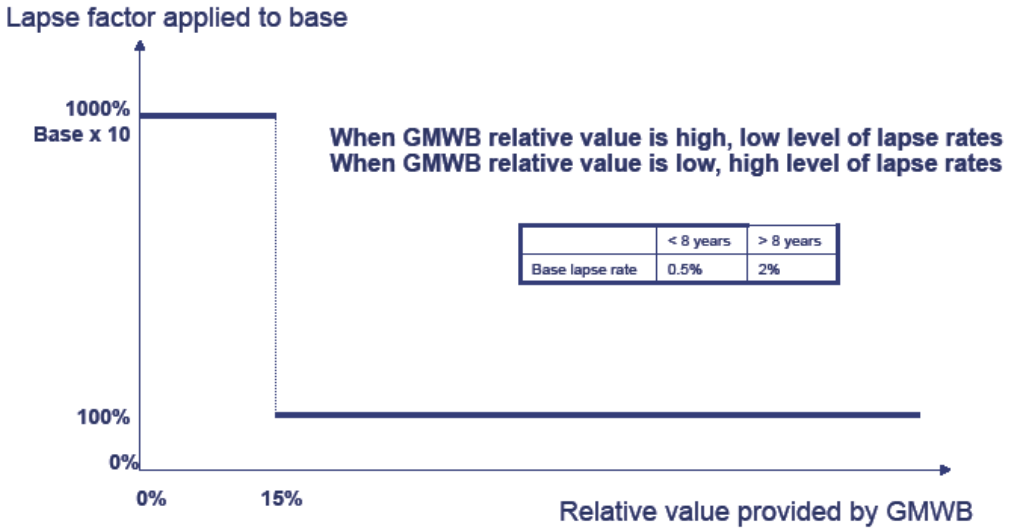
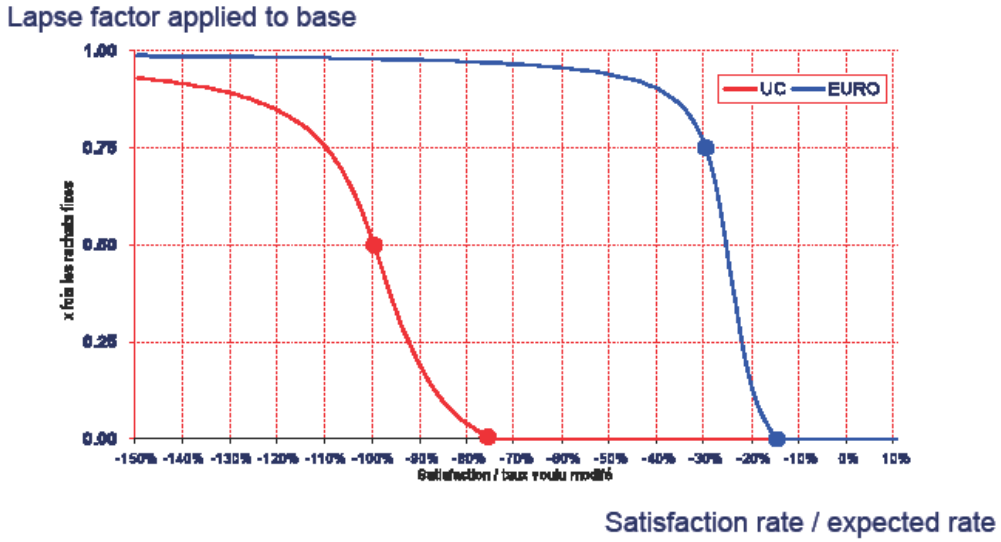


FIGURE 2 – Exemple de fonction dynamique des rachats en escalier.

FIGURE 3 – Exemple de fonction dynamique des rachats de forme Arctangente.



Fonction exponentielle La fonction exponentielle qui s'applique au risque de base est construite suivant la valeur du *spread* (rendement espéré - taux crédité) :

- si ce *spread* est négatif, alors nous considérons la fonction *exp(slope* × *spread)*, où *slope* est la vitesse de convergence vers le rendement espéré de l'assuré (par défaut 50%)
- si ce *spread* est positif, alors nous considérons la fonction *2 - exp(-slope* × *spread)*.

Si l'on devait se représenter ces fonctions, nous verrions que c'est toujours la même idée sous-jacente : si le contrat est favorable à l'assuré, le taux de rachat est abaissé par rapport à son niveau de base et inversement.

Fonction arctangente L'idée ici rejoint la modélisation dynamique linéaire ou en escalier, mais introduit une subtilité quant à la sensibilité des assurés par rapport à la différence entre leurs attentes et ce qu'ils reçoivent. Cette sensibilité non linéaire (figure 3) est caractérisée par un seuil à partir duquel le facteur multiplicatif "explose". C'est la modélisation dynamique la plus courante, et c'est celle dont nous discuterons les hypothèses au chapitre 2 (courbe en S).

L'importance du risque de rachat diffère suivant les lignes d'affaire. D'autre part, il faut distinguer deux types de contrats en épargne : les contrats avec garantie de taux (de type fonds euros ou garantie plancher), et les contrats sans garantie de rendement avec les UC classiques par exemple. Les observations tendent à montrer qu'une crise provoque globalement une baisse du nombre de rachats pour les contrats de la première catégorie (taux garantis), et une hausse du nombre de rachats pour les produits sans garantie de taux (la perception du risque de l'assuré change et il va réallouer les nouveaux flux d'épargne vers des supports garantis). La question du rachat est donc centrale pour tous les produits d'épargne individuelle ; mais ne l'est pas vraiment dans le domaine de la prévoyance. En effet, les contrats sont majoritairement (environ 70%) collectifs dans ce cadre : l'assuré ne connaît pas vraiment ses droits et ne se pose donc généralement pas la question du rachat.

Trois principales dimensions expliquent les rachats selon les équipes de vente : la fiscalité, les pénalités de rachat et le réseau de distribution. L'effet de la fiscalité n'est évidemment

visible que dans les pays pour lesquels certaines contraintes existent (ex : la France), les pénalités de rachat agissent sur l'assuré de la même manière qu'une contrainte fiscale (l'ancienneté du contrat guide souvent le profil de ces pénalités), et le réseau de distribution est en fait fortement lié au commissionnement des agents de vente. Un agent agréé sera tenté de provoquer le rachat aussitôt que les pénalités de rachat ou la fiscalité seront favorables à l'assuré, ou dès que le commissionnement qu'il reçoit pour la souscription de nouvelles affaires sur de nouveaux produits lui est favorable. Cette dernière remarque constitue d'ailleurs un véritable écueil en termes de modélisation, car il est impossible de prévoir à long terme la sortie de nouveaux produits et le comportement des agents de vente (bien que l'impact en soit majeur). Il est également important de garder en tête que le profil de rendement du produit va pousser l'assureur à favoriser ou non les rachats à un certain moment du contrat.

Enfin, rappelons que le rachat est un problème aussi bien à la hausse qu'à la baisse. Au Japon par exemple, AXA a dû transformer ses produits suite à l'explosion des taux de rachat dûe à la crise des taux d'intérêts (anormalement bas) à la fin des années 1990. Le niveau de la baisse ou de la hausse joue de manière relative car l'assuré réagit relativement à ce qu'il possède. Pour éviter les rachats massifs, certains produits (pour le moment marginaux) ont des clauses très spéciales comme un "facteur d'ajustement de marché" en cas de forte hausse des taux. Dans un futur proche, ce type de produit pourrait se développer dans la mesure où l'une des grandes craintes actuelles des assureurs concerne la remontée soudaine des taux au sortir de la crise (il n'y a jamais eu de taux si bas que les taux actuels en Europe, d'où des nouvelles affaires souscrites depuis trois ans à de faibles rendements). Cela pourrait entraîner une vague massive de rachats ayant des effets dévastateurs.

4 Contributions personnelles

Nous pouvons dégager de la revue bibliographique quatre grands types de modélisation : une modélisation financière sous forme de valorisation d'option (modélisation individuelle de la décision), une modélisation statistique sous forme de série temporelle (modélisation collective des décisions de rachat), une approche économique basée sur la théorie de l'espérance d'utilité et une modélisation probabiliste (individuelle) sous forme de modèle GLM. Nous avons privilégié cette dernière approche tout au long de la thèse, car elle permettait notamment de prendre en compte les caractéristiques individuelles et de modéliser la **décision** de rachat sans hypothèse préalable sur la rationalité des comportements.

Afin de mener à bien ce travail, nous avons lors d'une première étape tenté de segmenter le portefeuille Vie en classes de risque. Ceci a donné lieu à l'écriture et la publication de Milhaud et al. (2011), dans lequel nous comparons l'utilisation de deux modèles de classification dont les fondements théoriques sont radicalement différents. En formalisant de manière minimaliste notre démarche, le but était de partitionner un ensemble d'observations $\mathcal{X}$ en groupes homogènes en termes de caractéristiques (données par le vecteur X) et de comportements de rachat (données par la variable Y). La première approche, correspondante aux méthodes `CART`, consiste à utiliser un critère de mesure d'homogénéité qui puisse rendre compte de la qualité d'une division Δ de l'ensemble initial d'observations, et de maximiser ce critère :

$$\Delta^\star = \arg\max_{\Delta \in D}(\delta\, impur(\Delta)),$$

où $\delta\, impur(\Delta)$ désigne le gain d'homogénéité grâce à la division Δ. Ce processus réitéré étape par étape conduit à un partitionnement de $\mathcal{X}$ qui semble avoir de bonnes propriétés

classifiantes. L'approche paramétrique du modèle logistique nécessite quant à elle d'introduire des hypothèses sur la distribution des données Y. Dans notre cas, chaque observation Y_j de décision individuelle de rachat est de Bernoulli et est indépendante de sa "voisine". Les risques relatifs entre deux individus j et m sont supposés être proportionnels :

$$\frac{\dfrac{p_j}{1-p_j}}{\dfrac{p_m}{1-p_m}} = e^{\sum_{k=1}^{p}(X_{jk}-X_{mk})\beta_k},$$

avec p_j (respectivement p_m) la probabilité de rachat de l'individu j (resp. m) de caractéristiques X_j (resp. X_m). Cette hypothèse est forte dans la mesure où ces rapports de risque sont supposés ne pas évoluer au cours du temps, en plus de considérer les effets comme multiplicatifs. Notre principale contribution dans cette étude n'a pas été d'ordre théorique : nous montrons surtout que ces deux modèles de choix sont complémentaires mais peuvent amener à des conclusions erronées en fonction du contexte de leur application. Ce premier chapitre se veut volontairement pragmatique, et sert de point d'entrée dans le sujet : l'idée est de détecter des profils de risque sans présupposer d'hypothèses sous-jacentes aux données. En ce sens, notre "unique" contribution a été de fournir aux équipes produit quelques premières recommandations sur les profils risqués d'assurés. Toutefois, cette étude s'est aussi révélée déterminante dans l'approche développée au chapitre 3.

Dans le chapitre 2, nous discutons de l'hypothèse d'indépendance des comportements entre individus, en illustrant ses conséquences sur un exemple concret que nous avons publié dans Loisel and Milhaud (2011). L'alternative, qui introduit de la dépendance entre des variables aléatoires de Bernoulli, est réalisée par un modèle à chocs communs :

$$I_k = J_k I_0 + (1 - J_k) I_k^{\perp},$$

où J_k (de paramètre p_0) correspond à l'indicatrice de l'événement "le $k^{\text{ème}}$ assuré a un comportement moutonnier", I_0 est un consensus collectif de décision de rachat, et $I_k^{\perp}$ (de paramètre p) est une décision de rachat propre à l'assuré k. Bien sûr, ce modèle doit être calibré en fonction de données empiriques, notamment un écart de taux Δr entre la garantie et la concurrence qui provoque l'insatisfaction de l'assuré. Nos contributions théoriques sont d'abord le calcul de la distribution du nombre M de rachats en portefeuille par une approche combinatoire :

$$P(M = k) = p \sum_{i=0}^{k} a_{i,k} + (1-p) \sum_{j=0}^{n-k} b_{j,k},$$

avec pour $0 \leq i \leq k$,

$$a_{i,k} = C_n^i p_0^i (1-p_0)^{n-i} C_{n-i}^{k-i} p^{k-i} (1-p)^{n-k},$$

et pour $0 \leq j \leq n-k$

$$b_{j,k} = C_n^j p_0^j (1-p_0)^{n-j} C_{n-j}^{k} p^{k} (1-p)^{n-j-k}.$$

Ensuite nous prouvons que l'espérance, la *Value-at-Risk* à n'importe quel seuil $\alpha \in (0,1)$ et les primes stop-loss $\mathbb{E}\left[(M-m)_+\right]$ pour $0 \leq m \leq n$ sont croissantes en p et en Δr. Pour cela, nous formulons la proposition suivante :

Proposition 1. *Lorsque le paramètre de corrélation est fixé, le nombre de rachats est stochastiquement croissant en p : pour $p_0 \in (0,1)$ fixé, si $p < p'$ alors*

$$M_{(p,p_0)} \leq_1 M_{(p',p_0)}.$$

D'autre part, la variance et les primes stop-loss $\mathbb{E}\left[(M-m)_+\right]$ pour $0 \leq m \leq n$ sont également croissantes en p_0 (à p fixé). Ce résultat se retrouve grâce à la proposition :

Proposition 2. *Lorsque la probabilité individuelle de rachat p est fixée, le paramètre de corrélation induit un ordre 2-convex du nombre de rachats : pour $p \in (0,1)$ fixé, si $p_0 < p_0'$ alors*

$$M_{(p,p_0)} \leq_2 M_{(p,p_0')}.$$

Cette proposition montre aussi que si $p_0 < p_0'$, il existe un niveau $\alpha_0 \in (0,1)$ tel que pour $\alpha > \alpha_0$,

$$VaR_\alpha\left(M_{(p,p_0)}\right) < VaR_\alpha\left(M_{(p,p_0')}\right).$$

Enfin, nous déduisons de ces deux résultats que l'écart de taux (Δr) induit un ordre convexe croissant du nombre de rachats, ainsi que des propriétés intéressantes sur les moments de la distribution du nombre de rachats et certaines mesures de risque. D'un point de vue plus pratique, les propriétés sur la *Value-at-Risk* et les primes *stop-loss* seront les résultats d'intérêt pour un assureur dans une optique de gestion de risque.
La mise en oeuvre de ces théories permet par exemple de se rendre compte que le risque majeur en termes de besoin en capital se traduit lors de l'apparition de la corrélation dans un contexte ordinaire, ou même que la taille du portefeuille ne permet pas de mutualiser le risque de corrélation des comportements de rachat contrairement aux idées reçues.

Le chapitre 3 établit le lien entre les idées développées aux chapitres 1 et 2 : nous présentons la théorie des modèles mélanges appliquée à notre problématique, et proposons une méthodologie d'étude pour la modélisation probabiliste des comportements de rachat sur un portefeuille réel d'Assurance-Vie. Pour cela, nous effectuons des mélanges de régressions logistiques, dont la densité s'exprime comme

$$f(y_j) = \sum_{i=1}^{G} \pi_i f_i(y_j),$$

où π_i est la proportion de la composante i dont la densité est donnée par

$$f_i(y_j) = f(y_j; p_i(X_j)) = P(Y_j = y_j) = C_{N_j}^{y_j} p_i(X_j)^{y_j}(1-p_i(X_j))^{N_j - y_j}, \quad \forall i \in [\![1, G]\!].$$

y_j est le nombre de rachats observés dans le groupe homogène j (comportant N_j assurés) de la composante i, et $p_i(X_j)$ résulte du lien logistique

$$p_i(X_j) = \frac{\exp(X_j^T \beta_i)}{1 + \exp(X_j^T \beta_i)},$$

avec $X_j = (X_{j1}, ..., X_{jp})^T$ le vecteur des p covariables de l'individu j, $\beta_i = (\beta_{i1}, ..., \beta_{ip})^T$ le vecteur des p coefficients de régression de la composante i.

Notre principal apport provient du raisonnement que nous mettons en place dans la construction du modèle, et qui nous permet cette fois de modéliser correctement les décisions individuelles de rachat. Nous réduisons la dimension de l'espace des paramètres par des techniques d'arbres de classification, pour éviter de supposer d'entrée un lien entre variable observée Y et variables explicatives X. Des études empiriques viennent renforcer la connaissance de l'historique du produit considéré afin d'effectuer des choix dans la modélisation mélange, puis nous comparons cette approche avec celle du chapitre 1. C'est finalement ce modèle théorique qui est adopté pour répondre à la problématique initiale de l'entreprise, car il donne d'excellents résultats pour l'ensemble des produits (lorsque nous le confrontons à l'expérience d'AXA sous forme de *back-tests*). La clef vient de la manière de considérer les effets dans le mélange, ce qui constitue l'aboutissement de la thèse d'un point de vue opérationnel. Suite à cette trouvaille, nous avons dû implémenter un logiciel capable de mener cette étude sur n'importe quel produit dans quatre pays différents (Espagne, USA, Suisse et Belgique).

S'étant aperçu que certains problèmes subsistaient dans l'étape de sélection de modèle, le chapitre 4 se penche sur l'étude des propriétés des critères de sélection utilisés. En effet, nous observions fréquemment une potentielle surestimation du nombre de composantes du mélange, qui se traduisait par des composantes relativement similaires dans un même mélange. L'une de nos problématiques étant d'établir un clustering de notre population, il semblait inévitable de tenter de regrouper ces composantes afin de refléter les grandes catégories de comportement présents dans nos portefeuille. Après avoir repris dans le détail l'historique de la construction de l'estimateur du maximum de vraisemblance et des célèbres critères AIC et BIC, nous présentons un nouvel estimateur qui répond précisément à notre désir de partitionnement non-supervisé. Il s'agit de l'estimateur du maximum de vraisemblance classifiante conditionnelle, issue de la log-vraisemblance classifiante conditionnelle définie par

$$\log L_{cc}(\psi_G; y) = \log L(\psi_G; y) + \sum_{j=1}^{n}\sum_{i=1}^{G} \mathbb{E}_Z[z_{ij}|X] \log \tau_i(y_j; \psi_G).$$

Cette quantité n'est rien d'autre que la log-vraisemblance classique, pénalisée par un terme qui permet de mesurer la confiance du modèle lors de l'affectation des observations à telle ou telle composante (pour la formation de clusters). Par définition, l'estimateur du maximum de vraisemblance classifiante conditionnelle $ML_{cc}E$ dans un espace de paramètre Ψ_G satisfait

$$\hat{\psi}_G^{ML_{cc}E} = \underset{\psi_G \in \Psi_G}{\arg\max}\, \mathbb{E}_{f^0}[\log L_{cc}(\psi_G, Y)],$$

et est approché de manière empirique grâce à la loi des grands nombres par

$$\hat{\psi}_G^{ML_{cc}E} = \underset{\psi_G \in \Psi_G}{\arg\max}\, \frac{1}{n}\sum_{j=1}^{n} \log L_{cc}(\psi_G; y_j).$$

L'étude des propriétés de cet estimateur s'est faite dans le cadre des mélanges de GLMs (plutôt que l'unique cas des mélanges de logistiques) afin de garantir une plus grande généralité à nos résultats. C'est ainsi que la théorie que nous développons s'inscrit dans le cadre des mélanges définis par l'ensemble de modèles M_G, où

$$M_G = \left\{ f(.;\psi_G) = \sum_{i=1}^{G} \pi_i\, f_{glm}(.;\theta_i,\phi_i) \mid \psi_G = (\pi_1,...,\pi_G,\theta_1,...,\theta_G,\phi_1,...,\phi_G) \in \Psi_G \right\},$$

avec $\Psi_G \subset \Pi_G \times (\mathbb{R}^d)^{2G}$, et $f_{glm}(y;\theta_i,\phi_i) = \exp\left\{\dfrac{y\theta_i - b(\theta_i)}{a(\phi_i)} + c(y;\phi_i)\right\}$.

Par l'étude des propriétés de la vraisemblance classifiante conditionnelle ainsi que celles des lois de la famille exponentielle qui interviennent dans les modèles GLM, nous prouvons la convergence forte de l'estimateur $ML_{cc}E$ vers l'oracle, meilleur paramètre théorique pour l'optimisation de la vraisemblance classifiante conditionnelle avec des mélanges. Les conditions nécessaires à ces résultats sont des conditions de régularité sur la log-vraisemblance classifiante conditionnelle, couplées à des conditions de compacité et/ou convexité sur l'espace des paramètres Ψ_G. Il s'ensuit le théorème de convergence suivant (les notations sont introduites dans les chapitres) :

Théorème 1. *Plaçons nous dans l'ensemble de modèles M_G.*
Soit l'espace des paramètres Ψ_G de dimension K_G, tel que $\Psi_G \subset \mathbb{R}^{K_G}$.
Soit $\log L_{cc} : \Psi_G \times \mathbb{R}^d \to \mathbb{R}$. Soit $\Psi_G^{\mathcal{O}}$ un ouvert de $\mathbb{R}^{K_G}$ sur lequel $\log L_{cc}$ est bien définie, et tel que $\Psi_G \subset \Psi_G^{\mathcal{O}}$.

Si nous avons les trois hypothèses suivantes :

(H1-B) : Supposons Ψ_G compact. Alors $\Psi_G^b = \left\{\psi_G^b : \psi_G^b = \underset{\psi_G \in \Psi_G}{\arg\max}\, \mathbb{E}_{f^0}\left[\log L_{cc}(\psi_G; Y)\right]\right\}$.

(H2-B) : Supposons que $L^{'}(y) = \underset{\psi_G \in \Psi_G^{\mathcal{O}}}{\sup} \left|\left|\left(\dfrac{\partial \log L_{cc}}{\partial \psi_G}\right)_{(\psi_G;y)}\right|\right|_\infty < \infty$.

(H3-B) : Supposons également que $\|L^{'}\|_1 < \infty$.

Alors,

$\exists n_0 \in \mathbb{N}, \forall n \geq n_0, \forall \psi_G^b \in \Psi_G^b$, *en définissant* $\hat{\psi}_G^{ML_{cc}E} = \hat{\psi}_G^{ML_{cc}E}(Y_1, ..., Y_n) \in \Psi_G$ *un estimateur qui maximise presque la vraisemblance classifiante conditionnelle tel que*

$$\frac{1}{n}\log L_{cc}(\hat{\psi}_G^{ML_{cc}E}; Y) \geq \frac{1}{n}\log L_{cc}(\psi_G^b; Y) - \xi_n$$

$$\text{avec } \begin{cases} \xi_n \geq 0 \quad p.s. \\ \xi_n \underset{n\to\infty}{\longrightarrow} 0 \ p.s. \end{cases} \quad \text{, nous avons} \quad d(\hat{\psi}_G^{ML_{cc}E}, \Psi_G^b) \underset{n\to\infty}{\longrightarrow} 0 \ p.s.$$

En ce qui concerne la problématique de sélection de modèle, nous nous intéressons au critère de sélection ICL, taillé sur mesure pour la question de classification et défini initialement comme

$$M_{ICL} = \underset{M_g \in \{M_1,...,M_m\}}{\arg\min} \left(-\log L(\hat{\psi}_g^{MLE}; y) - \sum_{j=1}^{n}\sum_{i=1}^{g} \hat{Z}_{ij}^{MAP} \log \tau_i(y_j; \hat{\psi}_g^{MLE}) + \frac{K_g}{2}\log n \right),$$

où les M_g sont des modèles mélanges emboîtés (un mélange à deux composantes est emboîté dans un mélange à trois composantes, et ainsi de suite).
Ce critère, qui n'était pas consistant pour l'estimation du nombre de composantes d'un mélange lorsqu'il était basé sur l'estimateur du maximum de vraisemblance, le devient sous certaines conditions en utilisant l'estimateur $ML_{cc}E$. Nous démontrons ainsi que le critère ICL appliqué

en l'estimateur $ML_{cc}E$ est consistant (au sens de la dimension) avec la modélisation mélange de GLMs, et permet donc de sélectionner un modèle dont le nombre de composantes tend à être le nombre théorique de composantes de la loi sous-jacente. Le théorème suivant illustre ce résultat :

Théorème 2. *Plaçons nous dans l'ensemble de modèles M_G.*
Soit $\{M_g\}_{1\leq g\leq m}$ une collection de modèles de paramètres $\{\psi_g\}_{1\leq g\leq m} \in \{\Psi_g\}_{1\leq g\leq m}$ et de dimension $\{K_g\}_{1\leq g\leq m}$, avec $\Psi_g \subset \mathbb{R}^{K_g}$. Ces modèles sont classés dans un ordre croissant de complexité, avec $K_1 \leq K_2 \leq ... \leq K_m$. Supposons que

(H1-D) $\forall g \in [\![1,m]\!]$, Ψ_G *est un ensemble compact.*
Alors pour g quelconque, posons $\Psi_g^b = \arg\max_{\psi_g\in\Psi_g} \mathbb{E}_{f^0}\left[\log L_{cc}(\psi_g;Y)\right]$.
Définissons $g^b = \min\left(\arg\max_{1\leq g\leq m} \mathbb{E}_{f^0}\left[\log L_{cc}(\Psi_g^b;Y)\right]\right)$;

(H2-D) *Supposons que* $\forall g \in [\![1;m]\!]$, $\forall \psi_g \in \Psi_g$, $\forall \psi_{g^b}^b \in \Psi_{g^b}^b$,
$\mathbb{E}_{f^0}[\log L_{cc}(\psi_g)] = \mathbb{E}_{f^0}[\log L_{cc}(\psi_{g^b}^b)] \Longleftrightarrow \log L_{cc}(\psi_g;y) = \log L_{cc}(\psi_{g^b}^b;y) \quad f^0 d\lambda$- *p.s.*

Pour g quelconque, soit $\Psi_g^{\mathcal{O}}$ un ouvert de $\mathbb{R}^{K_g}$ sur lequel $\log L_{cc}$ est définie, avec $\Psi_g \subset \Psi_g^{\mathcal{O}}$.

(H3-D) *Supposons que* $\forall g \in [\![1,m]\!]$, $\begin{cases} L_g(y) = \sup\limits_{\psi_g\in\Psi_g^{\mathcal{O}}} |\log L_{cc}(\psi_g;y)| < \infty \quad f^0 d\lambda\text{-p.s.}, \\ \|L_g\|_\infty < \infty. \end{cases}$

(H4-D) *Supposons que* $\forall g \in [\![1,m]\!]$, $\begin{cases} L_g'(y) = \sup\limits_{\psi_g\in\Psi_g^{\mathcal{O}}} \left|\left|\left(\dfrac{\partial \log L_{cc}}{\partial \psi_g}\right)_{(\psi_g;y)}\right|\right|_\infty < \infty \quad f^0 d\lambda\text{-p.s.}, \\ \|L_g'\|_2 < \infty. \end{cases}$

(H5-D) *Supposons que* $\forall g \in [\![1,m]\!], \forall \psi_g^b \in \Psi_g^b$, $I_{\psi_g^b} = \dfrac{\partial^2}{\partial \psi_g^2}\Big(\mathbb{E}_{f^0}[\log L_{cc}(\psi_g;y)]\Big)_{|\psi_g^b}$ *est inversible.*

(H6-D) *Supposons que* $\forall g \in [\![1,m]\!]$, $\begin{cases} pen(K_g) > 0 \text{ et } pen(K_g) = o_{\mathbb{P}}(n) \text{ quand } n \to +\infty; \\ \Big(pen(K_g) - pen(K_{g'})\Big) \xrightarrow{\mathbb{P}} \infty \text{ quand } n \to +\infty \text{ et } g > g'. \end{cases}$

Quel que soit g et n, soit $\hat{\psi}_g^{ML_{cc}E} = \hat{\psi}_g^{ML_{cc}E}(Y_1,...,Y_n) \in \Psi_g$ un estimateur tel que

$$\log L_{cc}(\hat{\psi}_g^{ML_{cc}E};Y) \geq \log L_{cc}(\psi_g^b;Y) - o_{\mathbb{P}}(n);$$

Sélectionnons $\hat{g}$ tel que

$$\hat{g} = \arg\min_{1\leq g\leq m}\{-\log L_{cc}(\hat{\psi}_g^{ML_{cc}E};y) + pen(K_g)\},$$

Alors

$$\mathbb{P}(\hat{g} \neq g^b) \underset{n\to\infty}{\longrightarrow} 0.$$

Les hypothèses nécessaires à un tel résultat sont tout à fait réalistes dans le champ des applications connues à ce jour. Enfin, la mise en pratique de ces théories sur nos différentes familles de produit montre que le critère de sélection ICL permet effectivement d'obtenir une modélisation mélange finale dont le nombre de composantes est inférieur à celui donné par une sélection BIC, tout en conservant une bonne qualité d'adéquation aux données mais aussi de bonnes propriétés prédictives.

Bibliographie

Albert, F. S., Bragg, D. G. W. and Bragg, J. M. (1999), 'Mortality rates as a function of lapse rates', *Actuarial research clearing house* **1**. 7

Atkins, D. C. and Gallop, R. J. (2007), 'Re-thinking how family researchers model infrequent outcomes : A tutorial on count regression and zero-inflated models', *Journal of Family Psychology* . 7

Bacinello, A. R. (2005), 'Endogenous model of surrender conditions in equity-linked life insurance', *Insurance : Mathematics and Economics* **37**, 270–296. 4, 7

Bacinello, A. R., Biffis, E. and P., M. (2008), 'Pricing life insurance contracts with early exercise features', *Journal of Computational and Applied Mathematics* . 7

Bluhm, W. F. (1982), 'Cumulative antiselection theory', *Transactions of Society of actuaries* **34**. 4

Costabile, M., Massabo, I. and Russo, E. (2008), 'A binomial model for valuing equity-linked policies embedding surrender options', *Insurance : Mathematics and Economics* **40**, 873–886. 7

Cox, D. (1972), 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society : Series B* (34), 187–220. 8

Cox, S. H. and Lin, Y. (2006), Annuity lapse rate modeling : tobit or not tobit ?, *in* 'Society of actuaries'. 6

Cummins, J. (1975), *An econometric model of the life insurance sector in the U.S. economy*, Lexington books, Health, Lexington/Mass u.a. 6

De Giovanni, D. (2007), Lapse rate modeling : A rational expectation approach, Finance Research Group Working Papers F-2007-03, University of Aarhus, Aarhus School of Business, Department of Business Studies. 7

Dutang, C. (2011), Regression models of price elasticity in non-life insurance, Master's thesis, ISFA. Mémoire confidentiel - AXA Group Risk Management. 7

Engle, R. and Granger, C. (1987), 'Cointegration and error-correction : Representation, estimation and testing', *Econometrica* (55), 251–276. 7

Fauvel, S. and Le Pévédic, M. (2007), Analyse des rachats d'un portefeuille vie individuelle : Approche théorique et application pratique, Master's thesis, ENSAE. Mémoire non confidentiel - AXA France. 7

Fum, D., Del Missier, F. and A., S. (2007), 'The cognitive modeling of human behavior : Why a model is (sometimes) better than 10,000 words', *Cognitive Systems Research* **8**, 135–142. 4

Hin, H. K. and Huiyong, S. (2006), 'Structural prepayment risk behavior of the underlying mortgages for residential mortgage life insurance in a developing market', *Journal of Housing Economics* (15), 257–278. 8

Kagraoka, Y. (2005), Modeling insurance surrenders by the negative binomial model. Working Paper 2005. 6

Kim, C. (2005), 'Modeling surrender and lapse rates with economic variables', *North American Actuarial Journal* pp. 56–70. 6

Kim, C. N., Yang, K. H. and Kim, J. (2008), 'Human decision-making behavior and modeling effects', *Decision Support Systems* **45**, 517–527. 4

Kuen, S. T. (2005), 'Fair valuation of participating policies with surrender options and regime switching', *Insurance : Mathematics and Economics* **37**, 533–552. 4, 8

Lee, S., Son, Y.-J. and Jin, J. (2008), 'Decision field theory extensions for behavior modeling in dynamic environment using bayesian belief network', *Information Sciences* **178**, 2297–2314. 4

Loisel, S. and Milhaud, X. (2011), 'From deterministic to stochastic surrender risk models : Impact of correlation crises on economic capital', *European Journal of Operational Research* **214**(2). 13

McNeil, A., Frey, R. and Embrechts, P. (2005), *Quantitative Risk Management*, Princeton Series In Finance. 4

Milhaud, X., Gonon, M.-P. and Loisel, S. (2010), 'Les comportements de rachat en assurance vie en régime de croisière et en période de crise', *Risques* (83), 76–81. 3

Milhaud, X., Maume-Deschamps, V. and Loisel, S. (2011), 'Surrender triggers in life insurance : what main features affect the surrender behavior in a classical economic context ?', *Bulletin Francais d'Actuariat* **11**(22), 5–48. 12

Nordahl, H. A. (2008), 'Valuation of life insurance surrender and exchange options', *Insurance : Mathematics and Economics* **42**, 909–919. 8

Outreville, J. F. (1990), 'Whole-life insurance lapse rates and the emergency fund hypothesis', *Insurance : Mathematics and Economics* **9**, 249–255. 6

Pan, X., Han, C. S., Dauber, K. and Law, K. H. (2006), 'Human and social behavior in computational modeling and analysis of egress', *Automation in Construction* **15**, 448–461. 4

Pesando, J. (1974), 'The interest sensibility of the flow of funds through life insurance companies : An econometric analysis', *Journal Of Finance* **Sept**, 1105–1121. 6

Renshaw, A. E. and Haberman, S. (1986), 'Statistical analysis of life assurance lapses', *Journal of the Institute of Actuaries* **113**, 459–497. 6

Shen, W. and Xu, H. (2005), 'The valuation of unit-linked policies with or without surrender options', *Insurance : Mathematics and Economics* **36**, 79–92. 8

Stanton, R. (1995), 'Rational prepayment and the valuation of mortgage-backed securities', *Review of Financial* **8**, 677–708. 8

Torsten, K. (2009), Valuation and hedging of participating life-insurance policies under management discretion, *in* 'Insurance : Mathematics and Economics Proceedings', Vol. 44, pp. 78–87. 8

Tsai, C., Kuo, W. and Chen, W.-K. (2002), 'Early surrender and the distribution of policy reserves', *Insurance : Mathematics and Economics* **31**, 429–445. 4, 7

Vandaele, N. and Vanmaele, M. (2008), 'Explicit portfolio for unit-linked life insurance contracts with surrender option', *Journal of Computational and Applied Mathematics* . 4, 8

Première partie

La modélisation comportementale, une problématique complexe

Chapitre 1

Segmentation du risque de rachat

Ce chapitre s'inspire de l'article Milhaud et al. (2011), publié dans le *Bulletin Français d'Actuariat 22, vol. 11, p 5-48*. Nous l'avons coécrit avec mes directeurs de thèse Véronique Maume-Deschamps et Stéphane Loisel.

Comme nous l'avons déjà évoqué, la compréhension de la dynamique des taux de rachat est cruciale pour les compagnies d'assurance qui doivent faire face à plusieurs problèmes qui y sont liés. Il y a tout d'abord la problématique des rachats anticipés qui entraînent l'impossibilité pour la compagnie de recouvrir ses frais d'émission, de gestion et d'administration du nouveau contrat (environ 3,5%). En effet, l'assureur paie ces frais avant ou à l'émission du contrat et espère faire des profits au cours de la vie du contrat, ces profits n'étant pas réalisés en cas de rachat précoce. Nous constatons ainsi que le profil temporel du rachat a une importance toute particulière, puisque de ce profil vont dépendre les coûts du rachat pour l'assureur. De plus, les assurés qui ont certains problèmes de santé et d'assurabilité auront tendance à ne pas racheter leur contrat, causant finalement plus de sinistres que prévu (phénomène d'anti-sélection). Enfin il existe toujours le risque de taux d'intérêts : au cours de la vie des contrats, ces taux varient. Plaçons nous par exemple dans un contexte de contrat d'épargne à taux garanti : si les taux d'intérêts s'effondrent, l'assureur doit tenir ses engagements et verser aux assurés un taux garanti supérieur au rendement de ses actifs, le risque étant donc que les rachats soient beaucoup moins nombreux que prévu et que l'assureur manque de liquidité. Inversement, les assurés seront plus à même de racheter leur contrat en cas de hausse des taux car les nouveaux contrats offriront de meilleurs rendements à niveau de garantie équivalent. L'assureur devra donc rembourser aux assurés la valeur de rachat de ces contrats dans un contexte où l'emprunt d'argent peut s'avérer très coûteux ! Finalement l'assureur peut subir une série d'effets indésirables en cascade : pas le temps de recouvrir ses frais, obligation d'emprunter à prix fort et nécessité de liquider ses actifs au pire moment (cependant les rachats ne sont pas qu'une mauvaise nouvelle pour l'assureur puisque celui-ci se débarrasse de garanties qui ont un coût). Heureusement les observateurs noteront que l'agent (assuré) n'est en général ni rationel ni optimal, même si ces comportements sous-optimaux tendent à disparaître du fait d'une information toujours plus accessible.

Toutes ces considérations montrent que l'enjeu d'une modélisation précise des comportements de rachat est primordiale en termes de profitabilité et de solvabilité pour l'assureur. Les praticiens fixent des hypothèses de rachat, fruit de l'étude statistique de la collection de données expérimentales rendue complexe par l'essence même de celles-ci : les différents types de données, leur dimension, la gestion des données manquantes... Le défi est donc de sélectionner

le minimum de données apportant le maximum d'information, ce que nous essayons de faire dans ce chapitre par l'usage de deux modèles complémentaires de segmentation : les arbres de classification et de régression (CART), et la régression logistique (LR).

La méthode CART développée par Breiman et al. (1984) et la LR (Hilbe (2009)) nous ont permis de prouver avec différents portefeuilles d'Assurance-Vie d'AXA le pouvoir discriminant de certaines caractéristiques sur la décision de rachat. Nous présentons rapidement dans un premier temps les fondamentaux de chacun des modèles ainsi que leurs hypothèses et limites (de nombreux détails théoriques sont donnés en annexe B). Nous discutons au final des différences entre les deux modélisations en termes de résultats numériques et d'un point de vue opérationnel, et justifions l'emploi d'autres modélisations plus ajustées dans la suite du manuscript. Le but de ce chapitre est donc de i) réduire la dimension de l'espace des variables à prendre en compte dans une future modélisation, ii) déterminer quelle méthode semble la plus adaptée en regardant les taux d'erreur de classification, iii) trouver les déclencheurs essentiels du rachat en régime de croisière (économique).

Nous gardons à l'esprit que cette segmentation ne représente pas la réalité en période de crise (financière, d'image) et introduit un biais non-négligeable car nous n'y considérons pas le contexte économique. Les effets "cohortes" ne sont également pas pris en compte puisque la date de rachat n'intervient pas en facteur explicatif. Nous reviendrons sur ces remarques pour proposer des extensions possibles lors de prévisions futures de taux de rachat incluant des facteurs dynamiques. Cette première segmentation est utile à plusieurs titres : elle permet de mieux comprendre les comportements assurés, de planifier une segmentation du risque de rachat et d'améliorer le design de nouveaux produits (hypothèse de taux moyen de rachat, clauses et options des contrats). L'exemple considéré sert ici à alimenter la théorie pour une meilleure compréhension, les résultats de la segmentation pour l'ensemble des produits sont fournis en annexe C.

1.1 Modélisation CART

La méthode CART, outil non paramétrique flexible, est basée sur un algorithme à la fois itératif et récursif. Développée par Breiman et al. (1984) dans le but de diviser les données d'origine à l'aide de règles déterministes, ses arbres binaires offrent une manière puissante et conviviale de fournir des résultats dans les problèmes de classification. La particularité de l'algorithme CART en comparaison à ses "congénéres" est qu'il n'existe pas de règle d'arrêt lors de la construction de l'arbre. De manière générale, les deux principaux buts d'un processus de classification sont de produire un bon classifieur et d'avoir un bon pouvoir prédictif. Nous entendons par "bon" une procédure qui amène à des erreurs acceptables, bien que nous verrons qu'un arbitrage entre ces deux notions doit être fait par l'utilisateur dans le sens où un gain en précision de classification entraîne généralement une perte de pouvoir prédictif. CART se révèle être très utile, mais l'utilisateur doit être conscient que plusieurs modèles de segmentation doivent être utilisés dans l'idéal pour obtenir un résultat robuste.

1.1.1 Le modèle

Nous présentons dans cette section comment construire l'arbre. La figure B.1 de l'annexe B.1 indique et détaille les différentes étapes à suivre, ainsi que les concepts sous-jacents pour le lecteur que cela intéresserait. Nous trouvons intéressant de fournir une chronologie claire de l'algorithme CART car elle n'apparaît pas explicitement dans la littérature.

Construction de l'arbre de classification

Notation Soit $\epsilon = (x_n, j_n)_{1 \leq n \leq N}$ un échantillon de taille N, où les j_n représentent les observations de la variable réponse Y ($Y \in C = \{1, 2, ..., J\}$ et les $x_n = \{x_{n_1}, x_{n_2}, ..., x_{n_p}\}$ sont les observations de X dans $\mathcal{X}$, ensemble des p variables explicatives ($\mathcal{X} = \prod_{i=1}^{p} \mathcal{X}_i$ où $\mathcal{X}_i$ est un ensemble de variables continues et/ou catégorielles). Soit

- $\forall x \in \mathcal{X}$, le classifieur $class(., \epsilon)$ classe x dans un groupe j $\in$ C.
- La probabilité a priori d'appartenir au groupe j vaut $\pi_j = \frac{N_j}{N}$ où $N_j = \text{Card}\{j_n | j_n = j\}$.
- Sachant $t \subset \mathcal{X}$ (t sous-ensemble fini de $\mathcal{X}$), notons $N(t) = \text{Card}\{(x_n, j_n) \in \epsilon,\ x_n \in t\}$.
- $N_j(t) = \text{Card}\{(x_n, j_n) \in \epsilon,\ j_n$=j sachant que $x_n \in t\}$.
- Un estimateur par substitution de P(j,t), noté p(j,t), est donné par p(j,t) = $\pi_j \frac{N_j(t)}{N(t)}$.
- Un estimateur par substitution de P(t), noté p(t), est donné par p(t) = $\sum_{j=1}^{J} p(j, t)$.
- Soit $P(j|t)$ la probabilité a posteriori d'appartenir à j, estimée par $\frac{p(j,t)}{p(t)} = \frac{N_j(t)}{N(t)} = \frac{p(j,t)}{\pi_j}$.

Comment débuter ? Le principe est de diviser $\mathcal{X}$ en q classes, où q n'est pas connu à l'avance (*a-priori*). La méthode construit une séquence croissante de partitions de $\mathcal{X}$; On passe d'une partition à l'autre en appliquant des *règles de division binaires* telles que :

$$x \in t, \text{ avec } t\ \subset \mathcal{X}.$$

Par exemple, la première partition de $\mathcal{X}$ peut être le sexe de l'assuré. L'assuré dont la caractéristique est x est soit une femme soit un homme (une spécification des règles binaires est détaillée en annexe B.1.3).

Nous commençons par diviser la *racine* $\mathcal{X}$ en deux sous-ensembles disjoints appelés *noeuds* et notés t_L et t_R. Chaque noeud est ensuite divisé de la même manière (s'il contient au moins deux éléments). Au final nous obtenons une partition de $\mathcal{X}$ en q groupes appelés *noeuds terminaux* ou *feuilles*. Dans la suite, nous notons $\tilde{T}$ l'ensemble des *feuilles* de l'arbre T ; T^t est l'ensemble des *descendants* de *l'ancêtre* t dans l'arbre T (voir l'illustration en figure B.2).

Nous mesurons la qualité de la division d'un noeud t en t_L et t_R grâce à un *critère d'impureté*. Ce concept est également expliqué en détail en annexe B.1.3. Dans notre cas, l'impureté du noeud t dans l'arbre T est la quantité

$$impur(t) = g(p(1|t), p(2|t), ..., p(J|t)), \tag{1.1}$$

où g est la fonction d'impureté. Par conséquent, l'impureté de l'arbre T est donnée par

$$Impur(T) = \sum_{t \in \tilde{T}} Impur(t) \tag{1.2}$$

où $Impur(t) = p(t) impur(t)$.
Une règle de division Δ d'un noeud t donne p_L=$p(t_L)/p(t)$ observations dans t_L, et p_R=$p(t_R)/p(t)$ observations dans t_R. Nous aimerions maximiser la *pureté*, dont la variation due à la division vaut :

$$\delta\, impur(\Delta, t) = impur(t) - p_L impur(t_L) - p_R impur(t_R) \tag{1.3}$$

La pureté de l'arbre est censée augmenter à chaque division, ce qui impose la contrainte naturelle suivante :

$$impur(t) \geq p_L\ impur(t_L) + p_R\ impur(t_R).$$

Respectons-nous toujours cette inégalité ? La réponse est "oui" si g est concave. Dans la plupart des applications (y compris la nôtre), nous considérons l'index de diversité de Gini, interprétable comme une probabilité de mauvaise classification. C'est la probabilité d'affecter la classe k à une observation choisie aléatoirement dans le noeud t, multipliée par la probabilité estimée que cette observation appartienne en réalité à la classe j. Il existe aussi d'autres fonctions d'impureté qui ont une interprétation encore plus simple (annexe B.1.3), mais il n'existe pas de justification particulière pour l'usage de telle ou telle fonction (en particulier elles sont toute concaves, et les propriétés de l'arbre final ne sont pas vraiment impactées par ce choix, comme décrit dans Breiman et al. (1984)). La division optimale $\Delta_t^\star$ d'un noeud t satisfait

$$\Delta_t^\star = \underset{\Delta \in D}{\arg\max}(\delta\, impur(\Delta, t)), \tag{1.4}$$

où $\arg\max(\delta\, impur(\Delta, t))$ désigne la règle de division Δ qui maximise $\delta\, impur(\Delta, t)$.

Le processus génère donc une décroissance d'impureté aussi rapide que possible à chaque étape. Intuitivement, cela signifie qu'un maximum d'observations doivent appartenir à la même classe dans un noeud donné, ce qui définit la règle de division à choisir. Maximiser le gain de pureté (ou d'homogénéité) par la division du noeud t revient à maximiser le gain de pureté de l'arbre T. Nous obtenons ainsi un arbre $T^{'}$ (figure B.2 en annexe) plus ramifié en partant de l'ancêtre t vers les descendants (t_L,t_R) par Δ, et (1.2) donne

$$Impur(T^{'}) = \sum_{w \in \tilde{T} - \{t\}} Impur(w) + Impur(t_L) + Impur(t_R).$$

Nous en déduisons une fluctuation d'impureté au niveau de l'arbre T de

$$\begin{aligned} F &= Impur(t) - Impur(t_L) - Impur(t_R) \\ &= \delta\, Impur(\Delta, t) \\ &= p(t)\, \delta\, impur(\Delta, t). \end{aligned} \tag{1.5}$$

Il s'agit donc de la probabilité d'être présent dans ce noeud multipliée par la décroissance d'impureté donnée par Δ. L'étape suivante consiste à définir quand arrêter les divisions, ce qui relève du choix de l'utilisateur. Certaines règles d'arrêt sont naturelles tandis que d'autres sont purement arbitraires : i) les divisions s'arrêtent évidemment lorsque les observations des variables explicatives dans une classe donnée sont identiques ; ii) on peut définir un nombre minimal d'observations dans un noeud (plus ce nombre est petit et plus le nombre de feuilles sera grand) ; iii) on peut choisir un seuil λ de décroissance minimum de l'impureté :

$$\text{Soit } \lambda \in \mathbb{R}_+^*, \quad \max_{\Delta \in D} \delta\, Impur(\Delta, t) < \lambda \Rightarrow \text{arrêter la division.}$$

Comme énoncé au début de cette section, il n'y a en fait pas de régle d'arrêt dans l'algorithme CART ; l'arbre le plus ramifié est construit, puis élagué par une procédure avancée que nous détaillons en annexe B.1.3.

La fonction de classification

Le but est de construire une méthode permettant de classer les assurés (sachant leurs caractéristiques x) dans un ensemble B_j, afin de prédire la réponse qui leur est associée. Ici,

nous entendons réponse par groupe d'appartenance, ce qui se traduira dans nos applications par le rachat ou non-rachat. Le classifieur, noté $class(.,\epsilon)$, s'exprime comme

$$\begin{aligned} class \quad &: \quad \mathcal{X} \to C \\ &\quad\quad x \to class(x,\epsilon) = j, \end{aligned}$$

avec $B_j = \{x \in \mathcal{X};\ class(x,\epsilon) = j\}$. Cette fonction doit si possible classer au mieux les données et avoir un pouvoir prédictif intéressant. Considérons que l'arbre optimal a été construit ; pour connaître la classe d'appartenance d'un noeud terminal, nous utilisons la règle

$$class(x,\epsilon) = \arg\max_{j \in C} p(j|t), \tag{1.6}$$

autrement dit la fameuse règle de *Bayes* qui maximise la probabilité *a posteriori* d'être dans la classe j sachant que nous sommes dans le noeud t. Ce processus nous permet ainsi d'effectuer des prévisions de classification. Une estimation de la mauvaise classification d'une observation dans le noeud t (par rapport à la classe observée) est donnée par

$$r(t) = 1 - class(x,\epsilon) = 1 - \max_{j \in C} p(j|t), \tag{1.7}$$

Soit $\hat{\tau}(t) = p(t)\, r(t)$ le taux de mauvaise classification du noeud t. Pour chaque noeud, c'est la probabilité d'être dans le noeud t multipliée par la probabilité de mal classer une observation sachant que nous sommes dans ce noeud t. Nous en déduisons immédiatement le taux global de mauvaise classification de l'arbre T, donné par

$$\hat{\tau}(T) = \sum_{t \in \tilde{T}} \hat{\tau}(t). \tag{1.8}$$

Finalement, nous pouvons résumer les quatre étapes essentielles de la procédure de construction de l'arbre :

1. un ensemble de questions binaires $\{x \in S\,?\}$, $S \in \mathcal{X}$,
2. une fonction d'impureté pour le critère de qualité d'ajustement (choix arbitraire),
3. une règle d'arrêt des divisions (choix arbitraire),
4. une procédure de classification permettant d'affecter à chaque feuille une classe.

De fait, le choix arbitraire concernant la règle d'arrêt des divisions est évitée puisque l'algorithme CART construit un arbre maximal T_{max} avant de procéder à un élagage.

Estimation de l'erreur de prévision

L'*erreur de prévision* est évaluée par la probabilité qu'une observation soit mal classée par $class(.,\epsilon)$, c'est-à-dire :

$$\tau(class) = P(class(X,\epsilon) \neq Y)$$

L'efficacité du prédicteur est basée sur l'estimation de cette erreur. Le taux de mauvaise classification réel $\tau^*(class)$ ne peut pas être estimé lorsque la procédure de classification est construite à partir de l'ensemble des données, mais il existe plusieurs estimateurs dans la littérature (Ghattas (1999)). L'expression du taux de mauvaise classification dépend évidemment de l'échantillon d'apprentissage (détails en annexe B.1.3).

- Estimation par **resubstitution** du taux de mauvaise classification de l'arbre : nous considérons toutes les observations ϵ pour l'échantillon d'apprentissage. Les résultats ne sont pas représentatifs car nous classons les mêmes données que celles qui ont servi à la construction du classifieur. C'est donc évidemment le pire estimateur.
- Estimation par **échantillon témoin ou de validation** : soit W $\subset \epsilon$ l'échantillon témoin de taille $N' < N$ (N est la taille de ϵ et en général N'=$N/3$). Nous construisons le classifieur avec l'échantillon d'apprentissage et validons son efficacité sur l'échantillon témoin. Cet estimateur est meilleur mais nécessite beaucoup de données.
- Technique des **validations croisées** : supposons ϵ divisé en K sous-groupes disjoints $(\epsilon_k)_{1\leq k\leq K}$ de même taille. Définissons K jeux de données d'apprentissage tels que $\epsilon^k = \epsilon - \epsilon_k$. L'idée est de construire une procédure de classification sur chaque ϵ^k telle que $class^k(.) = class(., \epsilon^k)$. La validation croisée est recommandée lorsque peu de données sont disponibles, car le taux d'erreur (moyenne des K taux d'erreur) s'avère bien plus réaliste.

Dans la suite $\tau(T)$ est l'erreur de prévision sur T ; et $\hat{\tau}(T)$, $\hat{\tau}^{ts}(T)$, $\hat{\tau}^{cv}(T)$ son estimation respectivement aux trois méthodes ci-dessus.

1.1.2 Limites, améliorations

La classification par arbre offre des avantages certains : i) pas de restriction sur le type de données (catégorielles ou continues) ; ii) les résultats finaux sont simples à interpréter et à visualiser ; iii) l'algorithme induit une méthode pas-à-pas automatique de sélection de variable et donc une réduction de la dimension de l'espace et de sa complexité. De plus, les transformations monotones des variables ne changent pas les résultats, et l'aspect non-paramétrique ne suppose pas de relation prédéterminée entre la variable réponse et les variables explicatives. Les intéractions entre prédicteurs sont en général bien identifiées.

Cependant, quelques inconvénients subsistent parmi lesquels le fait que les divisions soient basées sur une seule variable alors que nous pourrions penser que des combinaisons de variables seraient parfois plus adéquates, auquel cas l'algorithme serait mauvais pour représenter la structure des données. Nous pouvons également citer le fait que l'effet d'une variable peut être caché par une autre lors du choix des règles de division. Il existe des solutions pour éviter ce phénomène, comme classer l'effet potentiel de chaque variable explicative lors de la division : ce sont les *secondary* et *surrogate splits* dans la littérature (également utilisées pour des données manquantes, voir Breiman et al. (1984)). L'arbre final peut aussi être difficile à utiliser en pratique car trop ramifié (cas extrême : une observation par feuille, ce qui amène un taux de mauvaise classification nul qui n'est évidemment pas du tout réaliste !), mais l'utilisateur peut jouer sur la taille de l'arbre par l'introduction d'un coût de complexité dans l'algorithme d'élagage (annexe B.1.3) pour pallier ce problème.

Enfin, bien que CART donne une idée claire de l'importance de chaque variable explicative par lecture de l'arbre final depuis la racine jusqu'aux feuilles, Ghattas (2000*b*) critique un déficit de robustesse au regard des résultats : une légère modification des données peut engendrer différents classifieurs, une contrainte importante dans la problématique de prévision. Ainsi, une variable pourrait se révéler déterminante dans le processus de classification avec un certain jeu de données, tout en étant absente dans un jeu de données quasi-similaire ! Plusieurs solutions ont été développées face à ce problème, parmi lesquelles la validation croisée (Ghattas (2000*a*)), et les *bagging predictors* ou *arcing classifiers*. Ces deux dernières techniques sont une agrégation de type *bootstrap* de classifieurs construits sur des échantillons bootstrap. Leur

robustesse et significativité ont été validées dans diverses études ((Breiman (1996), Breiman (1994) et Breiman (1998)). Elles ont amené au développement des "forêts aléatoires" (Breiman (2001)), un algorithme que nous utiliserons dans nos applications. Pour plus de détails, consulter la page web de Breiman et la documentation de la librairie `randomForest` [3] du logiciel R, de même que Breiman et al. (1984).

1.2 Segmentation par modèle logistique (Logit)

La régression logistique (Hosmer and Lemeshow (2000), Balakrishnan (1991)) appartient à la classe des modèles linéaires généralisés (McCullagh and Nelder (1989)). Elle permet de modéliser la probabilité d'occurence d'un événement binaire à partir de covariables catégorielles ou continues, en ajustant une courbe logistique aux données. Ce modèle de choix est utilisé dans le cadre des régressions binomiales, principalement dans le domaine médical et dans le monde du marketing. Les actuaires l'utilisent également pour modéliser la mortalité (qui présente une croissance exponentielle en fonction de l'âge, non loin de la forme logistique pour de petites probabilités) à partir de données empiriques, avec pour but la segmentation de leur portefeuille. Dans notre contexte, l'objectif est de segmenter la population par rapport au risque binaire du rachat. La présentation théorique sera raccourcie étant donnée la popularité de cette modélisation ; quelques exemples d'application sont consultables dans Kagraoka (2005), ainsi que certains modèles similaires dont le modèle Tobit (Cox and Lin (2006)) ou le modèle de Cox (Cox (1972)). Pour de plus amples comparaisons de ces différents modèles, l'article d'Austin (2007) est une référence intéressante.

1.2.1 Pourquoi utiliser la régression logistique ?

La fonction logistique est très utile car elle permet d'obtenir une image $\Phi(z)$ dans [0,1] à partir d'un antécédent z prenant des valeurs sur l'ensemble de la droite des réels :

$$\Phi(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}. \tag{1.9}$$

Notre volonté de modéliser une probabilité de rachat entre complètement dans ce cadre-là, sachant de plus que la propriété de non-décroissance d'une fonction de répartition classique est respectée par la fonction logistique. L'exposition $\mathbf{z}$ à un ensemble de facteurs de risque est appelé *prédicteur linéaire*. Il est donné par l'équation de régression classique

$$z = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p = X^T \beta,$$

où les X_k sont les covariables (explicatives), par exemple l'âge. Ainsi $\forall k = 1, ..., p; \beta_k$ représente le coefficient de régression associé au facteur de risque k. Nous noterons les coefficients de régression $\beta = (\beta_0, ..., \beta_k)^T$ et le vecteur des variables $X = (1, X_1, ..., X_p)^T$.
Si l'on considère une approche stricte de régression, l'idée est de transformer la sortie d'une régression linéaire classique pour obtenir une probabilité en utilisant une fonction de lien (ici le "logit-link", mais il existe aussi d'autres liens comme le "probit", etc).

3. Disponible à http ://cran.r-project.org/web/packages/randomForest/index.html

1.2.2 Estimation des paramètres par maximum de vraisemblance

Nous avons vu que les rachats sont binomialement distribués (nombre de rachats $\sim \mathcal{B}(n, \Phi(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p))$, où n est le nombre d'individus). La méthode du maximum de vraisemblance (ML) nous permet d'estimer de manière classique les paramètres de la probabilité de rachat. Par définition, la fonction de vraisemblance pour une loi binomiale vaut

$$L(\beta; Y|X) = \prod_{j=1}^{n} \Phi(X_j^T \beta)^{Y_j} (1 - \Phi(X_j^T \beta))^{1-Y_j},$$

où Φ est définie dans (1.9), et j représente l'individu. La log-vraisemblance est donc

$$\ln(L(\beta; Y|X)) = \sum_{j=1}^{n} Y_j (X_j^T \beta) - \ln(1 + e^{X_j^T \beta}), \tag{1.10}$$

L'estimateur du maximum de vraisemblance, noté $\hat{\beta}^{MLE}$, annule le score ; lui même représenté par la dérivée en β de l'équation (1.10). Cette condition amène généralement à un système d'équations dont la solution n'est pas explicite et n'admet donc pas de formule fermée. La résolution de ce système passe par l'utilisation d'algorithmes d'optimisation tels que celui de Newton-Raphson, détaillé en annexes B.2.3 et B.2.4.

L'estimation de la probabilité individuelle de rachat découle directement des estimations des coefficients de régression par

$$\hat{p}_j = \Phi(\hat{\beta}_0^{MLE} + \hat{\beta}_1^{MLE} X_{j1} + ... + \hat{\beta}_p^{MLE} X_{jp}), \tag{1.11}$$

où les $\hat{\beta}_k^{MLE}$ sont les coefficients de régression estimés par ML. Chaque assuré a donc sa propre probabilité de rachat, dépendante de ses caractéristiques personnelles et contractuelles. Il est capital d'avoir une idée de la précision de cette estimation. Pour ce faire, le calcul d'un intervalle de confiance est indispensable tant sur le plan individuel que sur le plan collectif (au niveau du portefeuille) lorsque nous voulons agréger les résultats pour reconstruire le taux de rachat du portefeuille.

Plaçons nous dans le cadre collectif (à l'échelle du portefeuille). L'approximation normale de la loi binomiale pourrait constituer le point de départ de la construction de cet intervalle de confiance. Cependant, cette approximation requiert deux hypothèses sous-jacentes : i) $n \to \infty$ (grand nombre d'individus), ii) la probabilité individuelle p_j de rachat est comparable pour l'ensemble des individus (hypothèse d'homogénéité).
Le premier point n'est généralement pas un problème dans l'industrie de l'assurance (les portefeuilles sont souvent volumineux par nature). Le deuxième point est une condition nécessaire pour l'application du théorème de la limite centrale (TCL) : la somme de variables aléatoires i.i.d. suit une loi gaussienne. Le portefeuille est en réalité très hétérogène, mais on pourrait former i groupes d'assurés homogènes (en termes de caractéristiques), chacun de taille n_i. De plus, les assurés à l'intérieur de chaque groupe i sont considérés independants. Le nombre de rachats N_i^s du groupe i composé de n_i assurés est donc binomialement distribué (par propriété) ou normalement distribué (TCL, somme de Bernoulli i.i.d.), bien que l'hypothèse d'indépendance puisse paraître relativement discutable car l'environnement extérieur peut affecter un ensemble d'agents simultanément. Ainsi pour chaque groupe i,

$$\mathbb{E}[N_i^s] = \sum_{j=1}^{n_i} p_j = n_i p_j \ , \tag{1.12}$$

$$\mathbb{V}ar[N_i^s] = \sum_{i=1}^{n_i} p_j(1 - p_j) = \sum_{i=1}^{n_i} p_j q_j = n_i p_j q_j. \tag{1.13}$$

A partir de (1.12) et (1.13), nous obtenons l'intervalle de confiance de la probabilité de rachat $\hat{p_i} = N_i^s/n_i$ du i^e groupe homogène en utilisant celui de la distribution gaussienne. Le nombre total de rachats N^s du portefeuille est la somme des rachats de chaque sous-groupe homogène : $N^s = \sum_i N_i^s$. Or la loi Normale est stable par somme, donc N^s est toujours normalement distribué. Nous avons donc finalement une bonne approximation de la probabilité de rachat du portefeuille $\hat{p} = N^s/n$ par

$$\hat{p} = \frac{1}{n}\sum_i N_i^s \sim N\Big(\frac{1}{n}\sum_i n_i\, p_j, \frac{1}{n^2}\sum_i n_i\, p_j(1-p_j)\Big),$$

qui conduit logiquement à l'intervalle de confiance (au niveau 5%)

$$[A - 1.96\, B\ ,\ A + 1.96\, B] \tag{1.14}$$

où $A = \dfrac{\sum_i n_i\, p_j}{n}$, $B = \sqrt{\dfrac{\sum_i n_i\, p_j(1-p_j)}{n^2}}$, i est l'indice des sous-groupes homogènes, et p_j est la probabilité de rachat correspondante estimée.
Pour des soucis de concision, nous ne présentons pas ici les tests statistiques conduisant à la validation du modèle. Le test du ratio de vraisemblance (pour la validation du modèle) et de Wald (pour la pertinence des covariables) sont détaillés en annexe B.2.5.

1.2.3 Interprétations des résultats

Les valeurs estimées des coefficients de régression nous renseignent sur l'impact de chaque facteur de risque. L'ordonnée à l'origine β_0 correspond à la valeur de z pour le profil de risque de référence : c'est la moyenne de la réponse lorsque les covariables du prédicteur valent les modalités de référence pour les variables catégorielles, et sont nulles pour les variables continues (à condition d'avoir centré ces covariables continues en amont, sinon lorsqu'elles valent leur moyenne). Les coefficients β_k (k = 1,2,...,p) décrivent la contribution de chaque risque : un β_k positif signifie que si le facteur de risque augmente alors la probabilité de rachat augmente (corrélation positive), alors que s'il est négatif l'évolution se fait en sens inverse. Si la valeur absolue de $\beta_k/\sigma(\beta_k)$ (où $\sigma(\beta_k)$ est l'écart-type de l'estimation du coefficient) est grande, alors le facteur de risque k a une forte influence sur la probabilité de rachat, et inversement. Ces coefficients sont à comparer au profil de risque de référence, pour lequel $\beta = 0$ (sauf pour β_0).

Les praticiens aiment bien utiliser le rapport de côte (OR pour "odd-ratio"), car il exprime le rapport entre les chances de racheter ou non (p/(1-p)). Prenons un exemple d'illustration : la probabilité de rachat $P(Y{=}1|X)$ vaut $p = 0,7$. L'OR vaut donc p/q=0.7/0.3=2.33, ce qui veut dire qu'avec les mêmes caractéristiques X, le rachat a 2,33 fois plus de chance de se produire que le non-rachat. Cette idée se généralise lorsque les professionnels veulent évaluer la différence en termes de probabilité de rachat avec un changement des caractéristiques entre deux individus. Prenons comme exemple l'âge : grâce à l'équation (1.11), nous savons que $p/q = e^{\beta_0+\beta_1 X_{age}}$ est l'OR pour un assuré donné. Lors de la comparaison de deux individus ne différant que par leur âge (40 et 30 ans), tous les termes disparaissent excepté l'âge, ce qui donne un OR entre les deux individus de

$$\frac{P(Y=1|X_{age}=40)}{P(Y=0|X_{age}=40)} \Big/ \frac{P(Y=1|X_{age}=30)}{P(Y=0|X_{age}=30)} = \frac{e^{40\beta_1}}{e^{30\beta_1}} = e^{10\beta_1}$$

Nous constatons que la variation des valeurs de variables explicatives entraîne un effet multiplicatif du risque, par des constantes liées aux coefficients de régression. Ces OR sont un outil opérationnel très utile (car simple) pour la définition de classe de risque.

1.2.4 Limites de la modélisation LR et améliorations potentielles

Les hypothèses formulées pour la mise en place de la modélisation constituent les principales limites du modèle. En particulier, les assurés sont considérés comme conditionnellement indépendants deux à deux (formellement $Y_j|X_j \perp\!\!\!\perp Y_m|X_m$, $\forall m \neq j$) sachant leurs caractéristiques. Nous avons déjà évoqué le problème posé par cette hypothèse, qui n'est bien sûr pas totalement vérifiée en réalité. D'autre part l'estimation des paramètres du modèle est réalisable à condition que le coefficient de corrélation de Pearson ne vaille pas 100% dans l'étude des corrélations entre covariables, auquel cas la matrice des covariables (ou de "design") est singulière. La LR nécessite également un important volume de données afin d'en assurer la robustesse, mais cela ne semble pas être un écueil majeur en assurance. L'utilisateur doit cependant s'assurer de la pertinence de ses données d'origine, point qui peut s'illustrer par l'exemple suivant : considérons un très ancien (disons 50 ans) portefeuille en *run-off* (pas de nouvelles affaires). Presque tous les assurés auront racheté leur contrat, la régression n'aurait plus beaucoup de sens en termes de segmentation (ce n'est pas le cas dans nos applications) et les résultats seraient probablement moins utiles. Enfin un seuil est à définir dans une problématique de prévision en termes de classification : une amélioration possible réside dans le choix optimal du seuil d'attribution de la réponse binaire. Le seuil naturel de 0,5 que nous considérons signifie qu'une probabilité prédite plus grande que 0,5 se verra attribuée la réponse 1 (rachat), sinon 0. Ruiz-Gazen and Villa (2007), Liu et al. (2006) et Lemmens and Croux (2006) montrent dans leurs articles que ce choix n'est pas optimal en cas d'échantillon avec réponses largement deséquilibrées. Pour éviter d'obtenir des résultats qui ne seraient pas représentatifs de la réalité, leurs méthodes se basent sur des techniques de reéchantillonage de type *importance sampling*.

1.3 Illustration : application sur des contrats mixtes

Les informations dont nous disposons dépendent de l'entité pays d'AXA qui nous les fournit. La majorité des bases de données comprennent la date de naissance des assurés, leur sexe, leur lieu de résidence, la date d'émission du contrat, la date de fin du contrat (et le motif), le type de contrat, la fréquence de la prime, la somme assurée. Nous avons aussi dans certains cas des renseignements concernant le statut marital, le statut fumeur ou encore le réseau de distribution (cette liste ne se veut pas exhaustive). Dans cette partie, nous nous focalisons sur le portefeuille espagnol d'AXA et plus précisément les contrats mixtes. **Précisons d'ailleurs que l'ensemble des études de cas des chapitres 1, 2 et 3 sont réalisées avec ces mêmes contrats mixtes pour garder une certaine continuité dans le raisonnement**.

Nous essayons de segmenter notre population d'assurés en fonction de leur type de contrat, de leur sexe, de leur âge, de leur fréquence de prime, de leur somme assurée, et du montant de leur prime. Plus précisemment, la somme assurée (notée "face amount" dans la base de données) est un indicateur de la richesse de l'assuré, la prime contient la prime de risque (liée à la garantie du risque Vie considéré, ici le décès) et la prime d'épargne. La prime de risque est le produit actualisé de la somme sous-risque par la probabilité de déclencher la garantie. La prime d'épargne est l'investissement de l'assuré. Nous utilisons en partie la librairie `rpart` de `R` pour implémenter la méthode CART et obtenir nos résultats. Les fonctions servant à résoudre le problème d'optimisation dans l'estimation du modèle LR sont déjà intégrées au coeur des programmes `R`.

Analyse statique

Nous entendons par analyse statique une photographie de l'état du portefeuille en décembre 2007, pour les contrats mixtes de la base de données. Les résultats numériques ci-après concernent l'étude des 28 506 contrats mixtes, isolés des contrats pure épargne car les comportements de rachat doivent être distingués par grande ligne de produit. Ces contrats annuellement renouvelables (TAR) ne peuvent pas être rachetés avant un an d'ancienneté (sauf cas exceptionnel), et il est à noter qu'il n'y a pas de dispositif fiscal particulier en Espagne sur les contrats d'Assurance Vie avec une composante épargne. Cela signifie ici que les assurés peuvent racheter leur contrat à chaque date anniversaire sans frais, mais sont pénalisés en cas de rachat à un autre moment. Nous voyons en figure 1.1 que ceci est une incitation importante qui dicte le profil des rachats en fonction de l'ancienneté du contrat.

L'étude couvre la période 2000-2007, les caractéristiques des assurés et de leur contrat sont celles observées soit à la date de rachat, soit en Décembre 2007 (si pas de rachat). Rappelons que notre but est de trouver les principaux déclencheurs de rachat en se servant des variables explicatives observées, ce qui nous permettra de détecter des agents "risqués" en termes de décision de rachat **à une date donnée**. Certaines précautions sont à prendre dans ce type d'analyse faute de quoi les résultats peuvent être particulièrement biaisés. Nous pensons à la composition du portefeuille, à son degré de maturité, à la part des nouvelles affaires. Par exemple, si les rachats ne sont observés qu'après une certaine ancienneté, il est important de s'assurer que le portefeuille soit à maturité (sinon nous observerions un taux de rachat quasi-nul, ce qui nous amènerait à des conclusions erronées). Nous verrons que c'est l'un des gros défauts de l'analyse statique qui peut introduire un biais important, par opposition à l'analyse dynamique du chapitre 2 qui tente de corriger ceci.

En Décembre 2007, 15 571 des 28 506 contrats mixtes ont été rachetés, soit environ 55%. Les deux modèles de segmentation présentés nous apportent deux informations complémentaires :

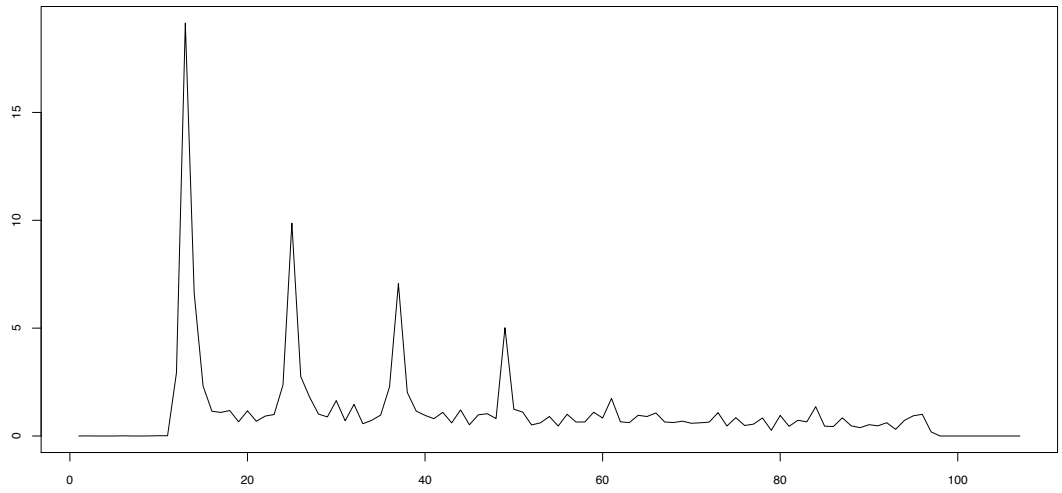


FIGURE 1.1 – Taux de rachat (%) VS ancienneté (en mois) pour les contrats Mixtes. Effet des pénalités de rachat (les contrats peuvent être rachetés sans frais à chaque date anniversaire, ce qui explique les pics de rachat observés).

TABLE 1.1 – Matrice de confusion (T_{max}), échantillon de validation.

	observed Y = 0	observed Y = 1
predicted Y = 0	4262	1004
predicted Y = 1	728	5644

TABLE 1.2 – Matrice de confusion (arbre élagué), échantillon de validation.

	observed Y = 0	observed Y = 1
predicted Y = 0	4188	1078
predicted Y = 1	664	5708

- CART nous donne les variables les plus discriminantes par rapport aux comportements de rachat en ordre décroissant (en lisant l'arbre depuis la racine jusqu'aux feuilles). Au final, nous classons un assuré comme "risqué" à l'aide d'une prévision binaire (bien qu'on puisse accéder aux probabilités précises de chaque classe et donc à la probabilité de rachat) ;
- LR offre un résultat plus numérique : la probabilité de racheter son contrat étant donné ses caractéristiques et la sensibilité de sa décision aux évolutions des covariables (avec les OR).

1.3.1 Résultats par les CART

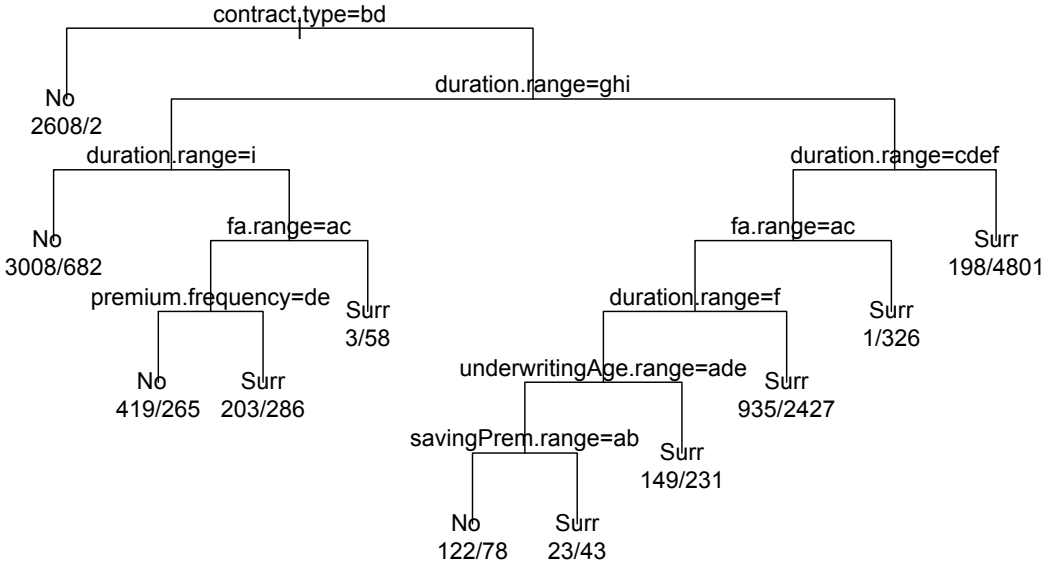
Nous réalisons l'analyse sous `R` grâce à la librairie `rpart` [4], et plus précisément par la procédure `rpart` qui construit l'arbre de classification. Par défaut `rpart` utilise l'indice de Gini pour calculer l'impureté d'un noeud, mais cette option n'est pas très importante puisque les résultats ne sont quasiment pas impactés. Il n'y a pas de coût de mauvaise classification introduit (voir annexe B.1.3) dans notre application.
Nous procédons comme en théorie et construisons d'abord l'arbre T_{max} sans coût de complexité (en posant l'option *cp* égale à 0), puis l'arbre est élagué ("pruned tree") pour diminuer son nombre de feuilles et simplifier les résultats. Le nombre minimal d'observations dans une feuille a été fixé à 1, le nombre de divisions concurrentes calculées est de 2. Nous créons aléatoirement les échantillons d'apprentissage et de validation, dont les tailles respectives sont de 16 868 et 11 638 assurés.

L'estimation par échantillon témoin de l'erreur de prévision de l'arbre maximal T_{max} calculée sur l'échantillon de validation est de 14.88%, correspondant aux termes non-diagonaux de la matrice de confusion du tableau 1.1. Cet arbre a trop de feuilles et admet une représentation trop complexe, ce à quoi nous remédions en l'élaguant. Le choix du paramètre de complexité α lors de l'élagage (annexe B.1.3) est un arbitrage entre la taille finale de l'arbre et le taux de mauvaise classification désiré par l'utilisateur. Le tableau B.1 et la figure B.3 en annexe B.1.2 trace l'erreur d'apprentissage en fonction du coût de complexité. Dans ce graphe, à chaque paramètre de complexité est associé un arbre optimal dont la taille est donnée, ce qui permet de choisir le α optimal par minimisation de l'erreur d'apprentissage. Nous obtenons $\alpha \in]1.04e^{-04}, 1.30e^{-04}]$, mais le nombre de feuilles correspondant (82) est encore trop élevé à notre goût. Nous avons donc choisi $\alpha = 6e^{-04}$, ce qui correspond à 11 feuilles pour une très faible perte de précision dans la classification. Cet arbre est visualisable en figure 1.2. La variable la plus discriminante semble toujours être le type de contrat (caractérisé en fait par le type de prime, unique ou périodique ; et l'option de participation au bénéfice), puis l'ancienneté du contrat et ainsi de suite.

Les variables sélectionnées dans la construction de l'arbre sont le type de contrat, l'ancienneté, la richesse de l'assuré ("face amount"), la fréquence de prime, la prime d'épargne et

4. r-partitionning : http ://cran.r-project.org/web/packages/rpart/index.html

FIGURE 1.2 – L'arbre final de classification. Variable réponse binaire : rachat. La première règle de division *contract.type = bd* signifie que le type de contrat est la variable la plus discriminante (*bd* correspond aux 2^e et 4^e modalités, comme dans l'ordre alphabétique). Les variables continues ont été catégorisées pour la modélisation.

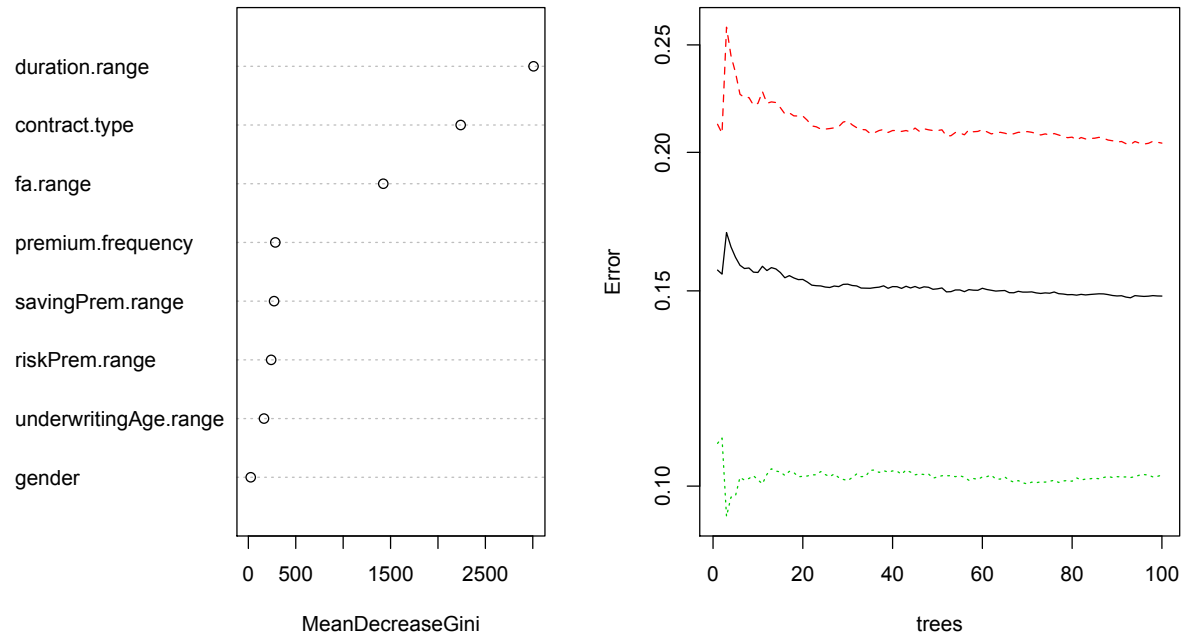


l'âge de souscription. Nous remarquons que le sexe et la prime de risque n'apparaissent pas dans cet arbre final, parce que leurs effets ne semblent pas être significatifs. La première règle de division est "L'assuré possède t'il l'option de participation au bénéfice ?". Si la réponse est négative alors descendre dans la branche de gauche, sinon descendre dans la branche de droite. Les classes prédites (rachat ou non-rachat) sont écrites sur les feuilles, les proportions qui y apparaissent sont le nombre d'assurés n'ayant pas racheté versus ceux qui ont racheté leur contrat. Plus la différence entre ces deux nombres est grande, meilleure est la segmentation. Ici, un assuré dont le contrat ne contient pas l'option de participation au bénéfice a 99,92% (2608/2610) de ne pas racheter, quel que soit le format de sa prime (PP sin PB et PU sin PB, voir légende du tableau 1.4). La classe attribuée est donc "No", équivalente à l'absence de rachat. Considérons un assuré dont les caractéristiques sont une prime périodique, un contrat avec clause de participation au bénéfice. Son ancienneté appartient aujourd'hui à la septième modalité de la variable catégorisée, et sa richesse se situe dans la deuxième classe. Selon l'arbre construit, cet assuré aurait 95% (58/61) de chance de racheter, et serait donc considéré comme potentiellement très risqué.

Il est évident que le facteur de risque le plus discriminant lorsque nous regardons la figure 1.2 est l'option de participation au bénéfice. Le taux de mauvaise classification (erreur d'apprentissage) de cet arbre est de 15% (33.1% $\times$ 45.4%, où 45.4% est l'erreur de la racine quand aucune division n'est réalisée) d'après le tableau B.1 des erreurs relatives, disponible en annexe B.1.2. L'erreur de prévision de 14.97% peut être estimée via la matrice de confusion du tableau 1.2, elle est relativement satisfaisante puisqu'elle reste proche de l'erreur de prévision de l'arbre maximal T_{max}. Le compromis est ici très intéressant : l'élagage d'un arbre de 175 feuilles à un arbre de 11 feuilles augmente l'erreur de prévision de moins de 1% !

Afin de consolider ces résultats, nous utilisons les *bagging predictors* dont l'implémentation est

FIGURE 1.3 – Sur la gauche : l'importance des variables explicatives. Sur la droite : le nombre d'arbres requis pour stabiliser l'erreur *out-of-bag* : la courbe noire est l'erreur globale, la verte est l'erreur pour la réponse "rachat" et la rouge l'erreur pour la réponse "pas de rachat".



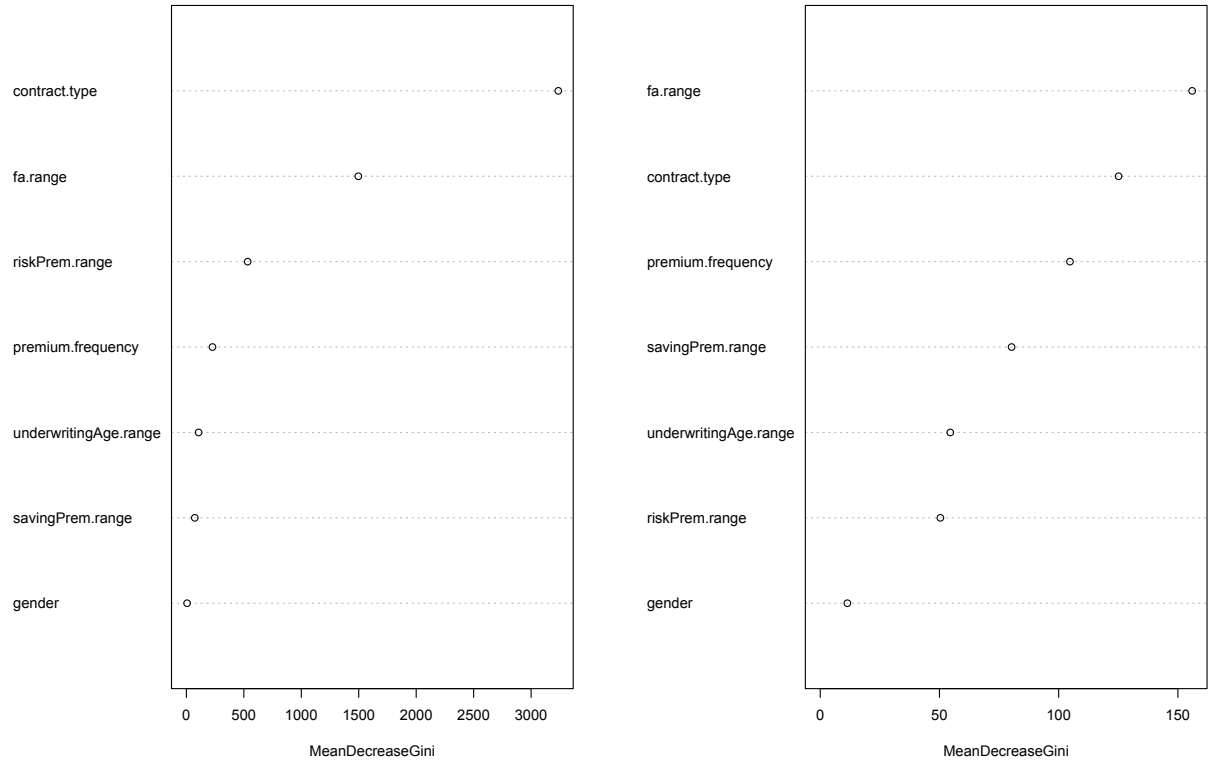
accessible via la librairie `randomForest`. Lors de l'utilisation de l'algorithme des forêts aléatoires, les étapes successives permettent de construire un ensemble de classifieurs bootstrap. L'agrégation de ces classifieurs amène à un classifieur final, qui ne peut cependant pas être représenté sous forme d'arbre mais qui fournit des résultats plus robustes (tous les concepts utilisés dans cet algorithme sont consultables sur la page web de Breiman [5]). Le tableau 1.3 résume les résultats de classification sur l'échantillon d'origine (pas d'échantillon d'apprentissage ni de validation car il s'agit déjà de méthodes bootstrap) : l'estimation de l'erreur sans biais appelée erreur *out-of-bag* est de 14.73%. L'importance des variables explicatives dans le processus de classification est visualisable en figure 1.3, de même que le nombre d'arbres nécessaires dans la forêt pour la stabilisation de l'erreur *out-of-bag* (environ 50 arbres ici). Ces résultats viennent confirmer nos attentes : l'ancienneté du contrat et son type sont encore une fois les variables les plus discriminantes pour expliquer les décisions de rachat des assurés. Afin de s'assurer que l'effet de l'ancienneté du contrat (effet temporel) ne biaise pas ces résultats, nous avons décidé de relancer l'analyse en isolant cet effet par une séparation dans notre jeu

5. See http ://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm

	observés Y = 0	observés Y = 1
prédits Y = 0	10327	2608
prédits Y = 1	1592	13979

TABLE 1.3 – Matrice de confusion du classifieur obtenu par les forêts aléatoires.

FIGURE 1.4 – Importance des variables explicatives en excluant l'effet de l'ancienneté. Sur la gauche les assurés dont l'ancienneté du contrat correspond aux pics observés en Figure 1.1, et autres assurés sur la droite.



de données : d'un côté les personnes dont le rachat s'est effectué pour des anciennetés de contrat correspondantes aux pics observés en figure 1.1, de l'autre les assurés restants. Nous regardons ainsi les rachats provoqués par les contraintes de frais à payer, mais aussi ceux qui ne le sont pas. La figure 1.4 montre que les facteurs discriminants principaux restent les mêmes quelle que soit la population étudiée (l'ordre diffère légèrement), ce qui signifie que l'effet de l'ancienneté n'est pas corrélé à un autre facteur de risque et n'introduit pas de biais dans les résultats que nous obtenons.

1.3.2 Classification par le modèle logistique (LR)

Le logiciel `R` et sa fonction `glm` nous permettent d'implémenter le modèle logistique sur nos données. Comme détaillé dans la partie théorique, les sorties du modèle sont l'effet de chaque covariable (facteur) par les coefficients de régression, l'écart-type de l'estimation de ces coefficients, et la déviance du modèle (cf annexes B.2.3, B.2.4 et B.2.5).

Lors de la résolution du système d'équation amenant à l'estimation des coefficients de régression, les variables catégorielles sont introduites par une suite de variables indicatrices (une par modalité) qui permet de définir la matrice de "design" qui sera inversée par la procédure `glm`. Cette fonction utilise un algorithme itératif pas-à-pas dans le but de comparer un modèle basé sur p' des p variables d'origine à n'importe quel sous-modèle (contenant une variable de moins), ou même à n'importe quel sur-modèle (avec une variable supplémentaire). La fonction

TABLE 1.4 – Rapports de côte (OR), contrats mixtes (ancienneté en mois, échantillon d'apprentissage). Types de contrat : PP con PB → prime périodique (PP) avec participation au bénéfice (PB), PP sin PB → PP sans PB, PU con PB → prime unique (PU) avec PB, PU sin PB → PU sans PB. Les variables continues (ex : ancienneté) ont été catégorisées.

OR	Référence	Autres modalités							
Duration	[0,12]	]12,18]	]18,24]	]24,30]	]30,36]	]36,42]	]42,48]	]48,54]	> 54
Rachats	3062	1740	1187	791	728	400	365	244	682
OR empirique		10.56	2.89	2.69	1.82	1.16	0.96	0.68	0.19
OR modélisé		0.27	0.07	0.06	0.05	0.03	0.02	0.02	0.004
Fréquence de prime	Mensuelle	Bi-mensuelle	Trimestrielle	Semestrielle	Annuelle	Unique			
Rachats	2790	12	323	92	595	5387			
OR empirique		2.22	0.93	0.66	2.39	1.60			
OR modélisé		2.52	0.97	0.80	1.55	0.75			
Age souscription	[0,20[	[20,30[	[30,40[	[40,50[	[50,60[	[60,70[	> 70		
Rachats	258	1719	2165	2002	1490	1088	477		
OR empirique		1.16	1.06	1.25	1.63	2.67	3.28		
OR modélisé		1.32	0.99	0.77	0.67	0.51	0.47		
Face amount	#1	#2	#3						
Rachats	5361	684	3154						
OR empirique		0.14	0.12						
OR modélisé		0.003	0.0008						
Prime de risque	#1	#2	#3						
Rachats	3941	2987	2271						
OR empirique		1.50	0.92						
OR modélisé		1.43	1.30						
Prime d'épargne	#1	#2	#3						
Rachats	3331	1762	4106						
OR empirique		1.90	2.09						
OR modélisé		2.55	3.78						
Type de contrat	PP con PB	PP sin PB	PU con PB	PU sin PB					
Rachats	3840	0	5357	2					
OR empirique		0	4.75	0.0008					
OR modélisé		5.6e-08	0.0006	3.9e-06					

`stepAIC` de la librairie `MASS` nous permet de sélectionner de manière adéquate les variables pertinentes, pour finalement obtenir un modèle optimal qui contient un minimum de variables explicatives pertinentes. A des fins de comparaison, les échantillons d'apprentissage et de validation sont les mêmes que dans l'application CART. Comme d'habitude les coefficients de régression ont été estimés sur l'échantillon d'apprentissage alors que les prévisions se sont effectuées sur l'échantillon de validation. Le tableau B.2 en annexe B.2.1 récapitule l'ensemble des résultats, à savoir les coefficients de régression et leur écart-type, les p-valeurs des tests de Wald (confiance dans l'estimation et significativité des coefficients, cf annexe B.2.5). Nous déduisons de ce tableau que les covariables qui semblent avoir le plus d'impact (grande valeur absolue) sont encore une fois l'ancienneté du contrat, le type de contrat, mais aussi la richesse de l'assuré ; ce qui est en ligne avec les résultats obtenus par la méthode CART. Le fait que l'ancienneté de contrat soit un facteur explicatif clef rend très important le profil des rachats en fonction de celle-ci, et notamment la prise en compte de changement de législation par exemple (taxation, fiscalité...).

Les rapports de côte (OR) (cf section 1.2.3) sont à comparer à la valeur 1 (modalité de référence). Nous constatons au regard du tableau 1.4 que les OR modélisés représentent assez mal la réalité car de grosses différences existent avec les OR empiriques (obtenus par statis-

TABLE 1.5 – Matrice de confusion (LR).

	observé Y = 0	observé Y = 1
prédit Y = 0	#correct rejections 4153	#misses 637
prédit Y = 1	#false risky policyholder 1113	#success 5735

TABLE 1.6 – Critères de performance.

	T_{max}	T_{pruned}	$T_{RandomForest}$	LR
Se	84.9%	84.1%	84.3%	90%
Sp	85.4%	86.3%	86.7%	78.9%
(1-Se)	15.1%	15.9%	15.7%	10%

tiques descriptives). Par exemple, le modèle prévoit qu'une personne âgée de 70 ans ait moins de chance de racheter qu'un jeune assuré de moins de 20 ans, toutes caractéristiques égales par ailleurs. L'expérience montre qu'ils sont en fait 3,28 fois plus susceptibles de racheter ! Par contre, les OR estimés varient très souvent dans la même direction que les OR observés, au sens qu'ils ont la même tendance. C'est le cas si l'on considère le facteur d'ancienneté : la figure 1.1 expose le profil des rachats en fonction de l'ancienneté (pourcentage des rachats pour chaque tranche d'ancienneté) et confirme les estimations des OR (tableau 1.4) liés à ce facteur : effectivement le risque est important à partir de la première date anniversaire et décroît dans le temps.

Le modèle a globalement une mauvaise qualité d'ajustement au regard de la significativité des estimations des coefficients de régression, c'est d'ailleurs ce qui explique que les OR empiriques et modélisés soient si éloignés. Toutefois, il faut garder en tête que notre problématique initiale est de segmenter notre population d'assuré et de faire des prévisions de profil de risque. Nous préférons donc privilégier le pouvoir prédictif à la qualité d'ajustement dans l'arbitrage naturel entre ces deux notions. La matrice de confusion du tableau 1.5 fournit le nombre d'assurés mal classés et représente le pouvoir prédictif de cette méthode, la lecture de cette table se faisant de la même manière qu'avec les CART. Une hypothèse supplémentaire est utilisée ici, à savoir que nous attribuons une réponse de rachat lorsque la probabilité de rachat modélisée excède 0,5 et inversement. Les bonnes prévisions représentent 84.96% de l'échantillon de validation ; ce qui donne une erreur de prévision de 15.04%, un résultat quasi-similaire à celui obtenu par l'algorithme CART (14,97% pour mémoire).

D'autres critères, couramment appelés critères de performance, servent à comparer les classifieurs : il s'agit de la sensibilité (Se) et de la spécificité (Sp). Appelons *success* la case correspondante à une réponse observée et prédite de rachat dans la matrice de confusion. Les *misses* correspondent à une réponse prédite de non-rachat tandis que l'observation était un rachat. Les *correct rejections* correspondent à une réponse observée et prédite de non-rachat, enfin les *false risky policyholder* désignent une réponse prédite de rachat alors qu'il n'y en pas eu dans la réalité. La sensibilité est définie comme le nombre de *success* sur le nombre de contrats rachetés observés, et la spécificité est le nombre de *correct rejections* sur le nombre de contrats non-rachetés observés. Le tableau 1.6 résume les critères de performance des différentes méthodes de classification ; sachant que ce qui nous intéresse avant tout dans ce contexte est de minimiser les *misses*. Les prévisions par la LR présentent moins de *misses* et plus de *false risky policyholders*, les résultats étant comparables et les erreurs équilibrées entre les trois applications de CART. Le compromis entre sensibilité et spécificité est meilleur avec CART mais le nombre de *misses* est plus élevé, ce qui nous conduirait ici à choisir le modèle LR ici (10%) pour plus de prudence.

1.4 Conclusion

L'objectif de ce chapitre était de présenter deux modèles de segmentation qui apportent des réponses sur le profil de risque des assurés, par la prise en compte de leurs caractéristiques individuelles et des options de leurs contrats. Qu'avons-nous appris ?

Cette étude a permis de mettre en exergue quelques types de profils risqués : **les jeunes ont tendance à racheter davantage que les autres**, comme ceux qui ont une **prime périodique** ("annuelle" et "bi-mensuelle" sont les pires cas). Le cycle de vie de l'assuré joue donc un rôle central dans les comportements de rachat, à savoir que la population des jeunes adultes est bien souvent en phase d'investissement et nécessite de disposer de fonds (acquisition d'une voiture, d'une maison, ...). Les assurés les plus pauvres (au vu de l'indicateur "face amount", ce qui ne veut pas forcément dire qu'ils le sont) rachèteront probablement leur contrat plus souvent : en effet ils doivent payer des frais et des primes régulières mais n'ont pas l'argent pour, alors que les personnes plus riches n'y prêtent pas vraiment attention (la corrélation entre l'âge et la richesse des assurés est négligeable dans notre portefeuille). La majeure concentration du risque se situe grosso modo sur les premiers jours (premières semaines) qui suivent la levée d'une contrainte fiscale ou d'une pénalité prévue par le contrat : **lorsque l'ancienneté atteint ce seuil, le risque est très élevé**. Dans une optique de segmentation de risque à la souscription, notons que ce facteur de risque est une information inexistante qui ne peut donc pas être prise, ce qui justifie le fait que nous ayons regardé le classement par importance des facteurs de risque en isolant cet effet. Enfin, la clause de participation aux bénéfices (PB) de l'entreprise semble jouer un rôle clef dans le processus de décision du rachat, l'étude ayant montré que les personnes **sans cette option ne rachètent que très peu** leur contrat alors que les autres le font tôt ou tard. Trois principales raisons pourraient expliquer ce phénomène : premièrement, les agents rachètent pour basculer leur épargne sur un nouveau produit offrant un taux de PB supérieur au leur ; deuxièmement un taux de PB attractif pendant les premières années du contrat permet à l'assuré de surperformer le rendement initial et l'inciter à racheter par la suite, dans le but de récupérer une valeur de rachat intéressante ; troisièmement, le simple fait de recevoir annuellement l'information sur le taux de PB qui sera versé par le contrat et la valeur de rachat associée peut jouer sur une décision de rachat. Le **sexe de l'assuré n'apparaît pas comme un facteur de risque à prendre en compte** car les différences entre hommes et femmes pour les décisions de rachat sont minimes.

D'un point de vue plus technique, nous avons vu que le processus de classification peut se réaliser soit par l'emploi du modèle logistique soit par l'emploi des méthodes CART, les résultats étant en adéquation. Des profils-types de risque se dégagent plus facilement à partir de statistiques descriptives ou des CART, tandis que le modèle LR donnent accès à des indicateurs supplémentaires tels que les rapports de côte. Les deux modèles apportent des résultats complémentaires et font intervenir des hypothèses bien différentes, mais servent globalement une même cause : une réduction de dimension de l'espace des données, autrement dit une sélection des variables les plus discriminantes en termes de rachat (dans le but de simplifier une future modélisation). Un outil informatique basé sur RExcel permettant d'obtenir un large panel de statistiques descriptives ainsi que l'usage de ces deux modèles de segmentation a été implémenté pour étudier les comportements de rachat à plusieurs niveaux d'échelle (numéro de produit, ligne de produit, famille de produit, pays) dans quatre entités d'AXA (Espagne, Etats-Unis, Belgique, Suisse). Quelques illustrations du fonctionnement de cet outil sont accessibles en annexe E.

Enfin cette analyse statique est utile pour comprendre quelles sont les caractéristiques des contrats et des assurés qui ont un rôle dans les décisions de rachat, mais elle présente l'inconvénient majeur de ne pas tenir compte de l'impact du contexte économique et financier sur les comportements de rachat. En effet, nous regardons l'état du portefeuille à une date fixée (fin 2007), en supprimant de ce fait les effets temporels. Nous pourrions argumenter que dans un contexte économique classique les comportements de rachat ne sont pas guidés par celui-ci, et donc que cette analyse suffit. Cependant, lorsque l'environnement (économie, image de la compagnie) change, il devient très difficile d'anticiper les comportements de rachat pour reconstruire le taux de rachat à l'échelle du portefeuille. La corrélation entre les décisions des agents est par exemple susceptible de fortement augmenter, un point que nous abordons dans le chapitre suivant. La modélisation devient de ce fait beaucoup plus compliquée, et nous verrons la nécessité de bien capturer les effets endogènes (facteurs idiosynchratiques et/ou structurels) **et** exogènes (conjoncturels) à travers les problèmes rencontrés ci-après lors de l'utilisation "dynamique" des GLM.

Bibliographie

Austin, P. C. (2007), 'A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting ami mortality', *Statistics in Medicine* **26**, 2937–2957. 29

Balakrishnan, N. (1991), *Handbook of the Logistic Distribution*, Marcel Dekker, Inc. 29

Breiman, L. (1994), Bagging predictors, Technical Report 421, Department of Statistics, University of California. 29

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* (24), 123–140. 29

Breiman, L. (1998), 'Arcing classifiers', *The Annals of Statistics* **26**(3), 801–849. 29

Breiman, L. (2001), 'Random forests', *Machine Learning* (45), 5–32. 29

Breiman, L., Friedman, J., Olshen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, Chapman and Hall. 24, 26, 28, 29

Cox, D. (1972), 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society : Series B* (34), 187–220. 29

Cox, S. H. and Lin, Y. (2006), Annuity lapse rate modeling : tobit or not tobit ?, *in* 'Society of actuaries'. 29

Ghattas, B. (1999), 'Previsions par arbres de classification', *Mathematiques et Sciences Humaines* **146**, 31–49. 27

Ghattas, B. (2000*a*), 'Aggregation d'arbres de classification', *Revue de statistique appliquee* **2**(48), 85–98. 28

Ghattas, B. (2000*b*), Importance des variables dans les methodes cart. GREQAM. 28

Hilbe, J. M. (2009), *Logistic regression models*, Chapman and Hall. 24

Hosmer, D. W. and Lemeshow, S. (2000), *Applied Logistic Regression, 2nd ed.*, Wiley. 29

Kagraoka, Y. (2005), Modeling insurance surrenders by the negative binomial model. Working Paper 2005. 29

Lemmens, A. and Croux, C. (2006), 'Bagging and boosting classification trees to predict churn', *Journal of Marketing Research* **134**(1), 141–156. 32

Liu, Y., Chawla, N., Harper, M., Shriberg, E. and Stolcke, A. (2006), 'A study in machine learning for unbalanced data for sentence boundary detection in speech.', *Computer Speech and Language* **20**(4), 468–494. 32

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models, 2nd ed.*, Chapman and Hall. 29

Milhaud, X., Maume-Deschamps, V. and Loisel, S. (2011), 'Surrender triggers in life insurance : what main features affect the surrender behavior in a classical economic context ?', *Bulletin Francais d'Actuariat* **11**(22), 5–48. 23

Ruiz-Gazen, A. and Villa, N. (2007), 'Storms prediction : logistic regression vs random forest for unbalanced data', *Case Studies in Business, Industry and Government Statistics* **1**(2), 91–101. 32

Chapitre 2

Crises de corrélation des comportements

Ce chapitre est inspiré de l'article Loisel and Milhaud (2011), coécrit avec Stéphane Loisel et publié dans l'*European Journal of Operational Research 214, p 348-357.* Cet article a reçu le 2ème prix du *Lloyd's Science of Risk Prize*, dans la catégorie *Insurance operations and markets including financial mathematics.*

Pouvoir recomposer précisément par date le taux de rachat à l'échelle d'un portefeuille d'assurance nécessite de conserver les caractéristiques individuelles des contrats et des assurés, car nous avons vu que certaines variables étaient fortement discriminantes. Par conséquent l'évolution de la composition du portefeuille est importante pour une modélisation adéquate du taux de rachat : en effet si le profil du portefeuille d'assurance change beaucoup entre deux dates données, les décisions de rachat et donc le taux seront radicalement différents. Cette suggestion nous amène à retenir une fois de plus l'usage de modèles de régression qui permettent d'intégrer ces considérations par l'intermédiaire de covariables. Nous verrons également dans ce chapitre qu'une modélisation logistique classique est insuffisante, ce qui nous permettra d'introduire le phénomène de corrélation entre comportements et de présenter qualitativement et quantitativement son impact potentiel.

2.1 Problème de la régression logistique dynamique

Comme vu dans le précédent chapitre, faire des prévisions de taux de rachat en se basant sur une analyse statique peut entraîner des erreurs importantes. Les modèles de segmentation développés sont adaptés dans une optique de définition de classe (profil) de risque, mais il faut relativiser la pertinence de leur utilisation lors de l'étude d'un phénomène dont la modélisation dépend fortement d'un environnement mouvant. Dans ce cas, une analyse dynamique permet de prendre en compte certains facteurs "externes", et ainsi de mieux refléter leur influence sur les décisions des assurés. Nous modélisons dans la suite le taux de rachat du portefeuille sur un pas de temps mensuel par agrégation de décisions individuelles des assurés (les intervalles de confiance sont calculés avec le raisonnement détaillé en 1.2.2). Ces décisions sont retournées par le modèle logistique, auquel nous ajoutons en variables explicatives des facteurs de risque économiques et financiers.

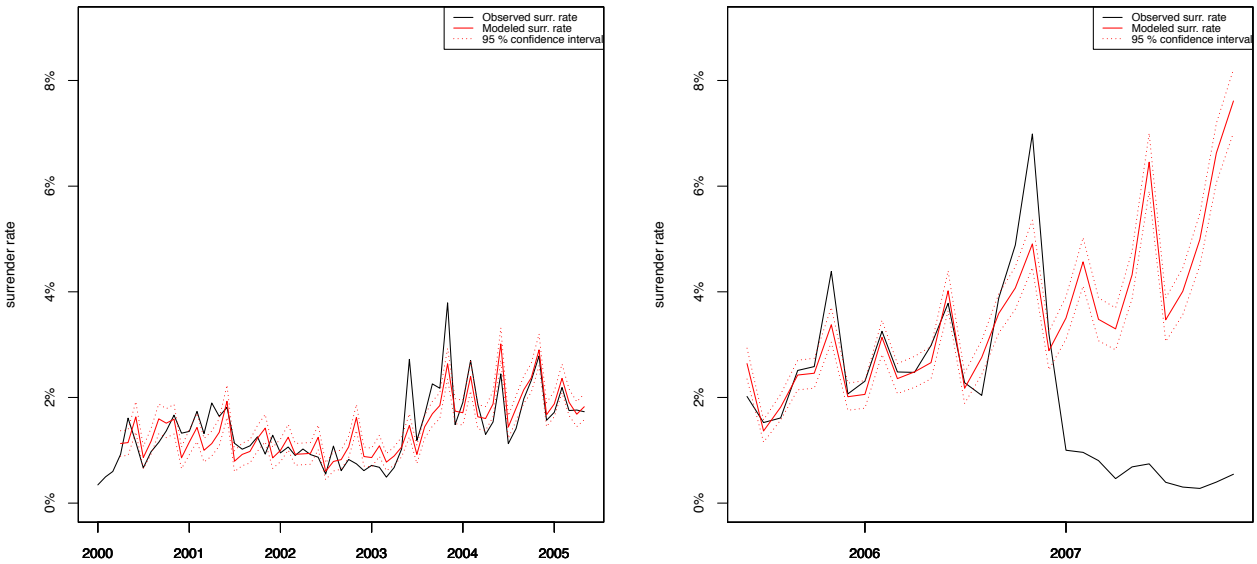
Hormis la prise en compte du contexte économique, cette analyse dynamique permet d'évi-

ter certains problèmes évoqués au chapitre 1 comme la durée de la période couverte pour étudier le phénomène. Si elle est trop courte, on n'observerait aucun rachat (puisqu'il faut généralement un minimum d'ancienneté avant que le rachat puisse se faire), si elle est trop longue on observerait un taux de rachat proche de 100% ; les deux situations n'étant pas réalistes. De plus, le fait de modéliser mensuellement les décisions des assurés permet de rendre compte du fait que dans la réalité les assurés sont susceptibles de se poser fréquemment la question du rachat de leur contrat (en tout cas plusieurs fois dans l'année).
Grossièrement, cette modélisation dynamique pose deux problèmes majeurs : la stabilité et la robustesse. Ces écueils sont dûs à l'ajout d'une hypothèse très forte, l'indépendance temporelle entre les décisions. Nous considérons que la décision d'un assuré à la date $t+1$ est indépendante de ce qui s'est passé avant, et notamment indépendante de sa décision à la date t. Pour construire la base de données nécessaire à cette analyse, nous dupliquons chaque assuré chaque mois où il est présent en portefeuille (ce qui nous donne un échantillon global de 991 010 lignes) et mettons à jour ses caractéristiques (indices économiques, ancienneté,...), à partir de la même base de données que celle utilisée au chapitre 1. Cette opération peut donner lieu à l'introduction d'un nouveau biais : les caractéristiques des assurés qui restent le plus longtemps en portefeuille sont sur-représentées. Ceci dit et après vérification, ce biais ne joue pas beaucoup sur nos résultats lorsque nous les comparons aux coefficients de régression d'un modèle de Cox, dans lequel les assurés et leurs caractéristiques ne sont pas dupliqués (car c'est un modèle de survie).

Après s'être assuré que nous lançons cette analyse sur une période représentative du portefeuille (à maturité), il est possible de voir la qualité de la modélisation en comparant le taux de rachat observé avec le taux de rachat prédit sur un pas mensuel. Les échantillons d'apprentissage et de validation sont construits différemment ici : l'apprentissage représente environ deux tiers de la période étudiée, soit de Janvier 2000 à Mars 2005 (629 357 observations) ; tandis que la validation s'effectue sur la période restante (Avril 2005 à Décembre 2007, 361 653 observations). Cette technique de validation "temporelle" permet de rendre compte de l'exposition des assurés à des contextes économiques différents, et ainsi de tester non pas uniquement la qualité d'adéquation du modèle mais aussi son réel pouvoir prédictif pour des simulations futures de taux de rachat. Les covariables introduites dans la régression logistique sont le mois d'observation (effet de saisonnalité) et le contexte économique (taux de chômage, taux crédité des contrats, Ibex 35, taux d'intérêt court terme 1Y et long terme 10Y [6]), en plus des covariables considérées dans l'analyse statique. Nous négligeons le décès des assurés lorsque nous réalisons les prévisions futures, car c'est un évènement rare (taux $\simeq 2e^{-4}$). La période d'observation a visiblement une grande influence sur le calibrage du modèle au vu des résultats de la figure 2.1 : nous constatons les bonnes qualités d'adéquation du modèle sur la période d'apprentissage, de même que ses mauvaises qualités prédictives observées en 2007. Les figures 2.1 et 2.2 montrent que le niveau moyen de rachat augmente lorsque le taux de participation aux bénéfices décroît fortement (2003-2004), dénotant une relation claire entre les taux crédités et les taux de rachat. Finalement les résultats ont l'air acceptable bien que le modèle marche très mal en situation extrême ; mais un modèle est-il censé marcher en régime extrême ? La crise financière qui s'est déclarée dans l'année 2007 a très certainement fait évoluer l'importance des facteurs explicatifs clef du rachat en accordant plus d'importance aux facteurs exogènes (indices boursiers, taux d'intérêts) qu'endogènes, ce qui ne semble pas forcément être le cas en régime de croisière. En période de crise, l'hypothèse d'indépendance

6. Données économiques et financières récupérées sur les sites Yahoo Finance et OCDE Stats.

FIGURE 2.1 – Prévisions du taux de rachat collectif (du portefeuille) avec l'inclusion de covariables économiques. Sur la gauche, les prévisions sur l'échantillon d'apprentissage et sur la droite les prévisions sur l'échantillon de validation.



(temporelle et entre agents) semble violée, ce qui fait chuter très nettement le taux de rachat du portefeuille sur les produits mixtes espagnols dans l'année 2007 : concrètement les taux garantis par les contrats sont très intéressants comparés aux taux d'intérêts qui baissent sans arrêt, incitant les assurés à prendre une décision commune (garder à tout prix leur contrat). Le modèle ne prévoit pas cette chute subite car il ne capte visiblement pas les effets dans leurs bonnes proportions en ce qui concerne les variables conjoncturelles (ou exogènes). Cet écart entre prévision et observation s'explique en partie par une hypothèse sous-jacente au modèle :

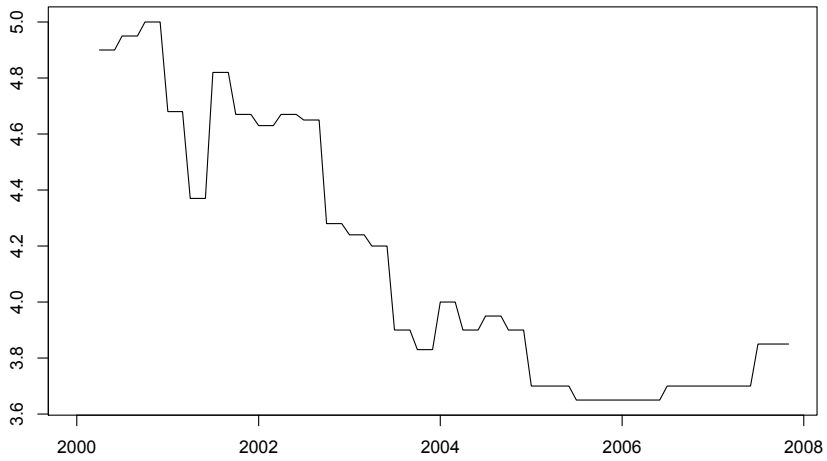


FIGURE 2.2 – Taux crédité mensuel des contrats Mixtes. Ce taux comprend le taux moyen garanti ainsi que taux de PB moyen servi.

quel sera le niveau moyen de rachat dans les mois à venir par rapport aux observations passées (période de référence pendant laquelle le modèle est construit) ? Le taux de rachat prévu sera ensuite ajusté en fonction de cette hypothèse de base. Le seuil d'affectation (0,5) de la réponse (cf section 1.2.4) pourrait aussi jouer sur ces mauvaises prévisions car la duplication des assurés crée un échantillon fortement déséquilibré avec 15 571 rachats sur les 991 010 observations, soit un taux de réponse nulle (non-rachat) de 98,43%. Néanmoins, cette hypothèse ne semble pas être à l'origine de la différence observée en 2007, pour la simple et bonne raison que la différence ne serait pas observée seulement en 2007.
L'aspect dynamique des rachats a également été traîté par des modèles fonctionnels (Ramsay and Silverman (2005) et Ramsay et al. (2009)) d'analyse de survie de type Cox et régression de Weibull (voir les excellents livres de Planchet and Thérond (2006) et Martinussen and Scheike (2006)) permettant de modéliser l'intensité de rachat à chaque moment de la durée de vie du contrat, mais sans succès quant à la bonne prise en compte de l'influence du contexte économique. Pourtant cette approche permettait d'éviter de catégoriser la principale variable d'intérêt, à savoir l'ancienneté du contrat.

Pour synthétiser, les prévisions de taux restent de qualité tant que les conditions économiques ne sont pas significativement différentes du passé, ce qui explique pourquoi l'usage de telles méthodes de prévisions ne s'est pas réellement popularisé dans la pratique actuarielle. La partie suivante illustre parfaitement le phénomène observé ici en 2007, et introduit une approche théorique pour le traitement de ce type de problématique. Cette théorie servira de point de départ dans la modélisation finale.

2.2 Impact de crises de corrélation des comportements

Les assureurs basent souvent leur modèle dynamique de taux de rachat sur une courbe déterministe en forme de S pour tenir compte de l'évolution des comportements de rachat en fonction des scénarios économiques (section 3 du premier chapitre). Cette courbe en S correspond au taux de rachat moyen exprimé en fonction de la différence entre deux taux, notée Δr. L'un de ces deux taux est le taux servi par l'assureur à ses assurés, tandis que l'autre peut valoir le taux du meilleur concurrent ou bien un taux d'intérêt (nous pourrions aussi imaginer que ce Δr représente une différence en termes de réputation...). L'idée courante est qu'un petit Δr ne provoque pas plus de rachats qu'à l'accoutumée, que le taux de rachat évolue de manière monotone et non-linéaire avec Δr, et que même si Δr est très grand certaines personnes resteront en portefeuille parce qu'elles ne prêtent pas spécialement attention à l'évolution des marchés (ou bien que cela les arrange). Le problème est que nous n'avons jamais observé les comportements de rachat dans la situation extrême où Δr est très grand, ce qui implique que la construction d'un modèle stochastique s'appuie davantage sur des jugements d'expert que sur des données statistiques (qui n'existent tout simplement pas !).
Une manière simple d'introduire des effets stochastiques à cette courbe déterministe en S est de supposer une distribution gaussienne autour de la valeur du taux de rachat, mais nous tentons d'expliquer ici pourquoi il serait préférable d'utiliser une distribution bi-modale qui permette de prendre en compte le changement des corrélations entre comportements en scénarios extrêmes. Ces crises de corrélation (Biard et al. (2008), Loisel (2008)) suggèrent de ne pas utiliser l'approximation normale basée sur le théorème central limite (TCL). En effet ce théorème repose sur l'hypothèse de base que les décisions sont indépendantes les unes des autres, or ce ne serait vrai qu'avec la connaissance d'un facteur qui rende compte du niveau

d'information des assurés, de la réputation de la compagnie et du secteur de l'Assurance. Ce facteur serait d'ailleurs déterminant pour comprendre la corrélation des risques de rachat avec d'autres risques tels que le risque de défaut via la matrice de corrélation d'un modèle interne. La crise du marché action et des produits dérivés a été suivie par une crise de corrélation : dans la plupart des cas, la corrélation grandit lors de scénarios défavorables. Il est probable qu'une situation extrême des taux d'intérêts conduise à des rachats massifs (tout du moins anormaux) suivant certaines déclarations politiques ou d'autres facteurs d'environnement (presse...). Par exemple l'une des premières phrases prononcée par les décideurs de pays développés suite au déclenchement de la crise fût : *Nous garantissons l'épargne des contribuables*. Cette attitude trahit leur crainte : ils anticipent des comportements extrêmes (loi binaire 0-1) plutôt qu'un comportement moyenné (gaussien).
Nous nous appliquons dans la suite à développer un modèle simple qui tienne compte de ces crises de corrélation : quand Δr grandit, la corrélation entre les décisions des assurés grandit et l'on passe d'une distribution en cloche en régime classique à une distribution bimodale quand Δr devient grand. Nous présentons en premier lieu le modèle et son interprétation, puis des simulations et des formules analytiques de calcul de la distribution des taux de rachats sont fournies. Des résultats qualitatifs de l'impact de la corrélation sur la distribution du taux de rachat sont développés via l'usage des ordres stochastiques, dans une optique de gestion de risque et de provisionnement basé sur un modèle interne partiel.

2.2.1 Le modèle

Supposons que les assurés se comportent indépendamment avec un taux moyen de rachat $\mu(0)$ quand Δr vaut zéro, que le taux moyen de rachat vaut $1-\epsilon$ avec ϵ très petit quand Δr est très grand (disons 15%), et que la corrélation entre les décisions individuelles vaut $1-\eta$, avec η très petit. Le modèle suivant capture ces notions : soit I_k une variable aléatoire qui prend la valeur 1 si le $k^{\text{ème}}$ assuré rachètent son contrat, 0 sinon. Supposons que

$$I_k = J_k I_0 + (1 - J_k) I_k^{\perp},$$

où J_k correspond à l'indicatrice de l'événement "le $k^{\text{ème}}$ assuré a un comportement moutonnier", I_0 est un consensus collectif de décision de rachat, et $I_k^{\perp}$ est une décision de rachat propre à l'assuré k. La variable aléatoire J_k suit une loi de Bernoulli dont le paramètre p_0 est croissant en Δr, et I_0, $I_1^{\perp}$, $I_2^{\perp}$, ... sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.), dont le paramètre p est aussi croissant en Δr. Ainsi la probabilité de rachat croît avec Δr, et la corrélation (τ de Kendall ou ρ de Spearman) entre I_k et I_l (pour $k \neq l$) est égale à $P(J_k = 1 \mid \Delta r = x)$ quand $\Delta r = x$. Sans conditionner et donc en toute généralité, la corrélation entre I_k et I_l (pour $k \neq l$) vaut

$$\int_0^{+\infty} P(J_k = 1 \mid \Delta r = x)\, dF_{\Delta r}(x).$$

En effet sachant $\Delta r = x$, I_k et I_l (pour $k \neq l$) admettent une copule de Mardia (somme linéaire de la copule indépendante et de la borne supérieure de Fréchet) [7]. L'hypothèse gaussienne est plutôt juste quand $\Delta r = 0$ pour un portefeuille de 20 000 assurés. Nous allons voir avec des valeurs réalistes pour la courbe en S comment lorsque Δr augmente, la densité des taux de rachat évolue progressivement d'une forme en cloche vers une densité bimodale à partir d'un

7. la copule d'I_k et I_l (pour $k \neq l$) n'est pas unique car leurs distributions ne sont pas continues.

certain seuil $\Delta r = x_0$. McNeil et al. (2005) détaillent précisément les problèmes de corrélation et leurs impacts sur la queue de la distribution de probabilité dans un contexte général.

2.2.2 Interprétation

La courbe en S du taux de rachat en fonction de Δr de la figure 2.3 signifie que moins le contrat est attractif et plus l'assuré a de chance de le racheter. La moyenne du taux de rachat est basse en régime économique de croisière (région 1, petit Δr sur la figure 2.3), et augmente significativement quand Δr croît. C'est la traduction d'une opportunité d'arbitrage que l'investisseur peut saisir : un contrat nouvellement acquis offre les mêmes garanties à un prix inférieur en cas de hausse des taux (toujours dans un contexte de rendement contractuel fixe), ce qui mécaniquement améliore la rentabilité. Si à l'inverse les taux d'intérêt chutent, alors l'assureur peut aussi choisir d'abaisser le taux crédité à l'assuré (suivant les modalités du contrat) pour inciter les assurés à racheter avant que ces derniers ne chutent davantage).
Par conséquent la région 1 de la figure 2.3 correspond à la zone dans laquelle les décisions des assurés sont indépendantes (la corrélation tend vers 0), alors que la région 2 est celle des comportements corrélés (la corrélation tend vers 1). En fait la corrélation entre les comportements des assurés est quasi-nulle aussi longtemps que l'économie est en "bonne santé", le taux de rachat peut donc être modélisé par une loi normale dont la moyenne et l'écart-type sont observés. C'est pourquoi la gaussienne (figure 2.4) est la distribution adaptée en région 1.
Inversement, la forte pente de la hausse du taux de rachat pour un certain niveau Δr en figure 2.3, suivie d'un plateau qui est le taux de rachat maximal atteignable (borne issue d'un jugement d'expert puisque jamais observée en pratique par principe), reflète la détérioration des conditions économiques. Le point crucial consiste à réaliser que l'hypothèse d'indépendance est largement erronée dans un tel contexte : la corrélation entre les décisions des assurés fait changer la distribution du taux de rachat. C'est la conséquence de deux comportements extrêmement risqués dans lesquels presque tout le monde rachète ou quasiment personne ne rachète. La distribution la plus adaptée pour l'expliquer est la loi bimodale, illustrée en figure 2.4. La différence majeure avec le modèle gaussien est que la moyenne du taux résulte de deux pics de densité placés aux extrémités du domaine de définition du taux.

Remarquons qu'un comportement irrationnel des assurés peut également mener à des crises de corrélation même dans le cas où Δr est petit, ce qui (nous le verrons en section 2.2.5) est d'ailleurs la situation qui a le plus d'impact sur les besoins en capitaux ou augmentation des réserves de l'assureur. Un comportement irrationnel désigne ici un comportement atypique

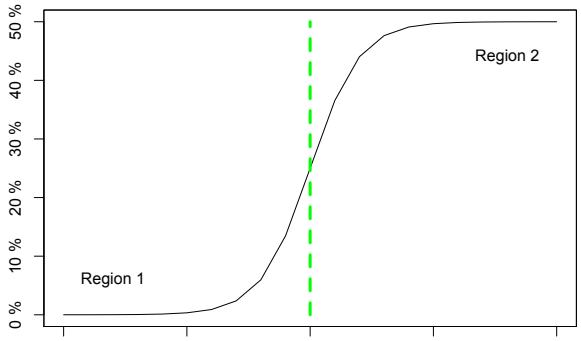
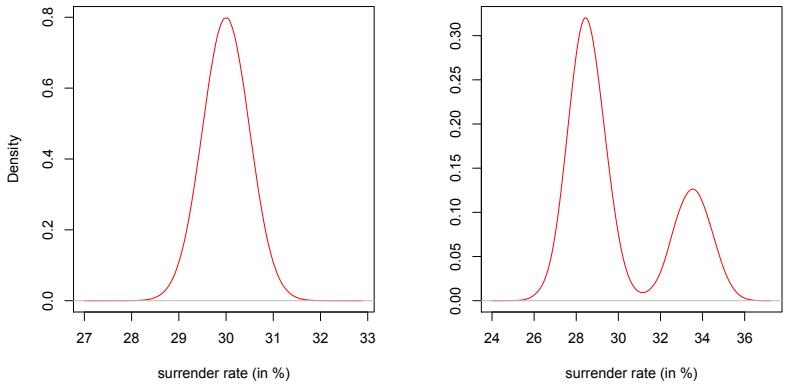


FIGURE 2.3 – Taux de rachat versus Δr.

FIGURE 2.4 – Sur la gauche, la densité de la loi normale et sur la droite la densité bimodale (la moyenne vaut 30 dans les deux cas).



par rapport à l'expérience qu'a la compagnie, dû à des rumeurs ou des recommendations de journalistes ou brokers. D'un point de vue financier, un assuré adopte un comportement irrationnel s'il ne rachète pas son contrat bien qu'il soit gagnant dans cette opération. Mais ce comportement d'irrationalité (financier) n'est pas si rare que ça à cause des contraintes fiscales et de la complexité des contrats actuels d'Assurance-Vie, munis de garanties et d'options de plus en plus compliquées. Nous pouvons cependant remarquer que les agents semblent de plus en plus rationnels sur le marché américain (qui contient beaucoup de "variable annuities"), et que quelquepart l'incertitude concernant la rationalité future des assurés est capturé par notre modèle de crise de corrélation.

2.2.3 Distribution des taux de rachat

Approche combinatoire

Considérons un portefeuille de $n \geq 2$ assurés. Soit

$$N = \sum_{k=1}^{n} J_k$$

le nombre de personnes ayant un comportement moutonnier, et

$$M = \sum_{k=1}^{n} I_k$$

le nombre de personnes qui rachètent leur contrat. Rappelons que

$$I_k = J_k I_0 + (1 - J_k) I_k^{\perp},$$

où J_k correspond à l'indicatrice de l'événement "le $k^{\text{ème}}$ assuré adopte un comportement moutonnier", et J_k a une distribution de Bernoulli de paramètre p_0, et où I_0, $I_1^{\perp}$, $I_2^{\perp}$, ... sont des variables aléatoires i.i.d. de paramètre p (et indépendantes des $(J_l)_{l\geq 1}$). Si le consensus

général est de racheter ($I_0 = 1$), alors pour M valant un entier $k \in [\![0, n]\!]$, le nombre N d'assurés "moutons" doit être inférieur ou égal à k, sinon nous aurions $M \geq N > k$. Par le même raisonnement, si le comportement moutonnier consiste à ne pas racheter ($I_0 = 0$), alors pour M égal à un entier $k \in [\![0, n]\!]$, le nombre N d'assurés "moutons" doit être inférieur ou égal à $n-k$, sinon nous aurions $M \leq n - N < n - (n-k) = k$. Nous obtenons à partir de la formule des probabilités totales que pour $0 \leq k \leq n$,

$$P(M = k) = P(M = k \mid I_0 = 0)P(I_0 = 0) + P(M = k \mid I_0 = 1)P(I_0 = 1)$$

$$= \sum_{i=0}^{k} P\left(M = k \mid I_0 = 1,\, N = i\right) P\left(I_0 = 1,\, N = i\right)$$

$$+ \sum_{j=0}^{n-k} P\left(M = k \mid I_0 = 0,\, N = j\right) P\left(I_0 = 0,\, N = j\right).$$

L'indépendance mutuelle entre les $(J_k)_{k\geq 1}$ et les $(I_l^{\perp})_{l\geq 1}$, avec $0 \leq k \leq n$ entraîne que

$$P(M = k) = p \sum_{i=0}^{k} a_{i,k} + (1-p) \sum_{j=0}^{n-k} b_{j,k},$$

avec pour $0 \leq i \leq k$,

$$a_{i,k} = C_n^i p_0^i (1-p_0)^{n-i} C_{n-i}^{k-i} p^{k-i} (1-p)^{n-k},$$

et pour $0 \leq j \leq n-k$

$$b_{j,k} = C_n^j p_0^j (1-p_0)^{n-j} C_{n-j}^{k} p^{k} (1-p)^{n-j-k}.$$

Remarquons que pour k fixé, les $a_{i,k}$, $0 \leq i \leq k$ et les $b_{j,k}$, $0 \leq j \leq n-k$ peuvent être calculés grâce aux formules récursives suivantes : pour $0 \leq i \leq k$, nous avons

$$\frac{a_{i+1,k}}{a_{i,k}} = \frac{C_n^{i+1}}{C_n^i} \frac{p_0}{p\,(1-p_0)} \frac{C_{n-i-1}^{k-i-1}}{C_{n-i}^{k-i}} = \frac{k-i}{i+1} \frac{p_0}{p\,(1-p_0)}$$

et pour $0 \leq j \leq n-k$, nous avons

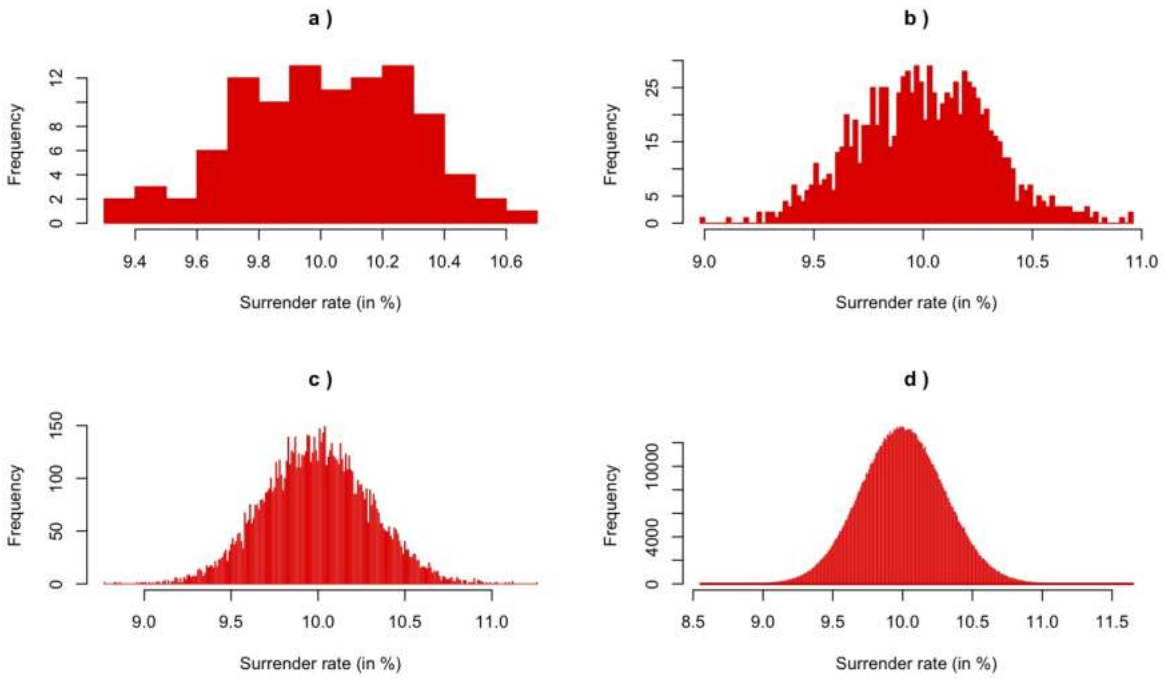
$$\frac{b_{j+1,k}}{b_{j,k}} = \frac{C_n^{j+1}}{C_n^j} \frac{p_0}{(1-p)\,(1-p_0)} \frac{C_{n-j-1}^{k}}{C_{n-j}^{k}} = \frac{n-j-k}{j+1} \frac{p_0}{(1-p)\,(1-p_0)}.$$

Notons qu'il est préférable de commencer avec a_{i_0} and b_{j_0} tels que a_{i_0} and b_{j_0} soient assez grands dans le but de minimiser les erreurs d'arrondis lors du calcul de

$$a_0 = b_0 = (1-p_0)^n\, C_n^k p^k (1-p)^{n-k}$$

qui sont en général assez petits. Viquerat (2010) propose des algorithmes efficaces (et leur précision) pour effectuer ce type de calcul.

FIGURE 2.5 – Effet du nombre de simulations sur la distribution du taux de rachat : a) 100, b) 1 000, c) 10 000 and d) 1 000 000. Pas de comportement moutonnier, probabilité individuelle de rachat égale à 10%, 10 000 assurés en portefeuille.



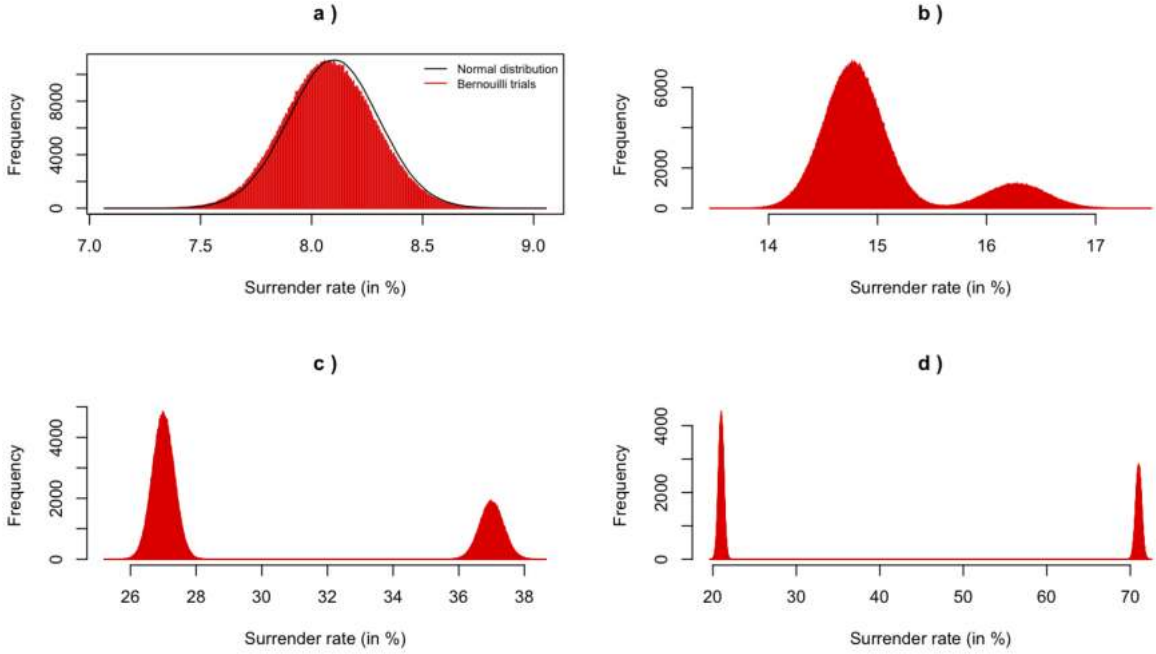
Approche par simulation

En pratique, l'utilisation de simulations est courante pour l'évaluation du risque de rachat, parmi les nombreux types de risque d'un modèle interne complexe. Le nombre de simulations est clef pour obtenir une approximation précise de la distribution du taux de rachat, quel que soit le contexte socio-économique. Le nombre d'assurés en portefeuille a son importance car il permet de diminuer la dispersion des valeurs de taux de rachat, bien que cela n'affecte pas vraiment la forme de la distribution. La figure 2.5 confirme ces remarques, nous prendrons donc dans la suite un nombre de simulations égal à 1000 000 et un nombre d'assurés de 10 000. Il va sans dire de l'effet capital de la probabilité individuelle de rachat p, qui joue directement sur la moyenne de la distribution et engendre un profil naturellement bien plus risqué en moyenne.

Concentrons nous maintenant sur l'effet de la corrélation, le coeur de ce chapitre. Le paramétre de corrélation p_0 (probabilité de suivre le consensus collectif) joue également un rôle crucial : la hausse de p_0 remodèle la forme de la distribution du taux de rachat. Une crise économique provoque naturellement l'augmentation simultanée de la probabilité de rachat et de la corrélation, ce qui est une très mauvaise nouvelle pour l'assureur qui doit faire face à une situation dans laquelle les modes s'équilibrent (risque élevé dans les deux cas) et s'écartent. La figure 2.6 illustre cette déformation.

Pour un certain Δr (et donc pour un p_0 donné en théorie), plus le paramètre de corrélation p_0 est grand et plus la forme de la densité de rachat devient bimodale. Tout l'intérêt de

FIGURE 2.6 – Evolution d'une distribution gaussienne de taux de rachat à une bimodale. De haut en bas et de gauche à droite, les probabilités de rachat et de comportement moutonnier valent : a) 8% et 0%, b) 15% et 1.5%, c) 30% et 10% et d) 42% et 50%. 1 000 000 simulations et 10 000 assurés.

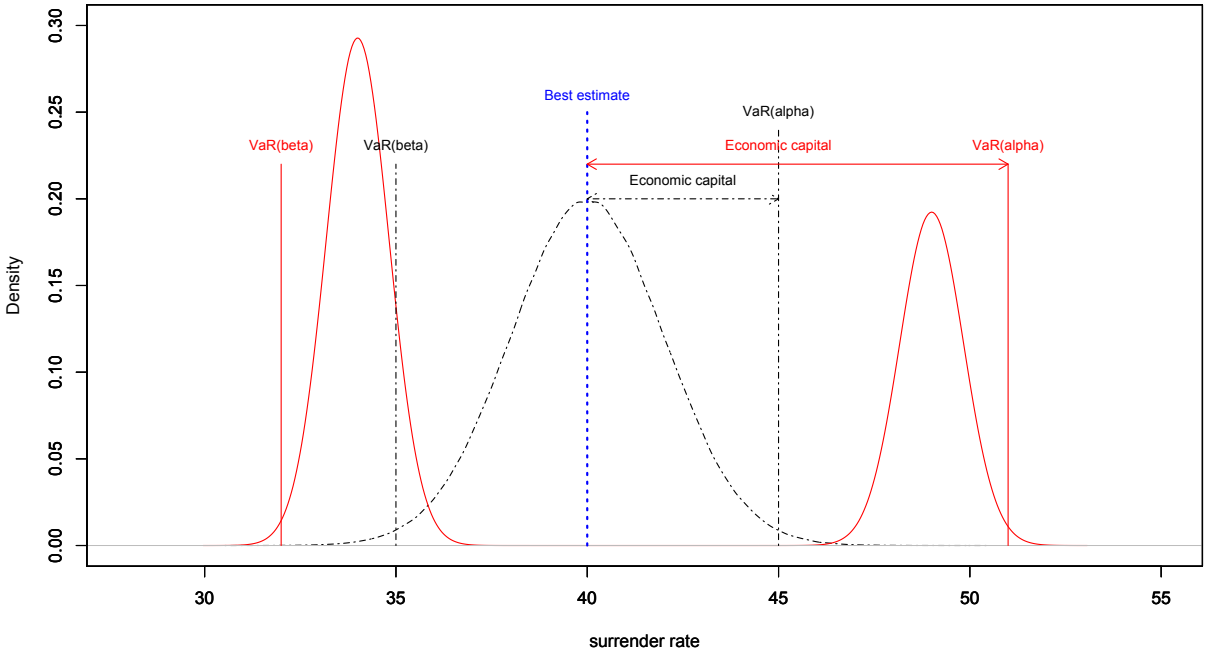


l'assureur se porte sur la quantification de la différence entre ces distributions en termes de risque de comportement. A cette fin, certaines mesures de risque dont la *Value-at-Risk* (ou *VaR*) peuvent servir d'indicateurs de cet écart. La *VaR* est définie pour une variable aléatoire X et un seuil α par :

$$VaR_\alpha(X) = \inf\{x \in X,\ F_X(x) \geq \alpha\}.$$

La variable aléatoire X est le taux de rachat dans notre application, ce qui signifie que l'assureur s'attend à subir un taux de rachat inférieur à *VaR* avec α% de confiance. En Assurance-Vie, nous posons en général $\alpha = 99.5\%$, car ce sont précisemment les recommandations de la directive européenne *Solvabilité II* pour la prise en compte des chocs.
Parfois la situation antagoniste (chute du taux de rachat) préoccupe également l'assureur car son exposition peut devenir trop grande par rapport à ses contraintes de capitaux (par exemple les garanties décès sont évaluées avec certaines prévisions d'exposition au risque basées sur les rachats du passé), ou parce que les taux d'intérêts lui sont défavorables (sur des produits à taux garantis). Dans ce cas, nous adaptons le raisonnement en considérant les VaR_α (côté droit, risque de rachats massifs) et VaR_β (côté gauche, très peu de rachat comparé aux prévisions) illustrées en figure 2.7.

FIGURE 2.7 – Densité du taux de rachat pour des comportements indépendants (en noir et pointillé) et pour des comportements corrélés (en rouge et trait plein) : le capital économique associé (qui vaut la différence entre la VaR_α et le "best estimate") s'accroît avec la corrélation.



2.2.4 Comparaison qualitative par ordres stochastiques

Soit $M_{(p,p_0)}$ le nombre d'assurés qui rachètent leur contrat dans le modèle où $P(J_1 = 1) = p_0$ et $P(I_1 = 1) = p$. Examinons comment les valeurs de p et p_0 affectent la distribution du taux de rachat conditionnel. Nous utilisons pour cela l'ordre stochastique $s-$convex (Lefèvre and Utev (1996) et Denuit et al. (1998)).
Sachant deux variables aléatoires X et Y, pour $s = 1, 2, \ldots$, nous avons

$$X \leq^{\mathcal{D}}_{s-cx} Y \text{ si } \mathbb{E}[\phi(X)] \leq \mathbb{E}[\phi(Y)] \text{ pour tout } s\text{-fonction convexe } \phi : \mathcal{D} \to \mathbb{R}, \tag{2.1}$$

dont la s-eme dérivée existe et satisfait $\phi^{(s)} \geq 0$. Les (s-1) premiers moments de X et Y sont d'ailleurs nécessairement égaux. L'ordre $\leq^{\mathcal{D}}_{1-cx}$ correspond à l'ordre stochastique simple $\leq_1$, $\leq^{\mathcal{D}}_{2-cx}$ est l'ordre convexe usuel $\leq_2$ (qui implique en particulier que $Var(X) \leq Var(Y)$). De plus, X est dit plus petit que Y dans l'ordre convexe croissant (noté $\leq_{icx}$) si

$$\mathbb{E}\left(f(X)\right) \leq \mathbb{E}\left(f(Y)\right)$$

pour toute fonction convexe croissante telle que l'espérance existe.

Proposition 3. *Lorsque le paramètre de corrélation est fixé, le nombre de rachats est stochastiquement croissant en p : pour $p_0 \in (0, 1)$ fixé, si $p < p'$ alors*

$$M_{(p,p_0)} \leq_1 M_{(p',p_0)}.$$

Démonstration. Cette proposition découle de résultats élémentaires sur les ordres stochastiques impliquant des distributions Binomiales et de Bernoulli. □

Proposition 4. *Lorsque la probabilité individuelle de rachat p est fixée, le paramètre de corrélation induit un ordre 2-convex du nombre de rachats : pour $p \in (0,1)$ fixé, si $p_0 < p_0'$ alors*

$$M_{(p,p_0)} \leq_2 M_{(p,p_0')}.$$

Démonstration. Sachant que le nombre de comportements moutonniers N vaut k, le nombre total de rachats est

$$M_{(p,k)} = k.I_0 + 0.I_1^{\perp} + 0.I_2^{\perp} + ... + 0.I_k^{\perp} + 1.I_{k+1}^{\perp} + ... + 1.I_n^{\perp}.$$

Sachant que $N = k^{'}$ avec $k \leq k^{'}$, nous avons

$$M_{(p,k')} = k^{'}.I_0 + 0.I_1^{\perp} + ... + 0.I_k^{\perp} + ... + 0.I_{k'}^{\perp} + 1.I_{k'+1}^{\perp} + ... + 1.I_n^{\perp}.$$

Nous pouvons comparer les deux variables aléatoires $M_{(p,k)}$ et $M_{(p,k')}$ par un ordre de majorisation (voir par exemple Marshall and Olkin (1979)). Notons $Z^{\downarrow} = (z_1^{\downarrow}, \ldots, z_K^{\downarrow})$ le vecteur des mêmes composantes que Z (quelconque) classées en ordre décroissant. Connaissant deux vecteurs $Y = (y_1, \ldots, y_K)$ et $Z = (z_1, \ldots, z_K)$ de taille $K \geq 1$ tels que

$$\sum_{i=1}^{K} y_i = \sum_{i=1}^{K} z_i,$$

rappelons que Z est dit majorant Y si pour tout $j \leq K$,

$$\sum_{i=1}^{j} y_i^{\downarrow} \leq \sum_{i=1}^{j} z_i^{\downarrow}.$$

D'après Marshall and Olkin (1979), si le vecteur $\alpha = (\alpha_0, ..., \alpha_n)$ est plus petit que le vecteur $\beta = (\beta_0, ..., \beta_n)$ dans l'ordre de majorisation partielle, et si les X_i sont i.i.d., alors nous obtenons l'ordre convexe suivant :

$$\sum_i \alpha_i X_i \leq_2 \sum_i \beta_i X_i.$$

Nous posons $X_i = I_i^{\perp}$ et

$$(\alpha_0, ..., \alpha_n) = (k, \underbrace{0, ..., 0}_{k\ fois}, 1, ..., 1) \text{ et } (\beta_0, ..., \beta_n) = (k^{'}, \underbrace{0, ..., 0}_{k'\ fois}, 1, ..., 1).$$

Pour $k \leq k'$, le vecteur $(\beta_0, ..., \beta_n)$ majore clairement le vecteur $(\alpha_0, ..., \alpha_n)$. De plus, la variable aléatoire $(N \sim Bin(n, p_0))$ est stochastiquement croissante en p_0. Nous pouvons donc conclure que pour $p_0 \leq p_0'$,

$$M_{(p,p_0)} \leq_2 M_{(p,p_0')},$$

où $M_{(p,p_0)}$ est le nombre d'assurés qui rachètent quand la probabilité de se comporter en mouton vaut p_0 et quand la probabilité individuelle de rachat est p. □

Proposition 5. *Dans le modèle où p et p_0 augmentent simultanément Δr, Δr induit un ordre convexe croissant du nombre de rachats : soit M (resp. M') le nombre de rachats quand $\Delta r = x$ (resp. $\Delta r = x'$). Si $x < x'$ alors nous avons*

$$M \leq_{icx} M'.$$

Démonstration. Nous utilisons les mêmes arguments que dans les preuves des propositions 3 et 4. En combinant ces arguments, le résultat est immédiat car quand Δr croît, p et p_0 augmentent. □

Ces propositions permettent l'interprétation de deux résultats pratiques pour gérer le risque :

- la proposition 3 implique que l'espérance, la *VaR* à n'importe quel seuil $\alpha \in (0,1)$ et les primes stop-loss $E\left[(M-m)_+\right]$ pour $0 \leq m \leq n$ sont croissantes en p et en Δr.
- la proposition 4 dit que la variance et les primes stop-loss $E\left[(M-m)_+\right]$ pour $0 \leq m \leq n$ sont croissantes en p_0 (à p fixé). Elle montre aussi que si $p_0 < p_0'$, il existe un niveau $\alpha_0 \in (0,1)$ tel que pour $\alpha > \alpha_0$,

$$VaR_\alpha\left(M_{(p,p_0)}\right) < VaR_\alpha\left(M_{(p,p_0')}\right)$$

  car le critère de Karlin-Novikov énonce dans ce cas que les fonctions de répartition de $M_{(p,p_0)}$ et $M_{(p,p_0')}$ ne peuvent se croiser qu'une fois.

Il n'est évidemment pas surprenant d'apprendre qu'augmenter le paramètre de corrélation et/ou la probabilité marginale de rachat provoque une plus grande mesure de risque en général, mais le but ici est aussi de déterminer l'importance de l'impact de cette corrélation sur le besoin en capital économique et sur la valeur du compte (notion introduite par la suite) dans un contexte réel.

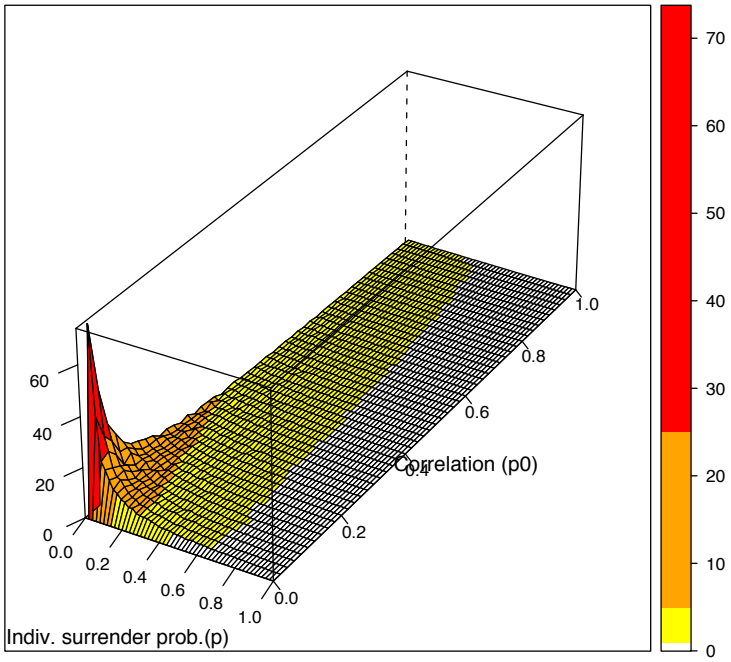
2.2.5 Ecarts de VaR et taille du portefeuille

Dans une perspective Solvabilité II, nous nous focalisons sur une analyse détaillée des écarts de VaR à 99.5 %. La figure 2.8 récapitule l'effet de l'accroissement de la corrélation entre décisions des assurés sur la VaR. Pour une probabilité individuelle de rachat donnée, disons d'1%, une corrélation passant de 0 à 1% augmente la $VaR_{99.5\%}$ de 30 à 50%. Nous pouvons notamment remarquer que les grands écarts **positifs** de VaR sont concentrés sur de petites valeurs de corrélation lorsque nous considérons une faible propension au rachat. Ce résultat remarquable nous suggère de définir des classes de risque en termes de *sensibilité* (par rapport à la corrélation) :

- *hyper sensible* (zone rouge dans la figure 2.8) : $p \in]0, 0.05]$ et $p_0 \in]0, 0.1]$;
- *sensible* (zone orange) : $p \in]0, 0.05]$ et $p_0 \in [0.1, 0.4]$, ou $p \in]0.05, 0.2]$ et $p_0 \in]0, 0.3]$;
- *peu sensible* (zone jaune) : autres situations.

Dans la configuration *hyper sensible*, la $VaR_{99.5\%}$ peut augmenter jusqu'à 70% ! Dans la configuration *sensible*, l'assureur peut voir sa $VaR_{99.5\%}$ augmenter de 5 à 25 %, ce qui est moins risqué mais reste très préoccupant. Enfin, la configuration *peu sensible* est une zone dans laquelle l'assureur peut être serein car il semble avoir déjà assez provisionné (la VaR est assez grande). Ces observations montrent que que la situation la plus dangereuse en termes d'écart de provisionnement pour l'assureur correspond à l'apparition de la corrélation dans des scénarios où la probabilité moyenne de rachat est très faible.

FIGURE 2.8 – Ecart relatif (en %) des VaR versus p_0 et p.



La taille de la compagnie pourrait également être un facteur-clef : pour tester son impact, nous avons simulé les écarts de capital économique (EC) lié à la VaR pour une probabilité annuelle moyenne de rachat réaliste de 8,08 % (BE pour *best-estimate* dans le tableau 2.4) et une corrélation passant de 0 % à 50 %. Etudier les écarts de capitaux économiques revient à étudier les écarts de VaR (voir graphe 2.7). Certaines compagnies d'assurance pourraient penser que leur taille leur évite des scénarios catastrophiques grâce à l'effet de la mutualisation. L'analyse du tableau 2.4 démontre que le nombre d'assurés en portefeuille n'a pas de forte influence sur le calcul des marges de risque. En effet, la différence de capital économique ΔEC nécessaire baisse de manière dérisoire, même en passant de 5 000 à 500 000 assurés ! Une application complète sur un exemple de produit est menée dans Loisel and Milhaud (2011).

Taille portefeuille	BE	EC ($VaR^{Normale}_{99.5\%}$)	Corrélation	EC ($VaR^{Bimodale}_{99.5\%}$)	ΔEC
Petite :			$p_0 = 0.05$	6.26%	4.5%
5000	8.08%	1.76%	$p_0 = 0.2$	20.42%	18.66%
assurés	($p_0 = 0$)	($p_0 = 0$)	$p_0 = 0.5$	48.54%	46.78%
Moyenne :			$p_0 = 0.05$	5.1%	4.59%
50 000	8.08%	0.51%	$p_0 = 0.2$	19.01%	18.5%
assurés	($p_0 = 0$)	($p_0 = 0$)	$p_0 = 0.5$	46.63%	46.12%
Grande :			$p_0 = 0.05$	4.73%	4.59%
500 000	8.08%	0.1426%	$p_0 = 0.2$	18.56%	18.41%
assurés	($p_0 = 0$)	($p_0 = 0$)	$p_0 = 0.5$	46.16%	46.01%

TABLE 2.1 – Impact de la taille du portefeuille sur la VaR (100 000 simulations).

2.3 Application sur un portefeuille d'Assurance Vie réel

Dans cette section, le but n'est pas forcément de fournir des résultats précis. Ces résultats dépendent fortement du portefeuille étudié en termes de calibration, et ne sont donc pas généralisables à tout bout de champ. Une bonne quantification de l'impact de la corrélation des comportements sur le capital économique repose sur des données relativement complètes et fiables. L'objectif est donc d'illustrer d'un point de vue pratique les notions que nous avons développées précédemment.

Les produits étudiés ici sont des produits d'épargne extraits du portefeuille espagnol d'un grand réassureur. Les données, mensuelles, couvrent la période allant de Février 2002 à Décembre 2007 ; et le nombre d'assurés varie de 291 au départ à 25 766 en Juillet 2006. En moyenne, il y a 17 657 assurés présents chaque mois dans le portefeuille. En principe, nous considérons dans le monde de l'assurance que le spread Δr représente la différence entre le taux crédité par la concurrence et celui du contrat (**?**). Malheureusement cette information est difficilement accessible, ce qui explique que l'on considère en général que le taux concurrentiel s'exprime comme le taux sans risque auquel on ajoute une constante c :

$$\Delta r = (\text{taux sans risque} + c) - \text{taux crédité}.$$

Pour simplifier car cela ne change rien à l'étude, nous posons $c = 0$. Le graphique 2.9 montre que le taux crédité du contrat est supérieur au taux sans risque sur toute la période étudiée, d'où $\Delta r < 0$. Un lissage exponentiel du taux de rachat en fonction de Δr sur ces données donne le graphique 2.10 : plus le spread est petit, plus le taux de rachat est grand. Cette observation est raisonnable puisqu'elle corrobore le fait que le contrat devienne moins avantageux (le taux crédité baisse). C'est ce lissage exponentiel qui nous permet de reconnaître la forme de courbe en S que nous avons évoquée plusieurs fois.
Au début du chapitre, nous soulignions que seules les conditions de marché correspondantes à la Région 1 du graphe 2.3 avaient été observées. Nous supposons dans la suite que le taux mensuel maximum de rachat atteignable est de 3,5% (avis d'expert pour définir le seuil du

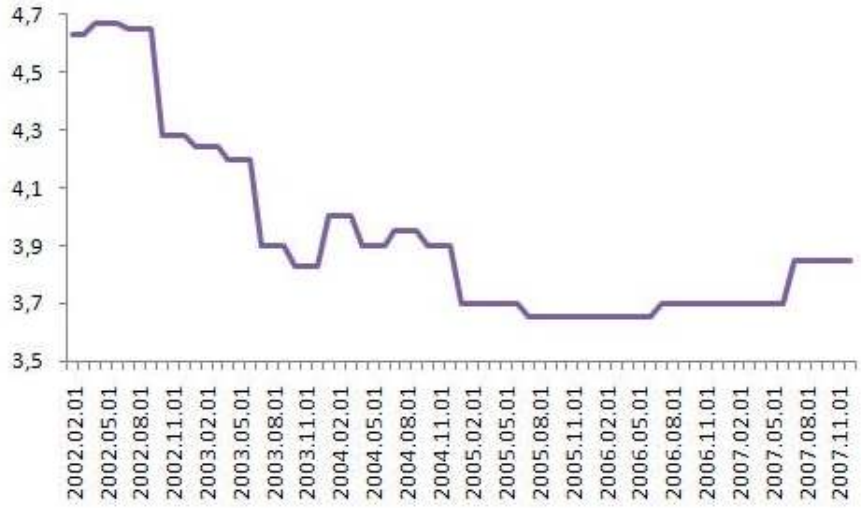
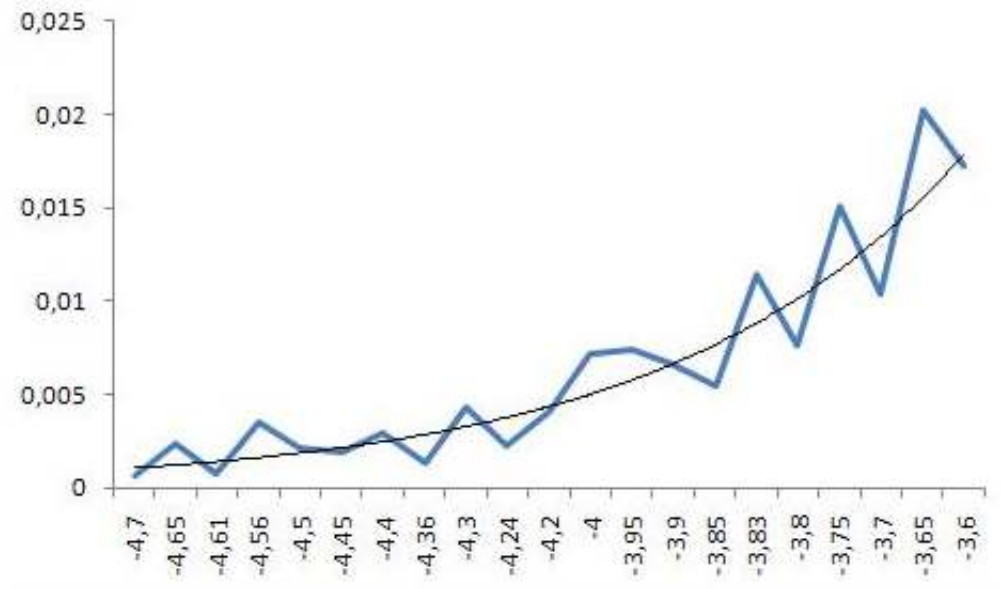


FIGURE 2.9 – Evolution du spread entre le taux crédité et le taux d'intérêt sans risque du marché.

FIGURE 2.10 – Taux de rachat mensuel versus Δr.



plateau en Région 2). C'est évidemment une première source d'erreur car ce choix est arbitraire, mais nous nous intéressons principalement à l'impact de la corrélation entre comportements ici. C'est la raison pour laquelle nous introduisons ce modèle comme exemple réel et illustratif, en gardant en tête que la prévision des taux de rachat va au-delà des objectifs de ce chapitre. Soit Y la variable aléatoire du taux de rachat. La moyenne empirique du taux de rachat en **Région 1**, notée $\tilde{Y}_{n_1}$, vaut

$$\tilde{Y}_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{i,1}, \text{ où } n_1 \text{ est le nombre d'observations } Y_{i,1} \text{ en Région 1.}$$

L'écart-type empirique, noté $\tilde{\sigma}_{n_1}$, est donné par l'estimateur non-biaisé usuel :

$$\tilde{\sigma}_{n_1} = \sqrt{\frac{1}{(n_1 - 1)} \sum_{i=1}^{n_1} (Y_{i,1} - \tilde{Y}_{n_1})^2}.$$

Les applications numériques donnent comme résultats $\tilde{Y}_{n_1} = 6.73 \times 10^{-3}$, et $\tilde{\sigma}_{n_1} = 4.28 \times 10^{-3}$. Ainsi le coefficient de variation, noté $\tilde{\nu}_{n_1}$ et défini par $\tilde{\nu}_{n_1} = \dfrac{\tilde{\sigma}_{n_1}}{\tilde{Y}_{n_1}}$, vaut 63,5% en Région 1. Cette valeur reflète une forte surdispersion des données.

Ces statistiques sont clairement discutables car basées sur peu d'observations, mais la surdispersion provient principalement d'effets cohortes conjugués aux contraintes de pénalité. La dépendance calendaire peut aussi jouer un rôle.

A cause du manque d'observation en Région 2, la deuxième hypothèse consiste à supposer le coefficient de variation constant quelle que soit la moyenne du taux de rachat. Cela nous permet d'approcher l'écart-type de la Région 2, donné par $\tilde{\sigma}_{n_2} = \dfrac{\tilde{Y}_{n_2}}{\tilde{Y}_{n_1}} \, \tilde{\sigma}_{n_1} \simeq 0.022$. Encore une fois, ce choix est arbitraire mais il a l'avantage d'être relativement objectif.

Nous simulons donc en Région 1 des expériences de Bernoulli de paramètre p_1 égal à $\tilde{Y}_{n_1}$, et vérifions que la distribution des taux de rachat est proche d'une gaussienne dont la moyenne

est $\tilde{Y}_{n_1}$ et l'écart-type vaut $\tilde{\sigma}_{n_1}$. Ainsi, le k^e assuré prend la décision $I_k \sim Bernoulli(p_1)$, avec la décision de suivre le consensus de marché donnée par l'indicatrice Bernoulli J_k de paramètre $p_0 = 0$ (décisions indépendantes car nous sommes en Région 1).
En Région 2 $I_k \sim Bernoulli(p_2)$ (avec $p_2 = \tilde{Y}_{n_2} = 0.035$), et la corrélation existante joue sur le paramètre p_0 de J_k. On pose $J_k \sim Bernoulli(p_0)$, avec $p_0 = 0.5$. Le nombre de simulations est fixé à 1 000 000 dans toute cette partie, et le nombre d'assurés vaut 17 657. Le graphique 2.11 expose les résultats de ces simulations : en haut à gauche, la simulation des comportements confirme l'approximation Normale en considérant des expériences de Bernoulli indépendantes. La gaussienne $\mathcal{N}(m = \tilde{Y}_{n_1}, \sigma = \tilde{\sigma}_{n_1})$ se superpose bien et la calibration de l'écart-type ne semble pas être trop mauvaise. En bas à droite, il est clair que l'approximation Normale n'est pas du tout appropriée pour modéliser les comportements assurés ! Les autres figures correspondent aux étapes de transition entre ces deux situations caricaturales.

Remarque 2.3.1. *Nous avons converti les taux mensuels en annuels car la VaR est estimée sur un horizon d'un an en Assurance. La dépendance temporelle pourrait aussi être considérée et aurait un impact fort sur le risque de rachat supporté par la compagnie. Ceci dit, nous nous focalisons ici sur une période d'un an, et laissons cette question pour de futures investigations. Cela mène à* $\tilde{Y}_{n_1} \simeq 0.0808 = 8.08\%$, $\tilde{\sigma}_{n_1} \simeq 0.05$, $\tilde{Y}_{n_2} \simeq 0.42 = 42\%$ *et* $\tilde{\sigma}_{n_2} \simeq 0.264$.

L'impact du contexte économique sur la *VaR* est impressionnant : par exemple, si la corrélation augmente de 0% à 50% dans un contexte économique classique, la $VaR_{99.5\%}$ augmente de 514% (différence notée $\Delta VaR_{99.5\%}$). Cela signifie que le coût du capital d'un assureur voulant se couvrir contre un taux de rachat très élevé à cause de crise de corrélation devrait augmenter de 510% !

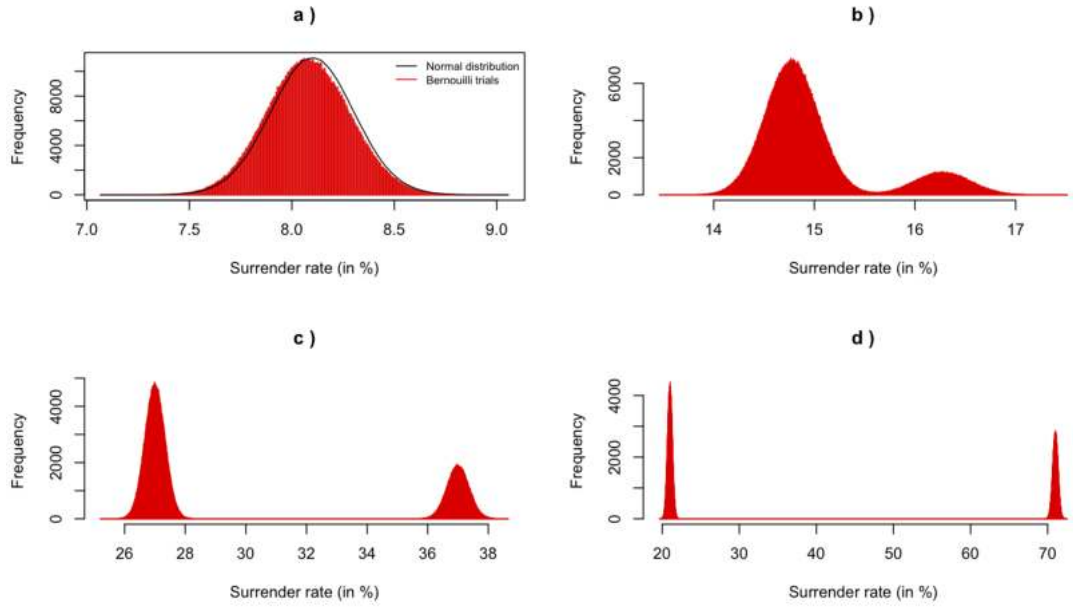


FIGURE 2.11 – Evolution de la distribution du taux de rachat selon le contexte économique (1 000 000 simulations, 17657 assurés). Probabilités individuelles de rachat et probabilités de comportement moutonnier : a) 8.08% et 0% (loi normale théorique en noir), b) 15% et 1.5%, c) 30% et 10% et d) 42% et 50%.

TABLE 2.2 – Estimations des *Value-at-Risk* (VaR) et *Tail-Value-at-Risk* (TVaR) du taux de rachat avec différentes corrélations. S signifie "soft" contexte économique ($p = 8.08\%$), M "Medium" contexte ($p = 20\%$) et H "Hard" contexte ($p = 42\%$). p_0 est la valeur de la corrélation, le nombre de simulations est 100 000.

Context éco	p_0	$VaR_{95\%}$	$VaR_{99.5\%}$	$\Delta VaR_{99.5\%}$	$TVaR_{90\%}$	$TVaR_{99\%}$	$\Delta TVaR_{99\%}$
S	0	8.823	8.993	0 %	8.844	8.996	0 %
	0.01	9.565	9.746	+8.4 %	9.576	9.758	+8.5 %
	0.02	10.471	10.658	+18.5 %	10.481	10.686	+18.8 %
	0.05	13.297	13.501	+50.1 %	13.314	13.514	+50.2 %
	0.15	22.659	22.886	+154.5 %	22.666	22.867	+154.2 %
	0.3	36.557	36.874	+310 %	36.577	36.909	+310.3 %
	0.5	54.992	55.218	+514 %	55.01	55.233	+514 %
M	0	21.028	21.204	0 %	21.046	21.22	0 %
	0.01	21.668	21.945	+3.5 %	21.688	21.939	+3.4 %
	0.02	22.416	22.693	+7 %	22.433	22.713	+7 %
	0.05	24.879	25.219	+18.9 %	24.905	25.214	+18.8 %
	0.15	32.933	33.301	+57 %	32.96	33.33	+57.1 %
	0.3	45.013	45.341	+113.8 %	45.038	45.368	+113.8 %
	0.5	61.018	61.301	+189.1 %	61.042	61.348	+189.1 %
H	0	43.201	43.563	0 %	43.226	43.513	0 %
	0.01	43.693	43.977	+1 %	43.719	44.054	+1.2 %
	0.02	44.231	44.503	+2.2 %	44.245	44.522	+2.3 %
	0.05	45.959	46.378	+6.5 %	45.983	46.374	+6.6 %
	0.15	51.735	52.053	+19.5 %	51.748	52.071	+19.7 %
	0.3	60.435	60.735	+39.4 %	60.449	60.768	+39.7 %
	0.5	71.965	72.237	+65.8 %	71.977	72.265	+66.1 %

La tableau 2.2 résume ces chiffres pour différents contextes économiques, niveaux de corrélation et niveaux de VaR. Nous exposons aussi les résultats sur une autre mesure de risque, la *Tail-Value-at-Risk* ($TVaR$) qui prend en compte la queue de distribution. La $TVaR$ est définie par

$$TVaR_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 VaR_\beta(X)d\beta.$$

On remarque que les résultats sur l'accroissement de la $TVaR$ par l'introduction de la corrélation ressemblent fortement à ceux de la VaR.

2.4 Ecart entre hypothéses standard et modèle réaliste

Dans quelle mesure le provisonnement est-il affecté par ces hypothèses ? Après avoir exprimé l'impact que pourrait provoquer l'hypothèse de non-indépendance des comportements sur la distribution des taux de rachat, nous commentons les conséquences potentielles des effets de corrélation sur les réserves d'un assureur. Les grandes compagnies d'assurance sont souvent des groupes dans lesquels chaque entité a sa propre gestion du risque de rachat, avec ses propres hypothèses sur le risque moyen pour définir le pricing du produit et sa rentabilité future. Il n'est donc pas surprenant que parfois, les idées et pratiques divergent au sein même d'un groupe : ici, nous simplifions l'étude en considérant que l'entreprise détient un compte dont la valeur (notée AV) est ajustée uniquement en fonction des rachats de

contrats. L'approche naturelle de gestion du risque de rachat pour les assureurs est de définir un risque de base, qu'ils ajustent dynamiquement ensuite. Le problème justement est que cet ajustement se fait sur des hypothèses d'indépendance des comportements, ce que nous voulons précisément éviter. Considérons donc le modèle simplifié mais réaliste suivant :

$$AV_{adj}(t) = AV_{adj}(t-1) \times \left[1 + \%\text{ periodical benefit} \times (1 - \text{surrender rate})\right]$$

Nous pouvons ainsi estimer la sensibilité de la valeur du compte au risque de comportement en utilisant des simulations de Monte Carlo, dont certains scénarios seront des scénarios de chocs. Cette méthodologie est d'ailleurs une des recommandations de Solvabilité II, qui préconise le développement d'un modèle interne et l'utilisation de multiples scénarios pour évaluer des marges de risque robustes (ou encore le *capital économiqie* noté EC). L'EC est la différence entre la VaR et le *best estimate* (BE) (le BE est le risque moyen attendu). La distribution bi-modale du taux de rachat implique des changements significatifs dans ces estimations, et donc dans le provisionnement : nous estimons d'abord la valeur du compte avec des hypothèses "best estimate", puis avec la $VaR_{99.5\%}$ sous l'hypothèse de normalité (comportements indépendants) et sous l'hypothèse bi-modale (comportements corrélés). La différence de capital économique entre ces deux situations représente à ce titre l'écart de réserves en termes de taux de rachat pour l'assureur.

Application numérique L'introduction de la dépendance entre les comportements se fait via le modèle présenté dans la première section. Nous formulons les hypothèses suivantes :

- l'horizon est fixé à un an ;
- les conditions économiques et financières se détériorent dans la période ;
- aujourd'hui, la valeur AV du compte est de 1 000 000 US$;
- le taux de rentabilité annuel est de 10% ;
- la VaR, l'EC et la AV sont calculés sur l'année à venir.

Les valeurs de ces paramètres sont censés refléter la réalité, sans pour autant trahir la confidentialité de nos données. Le portefeuille comporte 17 657 assurés et le nombre de simulations est fixé à 100 000. Le tableau 2.3 met en évidence l'impact potentiel de la modélisation des rachats

Scénario	BE	EC / AV($VaR^{Normal}_{99.5\%}$)	Corrélation	EC ($VaR^{Bi-modal}_{99.5\%}$)	ΔEC	ΔAV
Soft ($p = 8.08\%$)	8.06% ($p_0 = 0$)	0.84% / 1 091 029 ($p_0 = 0$)	$p_0 = 0.05$	5.45%	4.61%	5506.7
			$p_0 = 0.15$	14.85%	14.01%	14857.1
			$p_0 = 0.3$	28.7%	27.86%	28743.9
			$p_0 = 0.5$	47.26%	46.42%	47189.9
Medium ($p = 20\%$)	20.03% ($p_0 = 0$)	1.18% / 1 078 824 ($p_0 = 0$)	$p_0 = 0.05$	5.06%	3.88%	5179.8
			$p_0 = 0.15$	13.15%	11.97%	13182.3
			$p_0 = 0.3$	25.35%	24.17%	25290.8
			$p_0 = 0.5$	41.3%	40.12%	41233.5
Hard ($p = 42\%$)	42.02% ($p_0 = 0$)	1.48% / 1 056 459 ($p_0 = 0$)	$p_0 = 0.05$	4.28%	2.8%	4253.6
			$p_0 = 0.15$	10.13%	8.65%	10036
			$p_0 = 0.3$	18.76%	17.28%	18769.1
			$p_0 = 0.5$	30.21%	28.73%	30260.3

TABLE 2.3 – Estimations des différences de *capital économique* (en termes de taux de rachat) et valeur du compte pour différentes coorélations dans divers scénarios.

TABLE 2.4 – Effet de la taille du portefeuille sur la valeur du compte dans un contexte **soft** (100 000 simulations).

Portfolio size	BE	EC / AV($VaR^{Normal}_{99.5\%}$)	Correlation	EC ($VaR^{Bi-modal}_{99.5\%}$)	ΔEC	ΔAV
Little :			$p_0 = 0.05$	6.26%	4.5%	6280
5000	8.1%	1.76% / 1 090 140	$p_0 = 0.2$	20.42%	18.66%	20440
policyholders	($p_0 = 0$)	($p_0 = 0$)	$p_0 = 0.5$	48.54%	46.78%	48560
Medium :			$p_0 = 0.05$	5.1%	4.59%	5102
50 000	8.08%	0.51% / 1 091 408	$p_0 = 0.2$	19.01%	18.5%	19012
policyholders	($p_0 = 0$)	($p_0 = 0$)	$p_0 = 0.5$	46.63%	46.12%	46634
Big :			$p_0 = 0.05$	4.73%	4.59%	4731.2
500 000	8.08%	0.1426% / 1 091 777	$p_0 = 0.2$	18.56%	18.41%	18565
policyholders	($p_0 = 0$)	($p_0 = 0$)	$p_0 = 0.5$	46.16%	46.01%	46158.6

sur les estimations de réserves. Il résume les différences de provisionnement en $ pour divers contextes économiques et niveaux de corrélation. Ces résultats montrent que même dans un contexte sain, des comportements corrélés à 50% peuvent engendrer une perte surprise de 47 190$ à la fin de l'année par rapport à une hypothèse de comportements indépendants (4.72% de la richesse initiale) ! La différence entre le taux moyen de rachat espéré et le taux stressé est de 47,26% !
De plus, la valeur initiale du compte a été fixée à 1 000 000$, ce qui est peu comparé aux capitaux dont disposent les compagnies d'assurance en règle générale. Nous pouvons donc imaginer quelle serait la véritable perte dans une situation plus réaliste.
La taille du portefeuille joue souvent un rôle prépondérant lors de l'évaluation des risques (en faisant chuter par exemple la variance), ce qui rend intéressant le fait d'étudier les pertes en fonction de ce paramètre. Pour cela, nous extrayons la partie "contexte économique soft" du tableau 2.3. Au départ, rappelons que nous avions 17 657 assurés. Une idée reçue consiste à penser que parce que notre portefeuille est gros, nous sommes prémunis contre des scénarios extrêmes grâce aux effets de mutualisation. L'analyse du tableau 2.4 démontre que le nombre d'assurés n'a que peu d'influence sur le calcul de la marge de risque, surtout si les comportements sont très corrélés. Nous appelons ce type de risque un risque non diversifiable. Ainsi, mettre assez d'argent de côté pour prévenir le risque de rachat est très important dans la prévision des besoins en capitaux, quelle que soit la taille du portefeuille : la sous-estimation de ce risque pourrait provoquer des appels de marge aux actionnaires, et les rendre ainsi très mécontents.

2.5 Conclusion

Nous avons montré dans cette partie que l'impact du choix de la distribution des rachats est majeur tant pour les calculs de besoin en capital économique que pour les prévisions de taux de rachat. Cette distribution résulte de deux points de vue opposés sur la modélisation des comportements : de comportements supposés indépendants, nous évoluons vers une modélisation de comportements de rachat corrélés qui selon nous reflète davantage la réalité, comme l'illustre le graphique 2.1. Nous pourrions aussi investiguer l'impact rétroactif des rachats massifs sur les taux d'intérêt et d'inflation pour éventuellement détecter un cercle vicieux, ou

encore considérer les rachats partiels provenant d'options contenues dans les contrats qui permettent de basculer une partie de son épargne vers un autre type de support (ex : transférer de l'UC vers un fonds Euro, Loisel (2010)). Cette dernière information serait idéale mais n'est concrètement jamais accessible, car elle nécessite la récupération de données anciennes auprès d'autres services où le personnel a souvent changé.
En résumé, provisionner assez d'argent dans le but de couvrir le risque de corrélation des comportements est très important dans la prévention des besoins en capitaux. En effet il pourrait découler de la sous-estimation de ce risque des appels de marge aux actionnaires, ce qui serait très néfaste pour la compagnie. De plus, le caractère non-diversifiable de certains risques fait que la taille du portefeuille ne permet pas de réduire l'impact des crises de corrélation sur les quantités considérées.

L'approche théorique *supra* nous conduit naturellement à l'extension développée dans le chapitre suivant : les mélanges de régressions logistiques, permettant de capter à la fois les phénomènes de corrélation et l'hétérogénéité des comportements entre cohortes.

Bibliographie

Biard, R., Lefèvre, C. and Loisel, S. (2008), 'Impact of correlation crises in risk theory : Asymptotics of finite-time ruin probabilities for heavy-tailed claim amounts when some independence and stationarity assumptions are relaxed', *Insurance : Mathematics and Economics* **43**(3), 412 – 421. 48

Denuit, M., Lefèvre, C. and Shaked, M. (1998), 'The s-convex orders among real random variables, with applications', *Math. Inequal. Appl.* **1**(4), 585–613. 55

Lefèvre, C. and Utev, S. (1996), 'Comparing sums of exchangeable Bernoulli random variables', *J. Appl. Probab.* **33**(2), 285–310. 55

Loisel, S. (2008), 'From liquidity crisis to correlation crisis, and the need for quanls in enterprise risk management', pp. 75–77. in Risk Management : The Current Financial Crisis, Lessons Learned and Future Implications, Edited by the SOA, CAS and CIA. 48

Loisel, S. (2010), 'Contribution à la gestion quantitative des risques en assurance', *Habilitation Thesis, Université Lyon 1* . 65

Loisel, S. and Milhaud, X. (2011), 'From deterministic to stochastic surrender risk models : Impact of correlation crises on economic capital', *European Journal of Operational Research* **214**(2). 45, 58

Marshall, A. and Olkin, I. (1979), *Inequalities : Theory of Majorization and Its Applications*, Academic Press, New York. 56

Martinussen, T. and Scheike, T. (2006), *Dynamic Regression Models for Survival Data*, Springer. 48

McNeil, A., Frey, R. and Embrechts, P. (2005), *Quantitative Risk Management*, Princeton Series In Finance. 50

Planchet, F. and Thérond, P. (2006), *Modèles de durée : applications actuarielles*, Economica (Paris). 48

Ramsay, J., Hooker, G. and Graves, S. (2009), *Functional Data Analysis with R and MATLAB*, Springer. 48

Ramsay, J. and Silverman, B. (2005), *Functional Data Analysis, Second Edition*, Spinger, Springer Series in Statistics. 48

Viquerat, S. (2010), On the efficiency of recursive evaluations in relation to risk theory applications, PhD thesis. 52

Deuxième partie

Vers la création de groupes comportementaux “dynamiques”

Chapitre 3

Mélange de régressions logistiques

Ce chapitre correspond à l'article "*Exogenous and endogenous risk factors management to predict surrender behaviours*" (X. Milhaud), actuellement soumis pour publication.

Nous avons mis en évidence toute la difficulté d'étudier et de modéliser le risque de rachat, notamment car c'est un risque de comportement humain qui dépend de nombreux facteurs : les caractéristiques individuelles, les désirs et besoins personnels, les options du contrat, son ancienneté, le contexte économique et financier, les aspects socio-culturels (ex : comparaison expérience Japon / Etats-Unis), mais aussi les décisions du régulateur.
Les deux chapitres précédents nous ont permis de pointer du doigt les trois problématiques majeures : la dimension des données, les problèmes de corrélation entre comportements et l'hétérogénéité des décisions face à un environnement évolutif. Les rachats s'expliquent à la fois par des caractéristiques idiosynchratiques mais aussi par un ensemble de facteurs exogènes dont certains sont très difficilement quantifiables (réputation), voire inaccessible (politique de vente future). Une méthodologie quant à la réduction de la dimension des données se détache par l'utilisation des algorithmes CART qui ont l'avantage de ne pas supposer d'hypothèses fortes sous-jacentes aux données (relation linéaire, log-linéaire,...), et qui ont démontré leur robustesse. Les quelques variables sélectionnées (nous nous limitons à deux ou trois variables en général car un modèle surdimensionné donne souvent de mauvaises prévisions) présentant le plus fort impact sur l'événement de rachat seront ensuite introduites dans des modèles dynamiques. Nous avons également compris que l'association de facteurs exogènes et endogènes dans une même et unique équation de régression ne permet pas de capter les bons effets, le défi étant donc de trouver une manière fonctionnelle de considérer ces effets différemment. Ce chapitre propose donc une extension des premières modélisations évoquées, en associant les idées que nous avons développé jusqu'à présent pour parvenir à nos fins.

Les modèles mélange sont une technique populaire pour prendre en compte l'hétérogénéité non-observable de données, ou pour approximer une distribution générale de manière semi-paramétrique. Ils sont utilisés dans de nombreux domaines d'application tels que l'économie, l'astronomie, la biologie, la médecine. Historiquement, les mélanges ont été introduit pour la première fois il y a plus de cent ans par Pearson (1894). Karl Pearson utilise un mélange de lois normales dans le cadre de la modélisation de la longueur du corps des crabes. La modélisation de données asymétriques est aussi réalisable via des transformations de données, notamment la transformation en *log* (Box and Cox (1964)). Il est souvent difficile de faire un distinguo entre des données présentant une certaine asymétrie et des données provenant

d'un mélange (McLachlan and Peel (2000), p.15), bien que dans notre cas l'asymétrie est telle qu'il ne fait aucun doute qu'une simple transformation ne pourrait pas modéliser toute l'hétérogénéité présente dans nos données. Le cas le plus classique de la modélisation par mélange concerne les mélanges de lois normales, pour lesquels un grand nombre de résultats existe. Les modèles mélanges sont aussi régulièrement utilisés dans des problèmatiques de classification non-supervisée puisqu'ils assignent un composante donnée à chaque observation, ce qui constitue un clustering en soi.

La première partie développe l'aspect théorique des modèles mélange, les points importants à aborder lors de leur utilisation et les pièges à éviter. En ce qui concerne l'application, nous verrons ensuite les outils pratiques de présentation des résultats de la modélisation à travers un cas pratique (toujours basé sur les contrats mixtes en Espagne).

3.1 Formalisation de la théorie

La modélisation par mélange renvoie aux problèmes usuels suivants : identifiabilité, estimation des paramètres, propriétés de l'estimateur du maximum de vraisemblance, évaluation du nombre de composantes du mélange, application de la théorie asymptotique pour fournir une base de solutions à certains problèmes, critères de sélection et de performance du modèle. L'estimation des paramètres d'un mélange est un des axes de recherche ayant attiré le plus de chercheurs car de nombreuses questions subsistent encore aujourd'hui ; parmi lesquelles les valeurs initiales de l'algorithme d'optimisation qui maximise la vraisemblance, les critères d'arrêt de cet algorithme et les propriétés de la fonction de vraisemblance (convexité, bornitude). Nous allons dans cette partie tenter de résumer l'ensemble de ces problématiques afin de donner au lecteur une base théorique qui lui permette d'appréhender ce type de modélisation, qui servira de socle dans toute la suite de la thèse.

3.1.1 Généralités

Nous formalisons l'approche par mélange dans le cadre d'un mélange discret car elle correspond à notre cas d'étude opérationnelle, et demeure de plus bien plus intuitive lorsque nous nous intéressons à des questions de classification. Néanmoins, toutes les notions développées ci-dessous peuvent être adaptées au cas continu, ce qui veut dire que la distribution mélangeante est continue (nous verrons qu'elle est multinomiale dans le cas discret).

Soit $Y = (Y_1^T, ..., Y_n^T)^T$ un échantillon aléatoire indépendant et identiquement distribué (i.i.d.). Chaque enregistrement Y_j de cet échantillon contient q mesures, d'où un vecteur aléatoire q-dimensionnel ($q = 1$ pour nous car la décision de rachat est univariée). Dans le contexte des mélanges et par la formule des probabilités totales, il vient

$$f(y_j) = \sum_{i=1}^{G} \pi_i f_i(y_j), \tag{3.1}$$

où $f(y_j)$ est la densité de Y_j dans $\mathbb{R}^q$, π_i est la proportion (poids) **a priori** de la i^e composante du mélange, $f_i(y_j)$ est la densité de la i^e composante du mélange, avec la contrainte $\sum_i \pi_i = 1$. La matrice Y des observations est de taille $n \times q$. On dit que $f(y_j)$ est la densité d'un mélange fini à G composantes, et on note $F(y_j)$ la distribution du mélange. Chaque individu est donc censé provenir d'un des groupes composant le mélange.

En général, G est fini mais inconnu et doit donc être estimé inférentiellement à partir des données. Les probabilités d'appartenance à tel ou tel groupe doivent être estimées en même temps, de même que les densités $f_i(.)$ de chaque composante. Pour comprendre l'interprétation de la modélisation par mélange, une bonne méthode consiste à essayer de le générer. Pour simuler la variable Y_j, nous définissons la variable Z_j d'appartenance à une composante par

$$Z_j = \begin{cases} 1 & \text{avec probabilité } \pi_1 \text{ si l'individu j appartient au groupe 1,} \\ 2 & \text{avec probabilité } \pi_2 \text{ si l'individu j appartient au groupe 2,} \\ ... & \\ G & \text{avec probabilité } \pi_G \text{ si l'individu j appartient au groupe G,} \end{cases}$$

et la densité conditionnelle de Y_j est donnée par $f_{Y_j|Z_j=i}(y_j) = f_i(y_j)$. Nous pouvons donc voir Z_j comme le vecteur aléatoire $Z_j = (Z_{j1}, Z_{j2}, ..., Z_{jG})^T$ où

$$Z_{ij} = (Z_j)_i = \begin{cases} 1 & \text{si la composante d'appartenance de } Y_j \text{ dans le mélange est la } i^e, \\ 0 & \text{sinon.} \end{cases}$$

Ainsi, Z_j suit une loi multinomiale et l'on note $Z_j \sim Mult_G(1, \pi)$ avec $\pi = (\pi_1, ..., \pi_G)^T$. Nous avons donc

$$P(Z_j = z_j) = \pi_1^{z_{1j}} ... \pi_G^{z_{Gj}}.$$

Les mélanges peuvent être vus comme une alternative entre un modèle complètement paramétrique et un modèle non-paramétrique. Dans le cas non-paramétrique, nous retrouvons l'estimateur à noyau de la densité en prenant $G = n$ composantes (où n est le nombre d'observations), des poids tous égaux $\pi = 1/n$ et $f_i(y_j) = \frac{1}{h} k(\frac{y_j - y_i}{h})$ où $k(.)$ est une densité. A l'inverse, si l'on fixe $G = 1$ composante, alors le modèle devient complètement paramétrique. Nous nous intéressons dans la suite aux cas où $G \in [\![1; n]\!]$.

Nous l'avons dit en introduction : la multimodalité des données peut ne pas provenir d'un mélange. Il est possible de détecter ceci par l'usage du test du ratio de vraisemblance, mais la difficulté vient du fait que nous ne connaissons pas la distribution de la statistique de test sous l'hypothèse nulle dans ce cadre-là. Nous utilisons alors une approche de reéchantillonage qui permet d'obtenir une *p-valeur* de test sans connaître cette statistique (McLachlan and Peel (2000), p.75). La clef pour l'estimation des paramètres d'un mélange est de reformaliser le problème de données incomplètes sous forme d'un problème aux données complètes : en effet, nous ne connaissons pas le groupe d'appartenance de chaque observation dans la réalité, mais l'introduction de la variable Z_j va nous permettre de mener directement l'estimation par maximum de vraisemblance par l'algorithme espérance-maximisation (EM). Dans un contexte bayésien (le nôtre est fréquentiste), cette vision du problème permet d'estimer les paramètres par des méthodes de type MCMC (Monte Carlo Markov Chain).

En résumé, nous observons $y = (y_1, ..., y_n)$, réalisations de $Y = (Y_1, ..., Y_n)$ issues de la même densité mélange donnée par (3.1). Ces observations sont i.i.d. et nous avons

$$Y_1, ..., Y_n \sim F = F(Y_j).$$

Les données complètes, notées y_c, s'exprimeraient donc comme $y_c = \begin{pmatrix} (y_1, z_1) \\ (y_2, z_2) \\ ... \\ (y_n, z_n) \end{pmatrix}$.

Grâce à la formule de Bayes, nous pouvons calculer la probabilité **a posteriori** d'appartenir à telle ou telle composante du mélange :

$$\tau_i(y_j) = \pi_i \frac{f_i(y_j)}{f(y_j)},$$

pour $i = 1, .., G$ et $j = 1, ..., n$.
En pratique nous estimons les π_i par leur moyenne empirique (sauf au départ de l'algorithme où ils sont fixés arbitrairement, cf section 3.1.3), i.e. $\hat{\pi}_i = \sum_{j=1}^{n} z_{ij}/n$ où les z_{ij} sont déterminés par la règle du maximum a posteriori ; et les paramètres des composantes du mélange grâce aux données qui y appartiennent.
Une formulation paramétrique d'un modèle mélange peut s'écrire de la manière suivante

$$f(y_j) = f(y_j; \psi) = \sum_{i=1}^{G} \pi_i f_i(y_j; \theta_i), \tag{3.2}$$

avec $\psi = (\pi_1, ..., \pi_{G-1}, \xi^T)^T$ et $\xi^T = (\theta_1^T, ..., \theta_G^T)$. Nous noterons Ψ l'espace des paramètres de ψ. Faisons l'hypothèse que les composantes appartiennent à la même famille paramétrique (mais ce n'est pas une généralité), et considérons une distribution mélangeante discrète $H(\theta)$ définie par $H(\theta) = P(\theta = \theta_i) = \pi_i$ pour i=1,...,G. Alors le modèle mélange se reécrit comme

$$f(y_j; H) = \int f(y_j; \theta) dH(\theta).$$

Cette généralisation de l'écriture du mélange permet d'appréhender son expression avec une mesure de probabilité plus générale pour H, qui peut être une loi continue (souvent une loi Gamma (modèle Poisson-Gamma) ou une loi Beta (modèle Beta-Binomial)).

Il existe dans la littérature plusieurs techniques d'estimation de la distribution mélange : la méthode graphique, la méthode des moments, la méthode des distances minimum, l'approche bayésienne et le maximum de vraisemblance. Cette grande variété est due au fait que nous n'avons pas de formules explicites pour les estimateurs, qui sont calculés itérativement par divers algorithmes. D'autre part, la taille n de l'échantillon doit être relativement grande pour garantir les propriétés asymptotiques des mélanges.

3.1.2 Identifiabilité

L'estimation de ψ sur la base des observations y_j n'a de sens que si ψ est identifiable. La définition de l'identifiabilité dans le cadre des mélanges diffère un peu du cas classique dans la mesure où il y a la notion supplémentaire de composantes. Intuitivement, un modèle est identifiable si des valeurs distinctes de ψ déterminent des membres distincts de la famille paramétrique associée à ψ (il ne peut pas y avoir deux paramètres différents qui donnent le même modèle à l'arrivée). Formellement, $f(y_j; \psi)$ est une famille paramétrique de densité identifiable dans le cadre classique si

$$f(y_j; \psi) = f(y_j; \psi') \Leftrightarrow \psi = \psi'$$

Pour les mélanges, cette seule définition ne suffit pas car on peut avoir une classe de mélange identifiable sans pour autant avoir identifiabilité sur ψ. Il suffit pour comprendre cela de permuter les composantes d'appartenance (les "labels"), ce qui ne change pas la densité globale

du mélange mais qui rend ψ non-identifiable pour des densités composantes appartenant à la même famille paramétrique. Il faut donc ajouter une contrainte supplémentaire.
Soit $f(y_j;\psi) = \sum_{i=1}^{G} \pi_i f_i(y_j;\theta_i)$ et $f(y_j;\psi^\star) = \sum_{i=1}^{G} \pi_i^\star f_i(y_j;\theta_i^\star)$ deux membres d'une famille paramétrique de mélange. Cette classe de mélanges finis est dite **identifiable** pour $\psi \in \Psi$ si

$$f(y_j;\psi) = f(y_j;\psi^\star)$$
$$\Updownarrow$$
$G = G^\star$ et on peut permuter les indicatrices de composantes d'appartenance pour que $\pi_i = \pi_i^\star$ et $f_i(y_j;\theta_i) = f_i(y_j;\theta_i^\star)$.

En général, nous ajoutons des contraintes pour palier au manque d'identifiabilité dû aux permutations possibles entre composantes d'appartenance. Un détail important est à ajouter : le manque d'identifiabilité est un problème important en analyse bayésienne des mélanges lors de la simulation a posteriori de l'appartenance à un groupe donné, mais n'est pas préoccupant dans le cadre de l'estimation par maximum de vraisemblance.

En dehors de l'identifiabilité, un autre problème à ne pas confondre est celui de l'identification : est-ce facile de savoir à quelle composante appartient une observation donnée ? La réponse dépend évidemment de la répartition des données. Une multimodalité prononcée sera moins problématique que des données faiblement asymétriques, mais nous reviendrons justement sur ce point dans le dernier chapitre.

3.1.3 Algorithme espérance-maximisation (EM)

Cet algorithme offre des propriétés très intéressantes pour l'optimisation de fonction de vraisemblance complexe, sur un problème aux données manquantes. Ces propriétés ont été démontrées dans un article célèbre de Dempster et al. (1977), qui a permis avec la révolution informatique l'explosion de l'usage de ce type de modèle, qui jusque là demandait de complexes et fastidieux calculs pour maximiser la vraisemblance. Nous donnons ici la version originelle de cet algorithme et son idée, sachant qu'une multitude de développement ont depuis été proposés pour traiter des problématiques particulières (convergence vers le maximum global, dimension des données, y_j manquantes, ...).

Le principe de base de cet algorithme est de transformer le problème aux données manquantes en problème aux données complètes $Y_c = (Y^T, Z^T)^T$ où les $Z_j \sim Mult_G(1,\pi)$ et sont i.i.d. La vraisemblance des données complètes pour une observation j vaut

$$f(y_{jc};\psi) = \Pi_{i=1}^{G}[\pi_i f_i(y_j;\theta_i)]^{z_{ij}},$$

d'où la log-vraisemblance des donnés complètes sur l'échantillon entier $\log L_c(\psi;y) = \log(\Pi_{j=1}^{n} f(y_{jc};\psi))$ qui donne après développement :

$$\log L_c(\psi;y) = \sum_{j=1}^{n}\sum_{i=1}^{G} z_{ij}[\log \pi_i + \log f_i(y_j;\theta_i)]. \tag{3.3}$$

Etape espérance Traitement de la donnée z_j non observable par l'espérance conditionnelle de $\log L_c(\psi; y)$, sachant ce que nous observons y. Soit $\psi(0)$ la valeur initiale de ψ. Nous calculons

$$Q(\psi, \psi^{(0)}) = \mathbb{E}_{\psi^{(0)}}[\log L_c(\psi) \mid y],$$

or $\log L_c(\psi)$ est linéaire en z_{ij}, donc l'étape espérance requiert uniquement le calcul de $\mathbb{E}[Z_{ij} \mid y]$. Nous avons

$$\mathbb{E}_{\psi^{(k)}}[Z_{ij} \mid y] = P_{\psi^{(k)}}(Z_{ij} = 1 \mid y) = \tau_i(y_j; \psi^{(k)}), \tag{3.4}$$

avec τ_i la probabilité a posteriori à l'étape k d'appartenir à la composante i. En injectant (3.4) dans (3.3), nous pouvons calculer à l'étape $(k+1)$ l'expression de Q :

$$\begin{aligned} Q(\psi, \psi^{(k)}) &= \sum_{j=1}^{n}\sum_{i=1}^{G} \mathbb{E}[Z_{ij} \mid y] \times [\log \pi_i + \log f_i(y_j; \theta_i)], \\ Q(\psi, \psi^{(k)}) &= \sum_{j=1}^{n}\sum_{i=1}^{G} \tau_i(y_j; \psi^{(k)}) \, [\log \pi_i + \log f_i(y_j; \theta_i)], \end{aligned} \tag{3.5}$$

avec grâce à (3.1.1),

$$\tau_i(y_j; \psi^{(k)}) = \pi_i^{(k)} \frac{f_i(y_j; \theta_i^{(k)})}{f(y_j; \psi^{(k)})} = \pi_i^{(k)} \frac{f_i(y_j; \theta_i^{(k)})}{\sum_{h=1}^{G} \pi_h^{(k)} f_h(y_j; \theta_h^{(k)})}.$$

Etape maximisation A l'étape $k+1$, nous voulons maximiser globalement $Q(\psi, \psi^{(k)})$ par rapport à ψ sur Ψ, pour donner une estimation $\psi^{(k+1)}$. Dans le cas des mélanges finis, nous estimons séparément les proportions et les densités composantes. Il vient :

$$\begin{cases} \pi_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \tau_i(y_j; \psi^{(k)}) : \text{moyenne empirique des probabilités a posteriori,} \\ \xi^{(k+1)} \text{ est obtenu en résolvant } \sum_{j=1}^{n} \sum_{i=1}^{G} \tau_i(y_j; \psi^{(k)}) \dfrac{\delta \log f_i(y_j; \theta_i)}{\delta \xi} = 0. \end{cases}$$

La propriété de d'accroissement monotone de la vraisemblance des données complètes $Q(\psi, \psi^{(k)})$ à chaque étape garantit la convergence de la vraisemblance vers une valeur stationnaire (maximum local ou global). En effet, en notant L la vraisemblance d'un modèle mélange, on a :

Theorem 3.1.1. *(Propriété fondamentale de l'algorithme EM).*
$\forall \psi, \psi^{'} \in \Psi$

$$Q(\psi^{'}, \psi) \geq Q(\psi, \psi) \Rightarrow L(\psi^{'}) \geq L(\psi),$$

avec égalité si et seulement si $Q(\psi^{'}, \psi) = Q(\psi, \psi)$ *et* $\tau_i(y; \psi) = \tau_i(y; \psi^{'})$ *pour tout* i *et presque tout* x.

La log-vraisemblance observée étant la log-vraisemblance des données complètes moins un terme toujours négatif d'après l'inégalité de Jensen, le même résultat de convergence est applicable pour les données observées.
Nous répétons successivement ces étapes jusqu'à ce que le critère d'arrêt de l'algorithme soit

satisfait, en général $L_c(\psi^{(k+1)}) - L_c(\psi^{(k)})$ plus petit qu'un certain seuil. Toutefois cette procédure d'arrêt n'est pas toujours satisfaisante, c'est pourquoi Lindstrom and Bates (1988) et Bohning et al. (1994) proposent une amélioration basée sur le critère d'accélération de Aitken. Pour détecter la pertinence des estimations trouvées, il n'existe pas de méthode prédéfinie : la solution consiste à regarder à la fois la valeur de la vraisemblance, les valeurs des π_i estimées et les matrices de covariance (voir les exemples p.100 de McLachlan and Peel (2000)).
La vitesse de convergence de l'algorithme EM dépend de la proportion d'information manquante sur ψ, du fait que l'on observe seulement les réalisations de Y au lieu d'observer conjointement Y et Z. Plus cette proportion est grande, plus l'algorithme est lent. Nous ne discutons pas des variantes de l'EM qui permettent de contourner les problèmes de valeurs initiales de l'algorithme, mais le lecteur intéressé trouvera des références intéressantes dans McLachlan and Peel (2000).

3.1.4 Evaluation du nombre de composantes

L'évaluation du bon nombre de composantes d'un modèle mélange a toujours été difficile et le problème n'est pas encore vraiment résolu. Les mélanges ont principalement deux fonctions : fournir une classification basée sur une modélisation ; et définir une méthode semi-paramétrique permettant de modéliser des formes de distribution inconnues, comme une alternative à la méthode des noyaux. Mais dans ces deux approches, comment choisir G ?
Nous avons pu constater la séparation du problème de l'évaluation de G et celui de l'estimation des paramètres, dans le sens où l'on fixe d'abord G avant de lancer l'estimation. Nous faisons cela pour plusieurs valeurs de G. L'usage commun pour trouver G est :

- de considérer des critères de sélection tels que le critère d'information de Akaike (AIC) ou le Bayesian Information Criterion (BIC),
- de se servir du test du ratio de vraisemblance (LRT),

mais il existe aussi des techniques non-paramétriques, ou encore la méthode des moments, l'approche basée sur le Kurtosis de la distribution... Les références à toutes ces techniques sont disponibles dans l'ouvrage de McLachlan and Peel (2000). Nous ne discutons pas davantage des critères AIC et BIC car leur présentation exhaustive suivra dans le chapitre 4.

En revanche nous souhaitons préciser (très succintement) en quoi consiste le LRT dans le cadre des mélanges, sans pour autant entrer dans trop de détails. Ce test a pour but de trouver la plus petite valeur convenable de G, avec comme hypothèses nulle et alternative :

$$[\quad H_0 : G = G_0 \quad \text{contre} \quad H_1 : G = G_1 \quad] \text{, avec } G_1 > G_0.$$

En pratique nous prenons $G_1 = G_0 + 1$ et nous continuons d'ajouter des composantes tant que l'accroissement de la valeur de la vraisemblance est substantiel. Soient $\hat{\psi}_1$ l'estimateur par maximum de vraisemblance (MLE) de ψ sous H_1, et $\hat{\psi}_0$ le MLE sous H_0. Nous notons

$$-2\log\lambda = 2[\log L(\hat{\psi}_1) - \log(\hat{\psi}_0)] = 2\log\frac{\hat{\psi}_1}{\hat{\psi}_0}.$$

Si λ est suffisamment petit, ou si $-2\log\lambda$ est suffisamment grand, il parait logique de pouvoir rejeter H_0. Malheureusement ici, nous ne connaissons pas la distribution nulle de $-2\log\lambda$ dans le cas général, car les conditions de régularité nécessaires (Cramér (1946)) aux résultats asymptotiques du MLE ne sont pas satisfaites (voir Ghosh and Sen (1985)). Les travaux pionniers de Wolfe (1971) justifient par exemple l'usage de la simulation pour calculer la *p-valeur* de ce test. Plus globalement, ce problème complexe nécessite bien plus de détails :

le lecteur intéressé pourra alors consulter les travaux de Bohning and Seidel (2003) et Garel (2007), dans lesquels des références et des études théoriques (avec application) sont proposées.

3.1.5 Focus sur les mélanges de Logit dans le contexte des rachats

Les mélanges de régressions logistiques font partie intégrante des modèles mélanges semi-paramétriques. Un résumé et une revue bibliographique de ce type de modèles est disponible dans le papier de Lindsay and Lesperance (1995), et des applications dans le domaine de la biologie sont fournies dans Follmann and Lambert (1989) et Wang (1994). Nous avons vu au chapitre 1 que la régression logistique était adaptée aux données binomialement distribuées. Ainsi en notant $p_i(X_j)$ la probabilité de rachat de N_j individus homogènes (ayant les mêmes caractéristiques X_j) appartenant à la composante i du mélange, et en reprenant les notations ci-dessus avec $Y_j \sim Bin(N_j, p_i(X_j))$:

$$f_i(y_j) = f(y_j; p_i(X_j)) = P(Y_j = y_j) = C_{N_j}^{y_j} p_i(X_j)^{y_j}(1 - p_i(X_j))^{N_j - y_j}, \tag{3.6}$$

où y_j est le nombre de rachats observés dans le groupe homogène j, et $p_i(X_j)$ résulte du lien logistique

$$p_i(X_j) = \frac{\exp(\beta_i^T X_j)}{1 + \exp(\beta_i^T X_j)},$$

avec $X_j = (X_{j1}, ..., X_{jp})^T$ le vecteur des p covariables de l'individu j, $\beta_i = (\beta_{i1}, ..., \beta_{ip})^T$ le vecteur des p coefficients de régression de la composante i.
Considérons la moyenne et la variance du modèle mélange de régressions logistiques tel que

$$f(y_j; \psi) = \sum_{i=1}^{G} \pi_i(X_j^{'}) f(y_j; p_i(X_j)).$$

L'interprétation de ce modèle est la suivante : il existe plusieurs groupes qui suivent des distributions logistiques différentes, avec chacun une proportion $\pi_i(X_j^{'})$. Cette proportion peut donc dépendre de certaines variables explicatives, à condition que l'identifiabilité soit préservée (cf ci-dessous). La moyenne et la variance sont facilement calculables par les formules suivantes :

$$\begin{aligned}
\mathbb{E}[Y_j] &= \mathbb{E}_Z\left[\mathbb{E}[Y_j \mid Z_j]\right] = \sum_{i=1}^{G} P(Z_{ij} = 1)\mathbb{E}[Y_j \mid Z_j] = \sum_{i=1}^{G} \pi_{ij} p_{ij}, \\
\mathbb{V}ar[Y_j] &= \mathbb{E}\left[\mathbb{V}ar[Y_j|Z_j]\right] + \mathbb{V}ar\left[\mathbb{E}[Y_j|Z_j]\right] \\
&= N_j \left[\sum_{i=1}^{G} \pi_{ij} p_{ij}\right]\left[1 - \sum_{i=1}^{G} \pi_{ij} p_{ij}\right] + \frac{(N_j - 1)}{N_j} \mathbb{V}ar\left[\mathbb{E}[Y_j|Z_j]\right],
\end{aligned}$$

avec $\mathbb{V}ar\left[\mathbb{E}[Y_j|Z_j]\right] = N_j^2 \left[\sum_{i=1}^{G} \pi_{ij} p_{ij}^2 - (\sum_{i=1}^{G} \pi_{ij} p_{ij})^2\right]$.

Dans le cas général, nous considérons également des poids $\pi_i(X_j^{'})$ qui dépendent de certains facteurs endogènes ou exogènes, ce qui nous amène à les définir aussi comme des régressions logistiques multinomiales (rappelons que Z_j est multinomiale), soit :

$$\pi_i(X_j^{'}) = \frac{\exp(\gamma_i^T X_j^{'})}{\sum_{h=1}^{G} \exp(\gamma_h^T X_j^{'})}, \tag{3.7}$$

avec $X_j^{'} = (X_{j1}^{'}, ..., X_{jl}^{'})^T$ un ensemble de l covariables de l'individu j et $\gamma_i = (\gamma_1, ..., \gamma_l)^T$ le vecteur des l coefficients de régression du poids de la composante i. Ainsi le vecteur des paramètres à estimer vaut $\psi = (\gamma_1^T, ..., \gamma_G^T, \beta_1^T, ..., \beta_G^T)^T$. Pour effectuer l'estimation de ψ, il suffit donc d'insérer (3.6) et (3.7) dans les formules de l'algorithme EM qui sont valables en toute généralité. Dans le cadre de mélange de régressions logistiques, le lecteur intéressé par les problèmes d'identifiabilité pourra trouver son bonheur dans les travaux de Margolin et al. (1989) et Teicher (1963), qui donnent des conditions nécessaires et suffisantes pour leur résolution. Nos futurs choix de modélisation satisfont ces conditions.

En ce qui concerne les prévisions de taux de rachat, elles sont calculées par agrégation des décisions individuelles sur chaque pas de temps (les études seront trimestrielles). Ces décisions étant indépendantes, le principe de calcul de l'intervalle de confiance est identique à celui développé dans le chapitre 1. Pour connaître la décision individuelle d'un assuré, nous regardons les probabilités a posteriori d'appartenir à chacune des composantes : selon la règle de Bayes, l'individu appartient à la composante pour laquelle la probabilité $\pi_i(X_j^{'})$ d'appartenance est maximale. Etant donnée cette appartenance, nous calculons ensuite sa probabilité de rachat $p_i(X_j)$ associée à cette composante. Si une variable explicative est incluse dans le calcul des poids des composantes, alors la proportion de cette composante peut évoluer en fonction de l'évolution de cette variable (introduction d'une corrélation temporelle). Dans le cas classique où les poids ne dépendent d'aucune variable endogène ou exogène, un individu appartient à un seul et unique groupe au cours de la vie de son contrat.

3.2 Cas pratique d'utilisation de mélange de Logit

Nous utilisons dans cet exemple les mêmes données et la même méthodologie (leur description ayant déjà été faite) que dans l'analyse dynamique afin de construire le modèle mélange. La période d'apprentissage représente toujours les deux tiers de la période totale, sachant que nous validons le calibrage du modèle sur la période de validation. Nous exposons dans cette partie les différents résultats que nous retourne l'étude par mélange dans ce contexte ; à savoir une estimation des paramètres de chaque densité composante et leur robustesse, une estimation des paramètres de chaque poids, une comparaison du taux de rachat observé avec la projection par le modèle de ce taux sur l'échantillon de validation, et enfin un test de type Kolmogorov qui permet de valider ou non la qualité des prévisions. Nous commenterons ces résultats en y apportant une tentative de justification pratique. L'exposition de l'ensemble des résultats se trouve dans la dernière partie (et les annexes C pour l'analyse préalable), ce qui permettra d'illustrer la trouvaille faite dans le cadre de ce travail sur la manière de prendre en compte les différents facteurs de risque pour diverses grandes familles de produits d'épargne en Assurance-Vie.

Considérons donc les produits mixtes du portefeuille espagnol pour lesquels nous avons déjà tenté une modélisation sans succès, même par l'introduction de variables financières et économiques (graphique 2.1). La popularité des produits mixtes en Espagne n'est plus à démontrer. Comme déjà évoqué dans les chapitres précédents, le contrat mixte est un contrat d'épargne temporaire classique qui a l'avantage de retourner à son bénéficiaire un capital choisi lors de la souscription, et ce quel que soit l'état de l'assuré (en vie ou décédé, d'où l'appellation "mixte" de cette garantie). Ce type de contrat est très répandu sur le marché espagnol où il a rencontré un vif succès, bien que son prix soit plus élevé qu'un pur contrat d'épargne puisque le risque encouru par l'assuré est plus faible. Les variables disponibles et la période d'étude

(1/1/2000 au 31/12/2007) restent inchangées.
La construction des données qui vont nous servir à construire le modèle est identique, hormis le pas de temps qui passe de mensuel à trimestriel pour éviter une trop grande volumétrie de données.

Rappelons les informations dont nous disposons dans la base d'origine : le numéro du produit, la date d'émission, la date de sortie et sa raison (si sortie il y a), l'option de participation aux bénéfices de l'entreprise (PB), la date de naissance de l'assuré, son sexe, sa richesse, la fréquence de la prime, la prime de risque et la prime d'épargne. Les variables que nous pouvons potentiellement insérer dans la modélisation mélange de régressions logistiques sont donc l'ancienneté du contrat (par la variable "duration.range"), la clause de participation aux bénéfices de la compagnie (renommée "PB.guarantee" mais anciennement "contract.type"), la tranche d'âge de souscription (par "underwritingAge.range"), la tranche de richesse de l'assuré (par "fa.range"), la fréquence de la prime (par "premium.frequency"), les valeurs de prime (de risque par "riskPrem.range" et d'épargne par "savingPrem.range"), et les variables d'environnement que sont l'IBEX 35 (indice boursier espagnol) et le taux des obligations d'Etat 10 ans (par "rate10Y"). Nous considérons plus exactement un historique arbitraire de ces variables économiques puisque nous regardons leur valeur à la date de rachat comparée à leur valeur trois mois auparavant (une option de nos programmes permet de modifier ce critère : allonger la période *delta* de regard en arrière, ou considérer la moyenne de cette évolution).

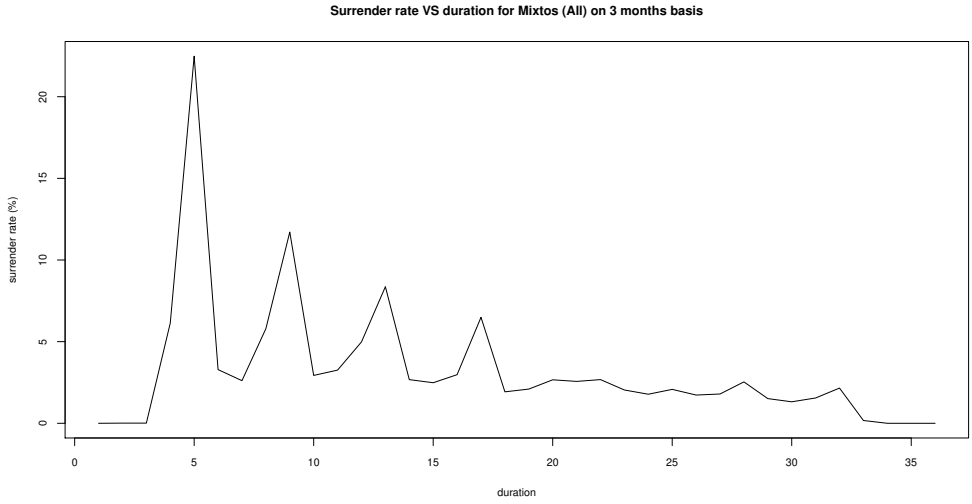
Nous présentons dans la suite quelques statistiques descriptives préalables sur la base de données des contrats mixtes, qui vont nous être fort utiles pour nous guider dans les choix de modélisation.

3.2.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille Nous constatons à travers l'étude du graphique 3.1 que la trajectoire du taux de rachat présente globalement trois phases : un plateau de niveau moyen de rachat bas entre 2000 et 2003, une hausse et une stabilisation entre 2004 et 2006, puis un pic de rachat suivi d'une chute vertigineuse

FIGURE 3.1 – Exposition et taux de rachat trimestriel du portefeuille de produits Mixtes.

FIGURE 3.2 – Rachat par ancienneté de contrat (en trimestre) pour les produits Mixtes.



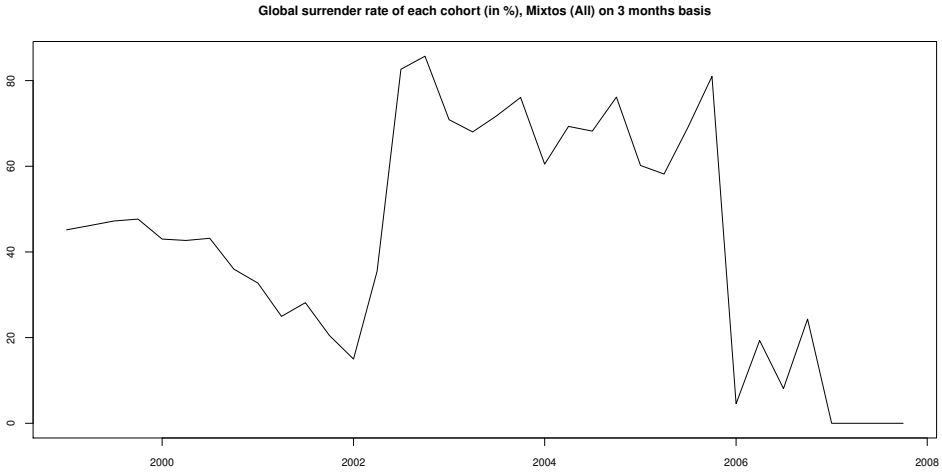
du taux dans l'année 2007 malgré une exposition encore conséquente (bien que les nouvelles souscriptions soient rares à partir de 2006). Sur l'ensemble de la période, nous observons un phénomène périodique avec certains pics et creux de même amplitude qui semblent revenir régulièrement (même si estompé entre fin 2002 et fin 2003), d'où la pertinence de considérer une saisonnalité dans la modélisation (par la variable "month" ou "end.month" parfois).

Profil des rachats par ancienneté de contrat La forme très spécifique du graphique 3.2 s'explique par les frais engendrés par des rachats entre les dates anniversaires du contrat, ce qui a déjà été discuté au chapitre 1. La catégorisation de l'ancienneté du contrat semble "obligatoire" pour rendre compte de l'aspect non-monotone de cette courbe, à moins de considérer un pas de temps annuel (et encore) qui ferait ressembler la courbe à une exponentielle décroissante. Ce choix de pas de temps annuel a été banni pour conserver un volume d'observations suffisant, nécessaire à la bonne construction et validation du modèle probabiliste.

Taux de rachat par cohorte Le taux de rachat global par cohorte pour les produits mixtes est tout à fait particulier et instructif car il représente le pourcentage d'assurés de la cohorte qui ont racheté leur contrat. Le graphique 3.3 montre clairement des comportements très hétérogènes entre cohortes. Au vu du graphique 3.2 qui présente une décroissance homogène des pics de rachat en fonction de l'ancienneté, il est difficile de comprendre pourquoi les anciennes cohortes (2000 à 2002) ont un taux global de rachat de l'ordre de 30 % alors que celles entre 2002 et 2006 s'approchent des 80 %. Le bas niveau de rachat des plus jeunes cohortes s'explique facilement par le fait que le rachat est interdit en première année de contrat. Cette hétérogénéité entre cohortes a sûrement une explication rationnelle, toutefois difficile à récupérer même s'il s'agit probablement d'une politique de vente (ou législation) changeante sur ces produits.

Taux de rachat par date et par ancienneté de contrat La vision 3D du graphique 3.4 fournit une information additionnelle : ce n'est que récemment que le profil spécifique des

FIGURE 3.3 – Pourcentage global de rachat par cohorte pour les produits Mixtes.



rachats en fonction de l'ancienneté est apparu. Cette information primordiale vient valider la mise en place récente d'une spécificité (rachat sans frais aux dates anniversaires) qui a aussi pour effet l'apparition du créneau de la figure 3.3. L'hétérogénéité vient du mélange en portefeuille de ces deux types de population, car les changements n'ont visiblement été effectifs que pour les nouvelles souscriptions à partir de 2002.

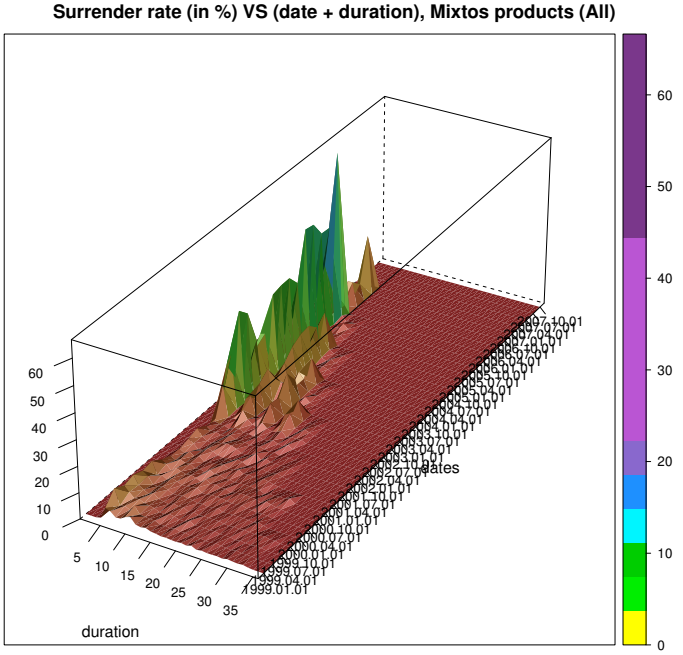


FIGURE 3.4 – Profil 3D du taux de rachat par date et par ancienneté de contrat (par trimestre), produit Mixtes.

3.2.2 Sélection des variables par CART

Taux d'erreur de classification de l'arbre Le classifieur par forêts aléatoires avec en variables d'entrée l'ensemble des variables à disposition (et pas seulement les variables catégorielles) donne d'excellents résultats. Le taux d'erreur de la matrice de confusion est de 4,6 %, avec une sensibilité de 99,5 % et une spécificité de 84 %. Ces statistiques nous sécurisent quant au classement (énoncé dans le paragraphe suivant) du pouvoir discriminant des variables.

	Rachats non-observés	Rachat observés
Rachats non-prédits	4599	877
Rachats prédits	85	15485

Importance des variables explicatives Comme énoncé dans la section 1.3.1, nous avons vérifié que le classement de l'importance des variables explicatives soit le même pour les périodes de pics de rachat comme pour les périodes creux (ce qui est le cas) lorsque nous regardons les comportements de rachat en fonction de l'ancienneté du contrat. Il n'y a donc pas de biais introduit dans les résultats de la figure 3.5, qui va nous servir de base pour la prise en compte des bons inputs lors de la modélisation. Nous prenons ainsi en priorité les variables de saisonnalité et d'ancienneté de contrat (catégorisée) déjà validées comme importantes, en y ajoutant l'option de PB et la prime de risque (catégorisée également car relation non-monotone encore une fois).

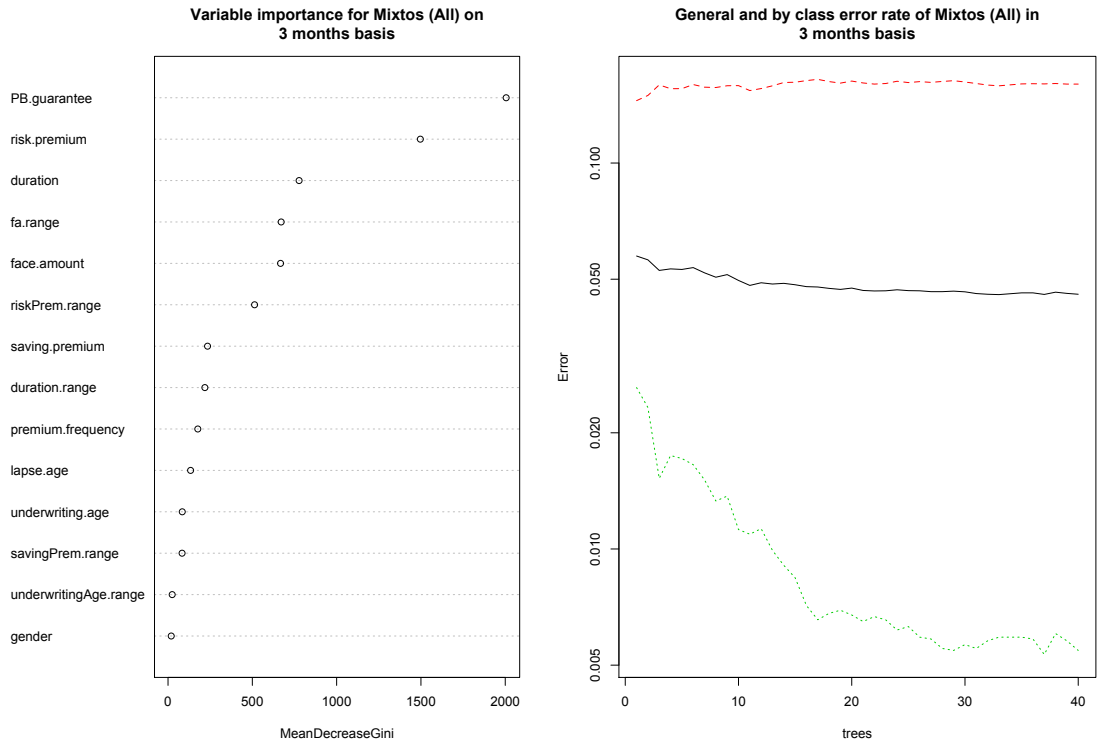


FIGURE 3.5 – Importance des variables explicatives, produit Mixtes.

3.2.3 Modélisation et prévisions par mélange de GLM

Nous ne commentons pas de nouveau le graphe 3.6 car cela a été fait longuement en section 2.1. Le pas trimestriel implique cependant une moins bonne estimation de la réalité sur la période d'apprentissage (en comparaison avec la figure 2.1), tout en conservant le problème majeur du changement de niveau du taux de rachat en 2007 qui n'est pas prévu par le modèle. Parmi des modèles mélange de régressions logistiques de deux à cinq composantes, nous choisissons celui qui minimise le critère BIC de sélection de modèle. Rappelons juste que ce critère prend en compte la complexité du modèle en pénalisant la vraisemblance d'un modèle qui contiendrait beaucoup de paramètres à estimer. Pour les produits mixtes, le modèle retenu a cinq composantes (nous verrons dans le chapitre suivant que ce n'est heureusement pas toujours le modèle avec le plus de composantes qui est retenu !), traduisant ainsi une forte hétérogénéité des données. Les résultats probants de la courbe 3.7 renforcent l'adéquation de la modélisation par mélange pour ce type de produit. En effet, nous constatons inévitablement le bon pouvoir prédictif de la méthode qui s'ajuste quasi-parfaitement aux observations tout en ayant un intervalle de confiance étroit, garantissant la robustesse de la modélisation. Nous avons vu comment choisir (CART) les variables explicatives à entrer dans la modélisation, de même que l'apport théorique des modèles mélange qui vont nous permettre de tenir compte de la forte hétérogénéité des comportements mise en évidence grâce à des statistiques descriptives ciblées. Il est maintenant grand temps de dévoiler notre intuition, celle qui nous guidera dans la modélisation de l'ensemble des familles de produit jusqu'à la fin de cette thèse. Selon nous, la logique voudrait que les effets structurels ne soient pas une source d'hétérogénéité entre comportements car ils s'appliquent à l'ensemble de la population sans distinction apparente.

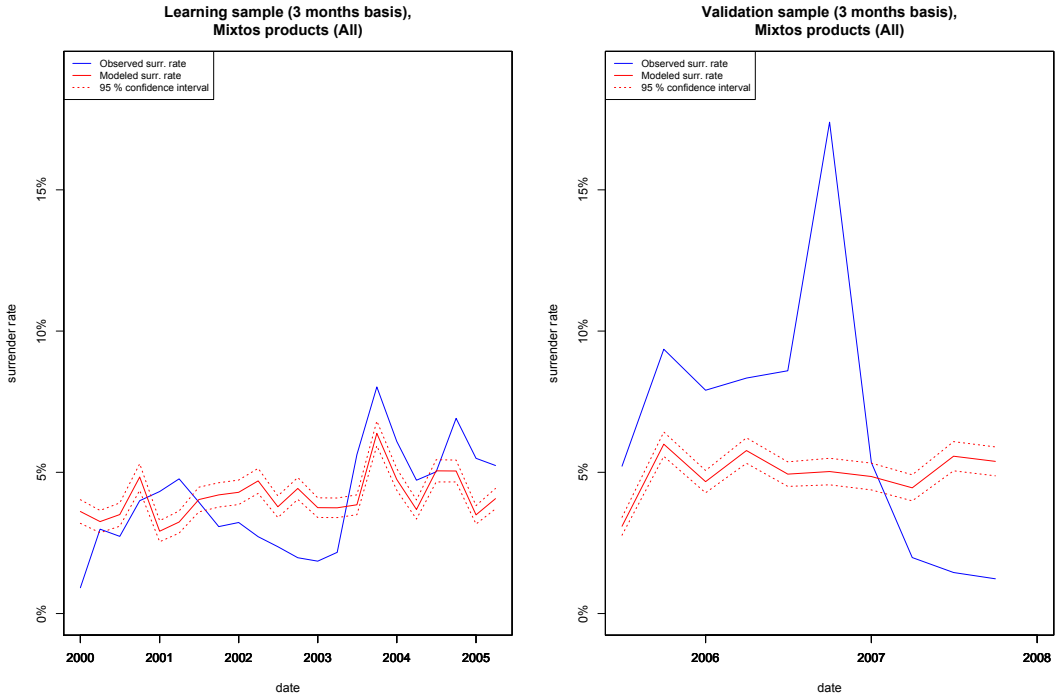
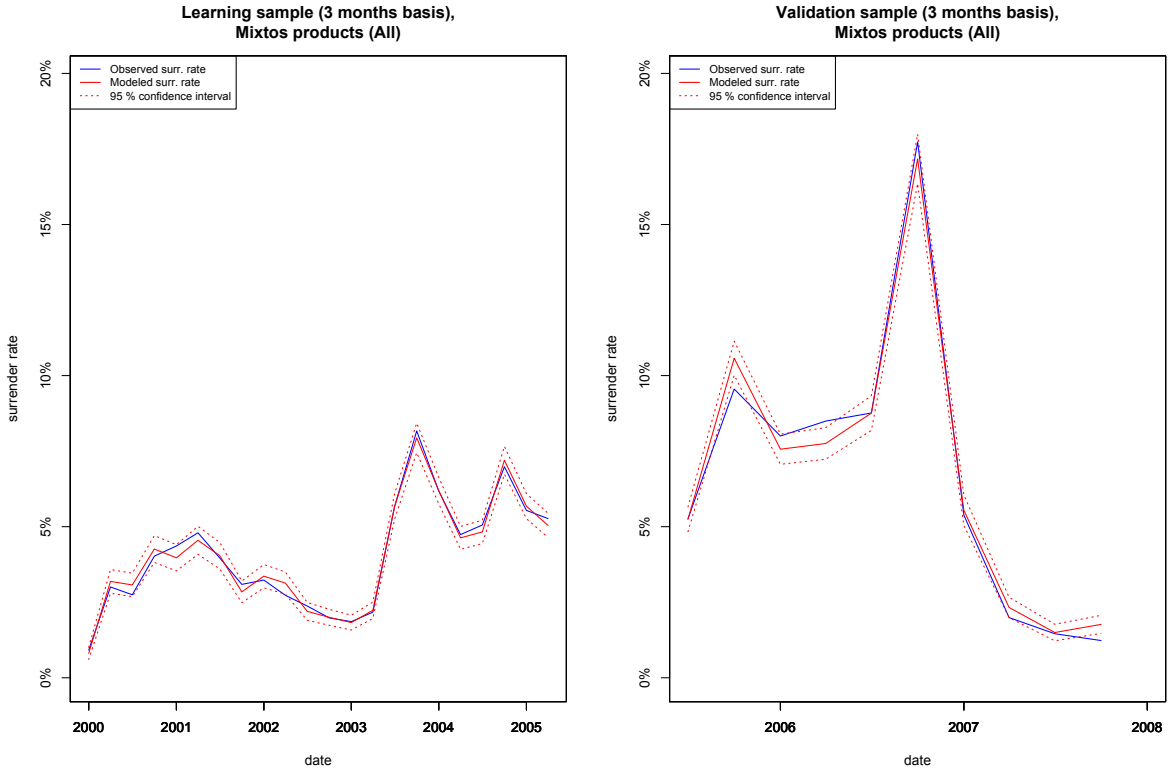


FIGURE 3.6 – Modélisation et prévision du taux de rachat des produits Mixtes par régression logistique dynamique.

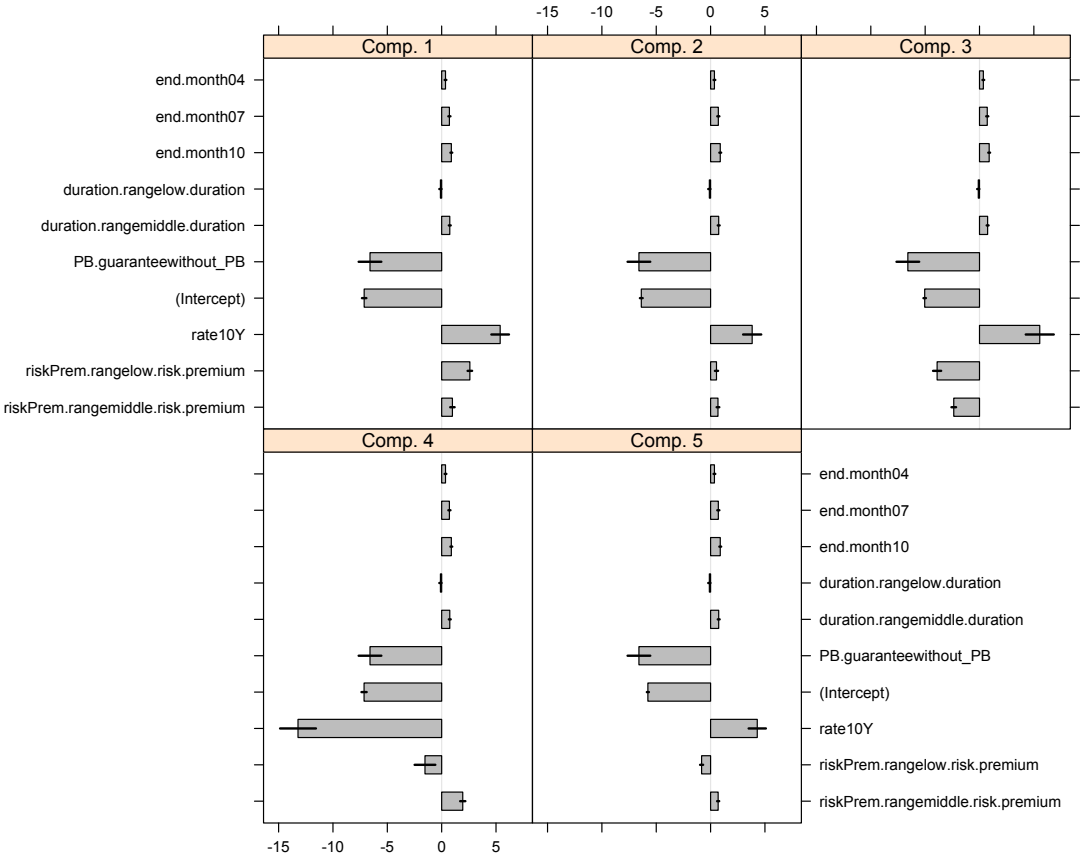
FIGURE 3.7 – Modélisation et prévision du taux de rachat par mélange de Logit (Mixtes).



Impact des variables explicatives par les mélanges de Logit A titre d'exemple, une contrainte fiscale reste une contrainte valable pour tous les assurés du portefeuille, et ce quelles que soient leurs autres caractéristiques. Par conséquent l'idée est de fixer les coefficients de régression constants entre les composantes pour les facteurs de risque associés à ces effets structurels, ce qui permet notamment de fortement limiter le nombre de paramètres à estimer. En revanche, la mixité de la population en termes de richesse et de sensibilité par rapport à l'environnement externe invoque des comportements de rachat totalement différents car très personnels. Il va ainsi se créer des groupes d'assurés pour lesquels il n'y a aucune raison que les paramètres de régression associés aux effets conjoncturels aient la même valeur, de même que pour le risque de base représenté par l'ordonnée à l'origine ("intercept"). Cette proposition fort logique et simple nous permet de prendre en compte une bonne part des différents comportements et de reconstruire correctement l'historique (en *back-testing*) du taux de rachat. L'estimation des coefficients de régression du modèle mélange appliqué aux contrats mixtes est disponible en figure 3.8. Les impacts des facteurs de risque sont les suivants :

– effets *structurels* (identiques quelle que soit la composante d'appartenance) : un faible effet de saisonnalité est détecté (valeur absolue des coefficients de régression associés faible) avec globalement de plus en plus de rachats en approchant de la fin de l'année civile. L'effet de l'ancienneté du contrat catégorisée en trois modalités ("low", "middle" et "high") est clair : c'est dans la tranche des anciennetés moyennes que se situent le plus grand nombre de rachats, et globalement les assurés rachètent assez rapidement. Le fait

FIGURE 3.8 – Coefficients de régression des composantes du mélange de Logit.

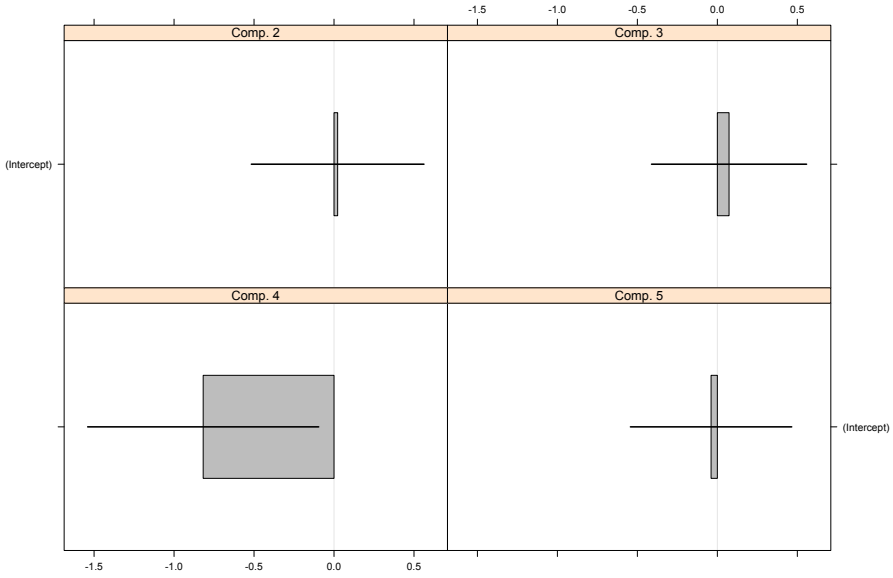


de ne pas avoir l'option de PB dans son contrat vient fortement diminuer la probabilité individuelle de rachat, ce qui confirme les observations faites auparavant.

– effets *conjoncturels* : l'effet de la prime de risque (liée à la richesse des assurés) est hétérogène et semble dicter en partie la sensibilité des agents aux mouvements de l'économie. Cela rejoint l'idée exprimée par les équipes Marketing, selon laquelle la réactivité des assurés est relative à ce qu'ils possédent. Nous pouvons remarquer que les groupes pour lesquels le niveau de richesse est fortement discriminant réagissent en plus forte proportion aux mouvements du taux 10 ans.
– effets de *corrélation* : introduite via le contexte économique. Le taux 10 ans a une importance prépondérante en termes d'impact (valeur absolue du coefficient associé élevée). La calibration montre que certains assurés rachètent plus lorsque le taux long-terme augmente (composantes 1, 2, 3 et 5) alors que les autres adoptent le comportement inverse, la rationalité étant illustrée par une augmentation des rachats en cas de hausse des taux sur ce type de produit (rendement garanti). Chaque trimestre, la hausse ou la baisse de ce taux va augmenter ou baisser **en même temps** la probabilité de rachat des assurés appartenant à un groupe donné, faisant évoluer dynamiquement la distribution du mélange au cours du temps.

Ces calibrations semblent robustes (l'écart-type des estimations est représenté par une proportion de la petite barre noire) puisque la valeur nulle n'appartient à aucun intervalle de

FIGURE 3.9 – Coefficients de régression des poids des composantes du mélange de Logit.



confiance de ces estimations. Pour ce qui est de la calibration des poids de chaque composante, les résultats semblent moins robustes et ce sera souvent le cas en pratique. La figure 3.9 résume les proportions de chaque composante dans le mélange (pas de variable explicative ici), nous obtenons par la formule (3.7) :

$$\pi_1 = 22\%, \quad \pi_2 = 23\%, \quad \pi_3 = 24\%, \quad \pi_4 = 10\%, \quad \pi_5 = 21\%,$$

ce qui semble indiquer qu'il n'y a pas de composantes inutiles, chacune ayant son importance dans le mélange. Néanmoins, l'estimation de ces poids trahit la confiance que nous pouvons avoir en leur valeur. De plus, il est possible que le nombre de composantes sélectionné soit légèrement sur-estimé si nous en croyons l'estimation des coefficients de régression : en effet, certaines composantes ont tendance à se ressembler (la première et la deuxième), quoique ce n'est pas forcément évident ici (mais plusieurs produits révèlent cette faiblesse, cf annexe C).

Pour vérifier la robustesse de cette approche autrement que par l'aspect visuel, nous appliquons deux tests : un test de normalité des résidus (Pearson), et un test sur les distributions (Wilcoxon Mann-Whitney). Nous ne détaillons pas le test de Pearson qui est un des plus connus ; le principe du test de Wilcoxon-Mann-Whitney est donné ci-dessous. Les résultats de ces deux tests pour un seuil de 5% suivent dans le tableau 3.1. Nous ne pouvons donc pas rejeter l'hypothèse nulle qui correspond au fait que la variable aléatoire "observée" et la variable aléatoire "prédite" aient la même distribution. Les sorties R des résultats numériques de ces tests sont disponibles en annexe C.1.

	Test de Pearson	Test de Wilcoxon-Mann-Whitney
p-valeur	0.8495	0.7394

TABLE 3.1 – p-valeur des tests de résidus et de distribution pour validation.

3.3 Extension au portefeuille Vie d'AXA

L'intérêt de cette partie réside dans l'application pratique des théories développées dans les chapitres précédents, dans le but de valider la méthodologie adoptée dans la section antérieure. Le portefeuille d'Assurance-Vie épargne d'AXA Seguros est utilisé dans toute sa "largeur", avec des résultats allant de produits de pur investissement à des produits alliant des composantes épargne à des garanties de prévoyance, en passant par des produits directement indexés sur les marchés financiers. Nous verrons que la modélisation proposée a un fort pouvoir d'adaptation et fournit des résultats très encourageants en termes de pouvoir prédictif, tout en conservant l'originalité de ne pas impliquer trop de facteurs explicatifs afin de ne pas trop complexifier le modèle. Chaque section de cette partie correspond à l'étude d'une famille de produit, avec toujours le même plan d'étude : une explication très succinte du type de contrat (car les produits sont agrégés), suivie des deux modélisations logistiques (mélange ou non) avec prévisions associées (l'analyse basée sur les statistiques descriptives, les résultats de la méthode CART qui permettent la sélection des facteurs de risque, et les tests se trouvent en annexe C pour ne pas trop alourdir ce chapitre). D'un point de vue granularité des données, il est nécessaire d'étudier les rachats par famille de produits au maximum (une agrégation encore plus grande n'aurait plus de sens) car les supports d'investissement et les options classiques varient d'une famille à l'autre, ce qui apporte des changements importants en termes de modélisation. Il va sans dire que l'idéal est d'affiner les études à l'échelle de lignes de produits, voire de produits. L'outil informatique que nous avons développé permet de choisir son niveau de granularité, mais nous préférions montrer que notre méthode fonctionne à une échelle d'agrégation importante (sachant qu'à l'échelle d'un produit, cette modélisation est souvent moins complexe car nous connaissons exactement toutes les clauses et options qui impactent le rachat ; il suffit alors de les inclure dans la modélisation). De plus, une étude par produit ne permettrait pas de modéliser les rachats globalement, car les corrélations entre produits seraient difficilement calibrables.

Pour les résultats de la modélisation par mélange de régressions logistiques, nous avons choisi de commenter les effets des variables explicatives au vu des estimations des coefficients de régression sans pour autant afficher les "boxplots" correspondants pour des soucis de concision. Le lecteur intéressé pourra consulter les annexes C pour accéder à ces informations plus précises. Toute l'étude est basée sur un pas de temps trimestriel et sur une période de retour *delta* (durée sur laquelle l'assuré regarde la performance des indices avant la date de rachat) de un trimestre, ces options pouvant être ajustée dans notre outil (pas mensuel, trimestriel ou annuel et *delta* doit être un entier positif).

3.3.1 Les contrats de pure investissement (Ahorro)

Les contrats "Ahorro" sont des contrats de pure épargne. Ils offrent un rendement différent suivant le produit considéré, mais tous sont des taux garantis (le risque de taux est donc porté par l'assureur). Nous pourrions comparer ces contrats à des contrats bancaires, avec la différence qu'ils offrent des avantages fiscaux et/ou des garanties supplémentaires. Les informations dont nous disposons pour ce type de contrats sont le numéro du produit, la date d'émission, la date de sortie et sa raison (si sortie il y a), l'option de participation aux bénéfices de l'entreprise (PB), la date de naissance de l'assuré, son sexe, sa richesse, la fréquence de la prime, la prime de risque et la prime d'épargne. Un aperçu des données formatées est disponible en annexe C.2.1. La période de données va de début 1999 à fin 2007 (certains contrats sont évi-

demment souscrits avant 1999), mais la période d'étude s'étend du 1/1/2000 au 31/12/2007 car les rachats n'ont été répertoriés qu'à partir de début 2000.

Modélisation et prévisions par mélange de GLM

Pour toutes les applications suivantes, les mêmes variables explicatives sont considérées en input de la modélisation dynamique et de la modélisation par mélange. L'approche par mélange permet de prendre ces variables en compte de manière différente, mais il est primordial de garder à l'esprit que nous prenons exactement les mêmes informations en entrée des modèles afin de comparer ce qui est comparable. Cette remarque justifiera le fait que certains modèles mélange ne sont pas optimisés (en termes de variables considérées, de nombre de composantes car parfois certaines composantes se ressemblent fortement...). Dans une optique où la volonté de l'utilisateur est de trouver la meilleure solution de modélisation, cette optimisation est tout à fait réalisable dans des délais raisonnables.

Le but est de comparer l'approche par mélange de régressions logistiques avec la régression logistique dynamique, et de voir s'il y a un apport conséquent de cette nouvelle modélisation. Nous discutons de l'impact des facteurs de risque suivant les groupes d'assurés dans le cadre de la modélisation mélange, et effectuons des comparaisons grâce aux prévisions des décisions individuelles qui nous permettent de reconstruire le taux de rachat par date.

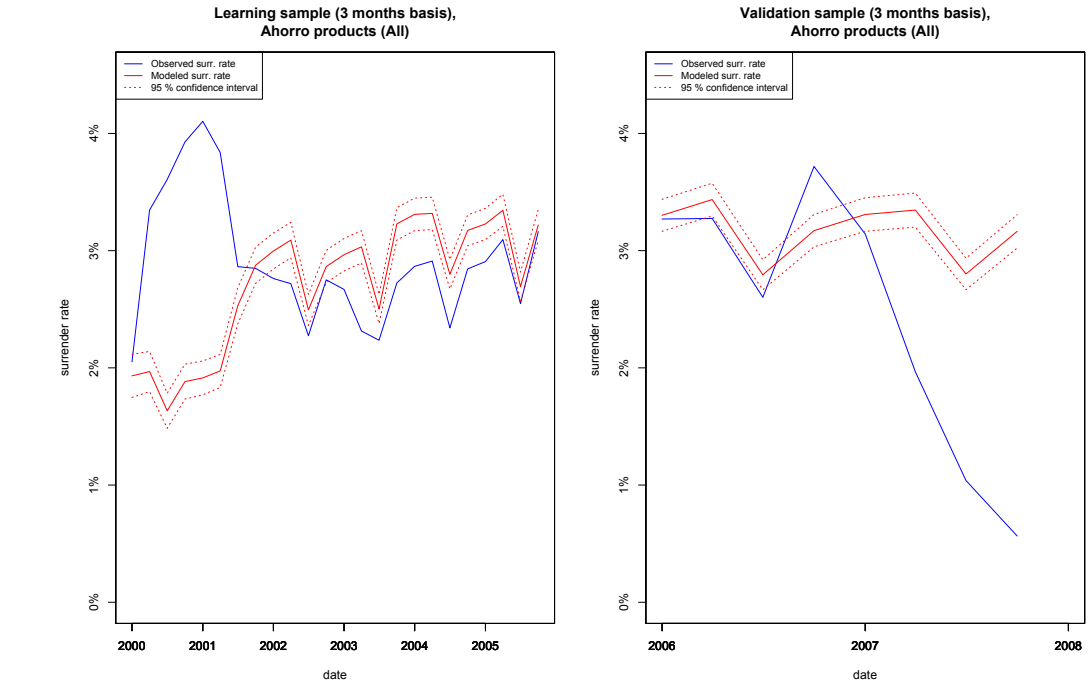
Comparaison et discussion Les mauvais résultats de la modélisation par régression logistique dynamique simple sont très frappants (graphe 3.10). La cause de cette "faillite" est l'environnement économique changeant qui est mal modélisé, pour preuve la valeur du coefficient de régression consacré à l'impact du taux 10Y qui est extrêmement faible (0,06). Cela signifie qu'une forte variation de ce taux n'a que peu d'impact sur la probabilité finale de décision individuelle de rachat, ce qui est évidemment très discutable. Nous constatons également que le modèle logistique dynamique modélise bien la périodicité.

De par la flexibilité permise par les mélanges, les prévisions s'avèrent nettement plus justes et précises aussi bien sur la période d'apprentissage que sur la période de validation (graphe 3.11). Ce changement se retrouve notamment dans la valeur des coefficients de régression correspondant au taux 10Y (entre 10 et 100 plus élevé suivant les composantes), traduisant un impact nettement plus réaliste de cette variable (voir figure C.6).

Impact des variables explicatives par les mélanges de Logit

Nous partons du postulat que l'hétérogénéité provient de facteurs de risque qui peuvent avoir un effet différent suivant les personnes. L'idée de base est donc que les effets structurels bien connus (ancienneté de contrat, saisonnalité) sont censés avoir un impact homogène et constant quels que soient les groupes d'assurés considérés, alors que les effets conjoncturels (environnement économique) jouent différemment suivant les assurés. La mise en oeuvre de cette idée requiert de spécifier une estimation identique des coefficients de régression correspondant aux effets structurels pour toutes les composantes, en permettant aux coefficients de régression dédiés aux effets conjoncturels de varier entre composantes. Les professionnels ont coutume de considérer un taux d'intérêt long terme pour les produits de pure épargne à rendement garanti, aussi nous avons pris le taux 10 ans (taux 10Y). C'est ainsi que nous obtenons après estimation du modèle les coefficients de régression donnés en annexe C.2.4. Détaillons maintenant les impacts respectifs des facteurs de risque :

FIGURE 3.10 – Modélisation et prévision du taux de rachat des produits Ahorro par régression logistique dynamique.



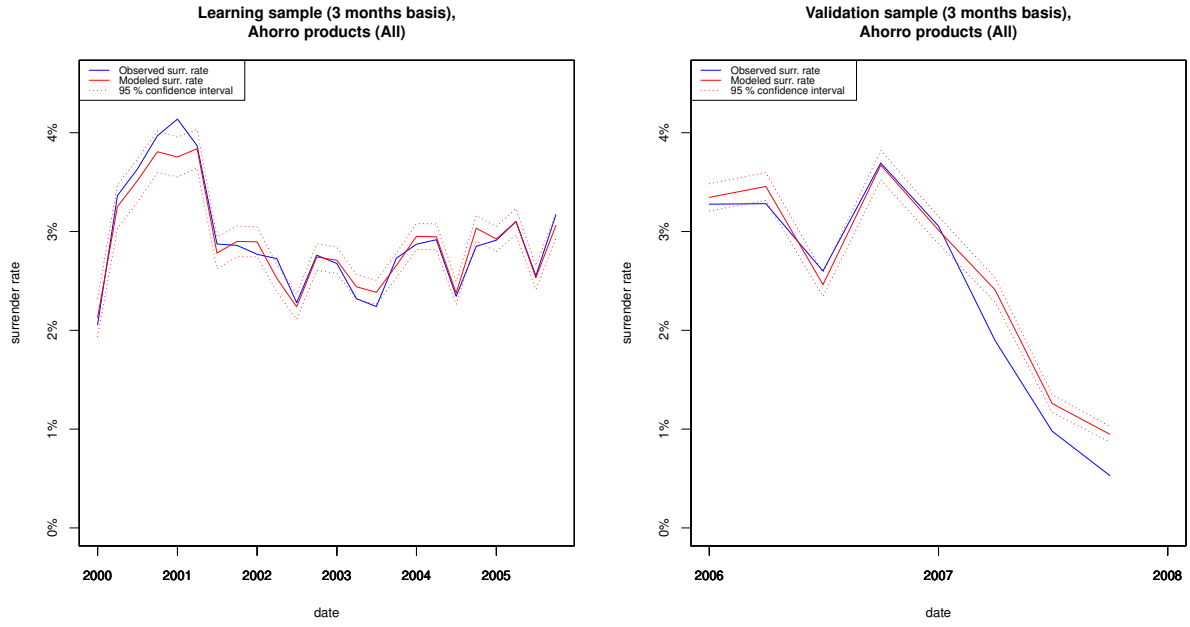


FIGURE 3.11 – Modélisation et prévision du taux de rachat par mélange de Logit, produits Ahorro.

- effets structurels : identiques à tout le monde. En ce qui concerne la saisonnalité, moins de rachats constatés en été et environ le même taux de rachat en début et en fin d'année civile. Le risque de rachat est fort lorsque l'ancienneté de contrat (catégorisée en 3 modalités : faible, moyenne et longue) est faible, ce qui confirme les pics constatés dans le graphe C.2. Plus les assurés sont âgés et moins la probabilité de rachat est grande, et l'effet de la fréquence de la prime (regroupée en 3 modalités : haute périodicité, moyenne et prime unique) est confirmé : plus la prime est fréquente et plus la probabilité de rachat est grande. Le fait de ne pas avoir l'option de PB abaisse fortement la probabilité de rachat.
- effets conjoncturels : deux groupes se distinguent. Pour le premier groupe, un taux 10Y qui augmente fait baisser la probabilité de rachat des assurés (composantes 1 et 3) alors que l'effet est inverse pour les autres groupes d'assurés. L'intensité de cette sensibilité caractérise ensuite les différentes composantes.

Nous constatons ainsi que les assurés réagissent différemment aux mêmes mouvements des effets du marché obligataire, venant confirmer l'irrationnalité et l'hétérogénéité des réactions.

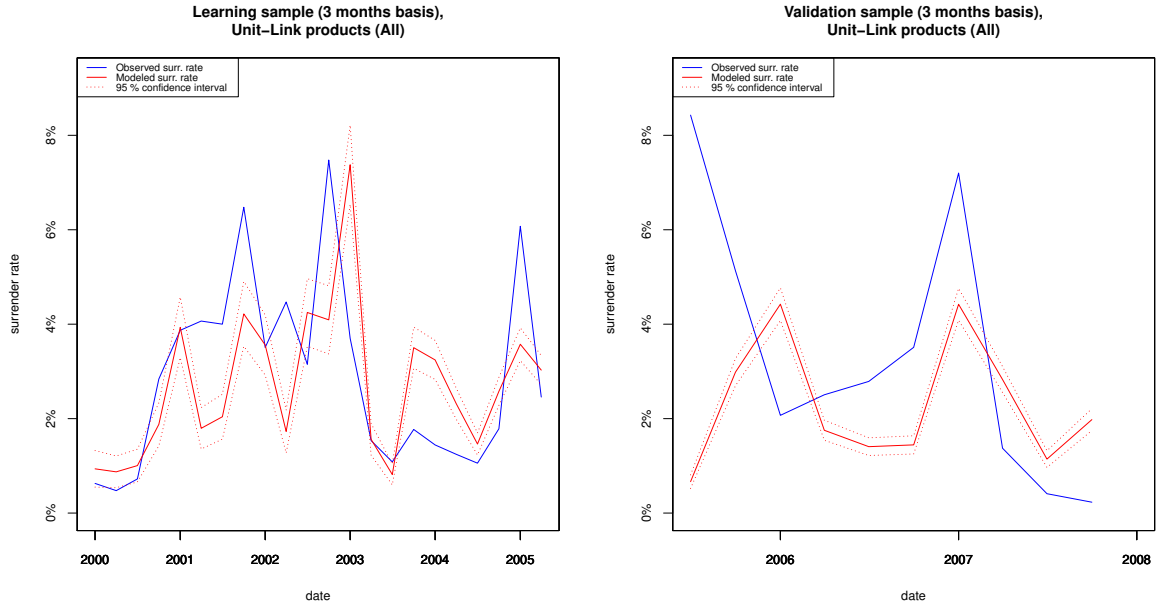
3.3.2 Les contrats en Unités de Compte (Unit-Link)

Les contrats en UC sont des contrats qui offrent un rendement variable suivant les performances des marchés financiers. La rentabilité n'est donc pas garantie, bien qu'on adosse à certains de ces contrats des garanties plancher, ce qui limite le risque porté cette fois-ci par l'assuré. En général, ces contrats d'épargne offre des garanties supplémentaires telles qu'une couverture contre le décès, et les unités de compte sont basées sur des obligations et actions de diverses entreprises. Les informations dont nous disposons pour ce type de contrat sont le numéro du produit, la date d'émission, la date de sortie et sa raison (si sortie il y a), l'option de participation aux bénéfices, la date de naissance de l'assuré, son sexe, sa richesse, la fréquence de la prime, la prime de risque et la prime d'épargne. En fait le type d'information est identique que pour la famille précédente car nous avons la même base de données originelle, les donnés formatées ont donc le même aperçu (annexe C.2.1). La période d'étude va du 1/1/2000 au 31/12/2007.

Modélisation et prévisions par mélange de GLM

Le contexte des produits en Unités de Compte est particulier puisque ceux-ci sont indexés sur les marchés financiers. Nous connaissons la volatilité du marché, qui a ainsi un impact direct sur la volatilité du taux de rachat lui-même. Mêlée aux effets "cohortes", une hétérogénéité très forte apparaît pour ce type de produit, pour lequel les comportements de rachat sont donc très difficilement prévisibles comme illustré par le graphe 3.12. C'est certainement dans ce contexte que la modélisation mélange a le plus d'apport. La sensibilité des assurés aux mouvements des marchés est évidemment très hétérogène. Le graphique 3.13 permet de constater que les principaux effets sont bien captés par le modèle, en bonne proportion et dans le bon sens. Le taux de rachat observé appartient à l'intervalle de confiance des prévisions sur toute la période (excepté fin 2005 et fin 2006), et ce malgré notre méthode de validation temporelle. La différence de la quantification de l'effet des marchés financiers entre la modélisation classique et la modélisation mélange est très importante, tant en termes de sens de l'impact que d'intensité.

FIGURE 3.12 – Modélisation et prévision du taux de rachat des produits UC par régression logistique dynamique.



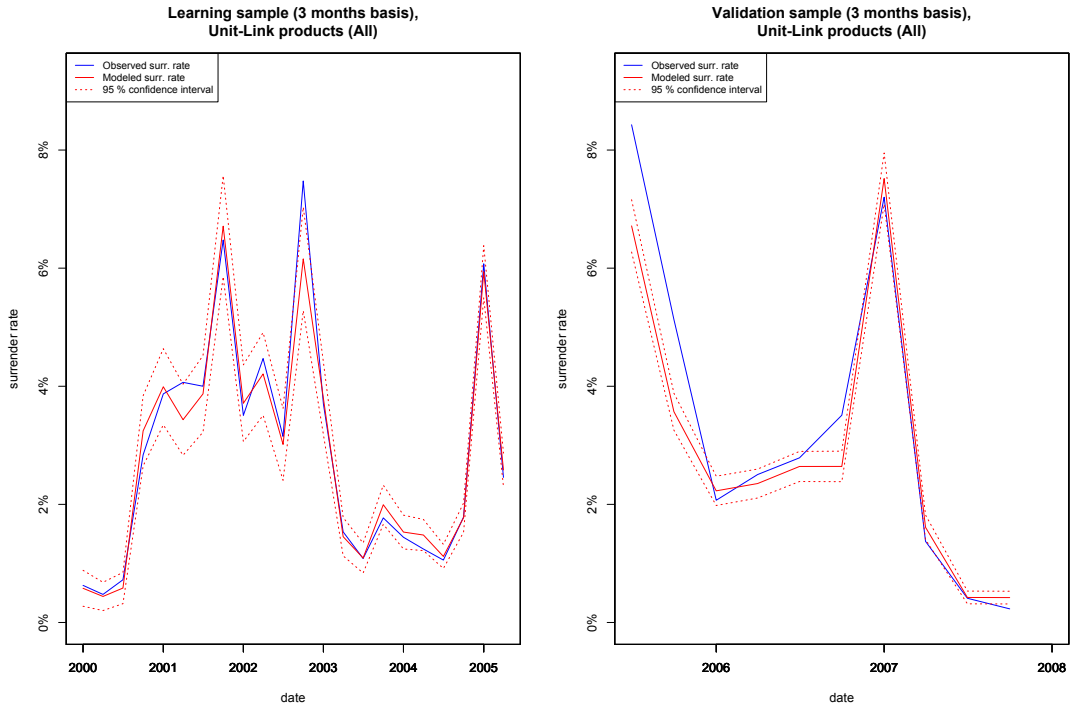


FIGURE 3.13 – Modélisation et prévision du taux de rachat par mélange de Logit (UC).

Impact des variables explicatives par les mélanges de Logit

L'annexe C.3.3 donne l'estimation des coefficients de régression du modèle mélange. Nous adoptons toujours la même méthode de choix d'estimation des variables (structurelles → coeff. identiques, conjoncturelles → coeff. variables) en espérant que les résultats soient probants. Détaillons maintenant les impacts des facteurs de risque :

- effets *structurels* : identiques à tout le monde. Pour la saisonnalité, les conclusions sont ressemblantes avec celles des contrats de pure épargne (cycle du marché de vente) : l'été est une période où très peu de rachats sont observés. L'effet de l'ancienneté du contrat est nettement moins évident comme le graphe C.9 l'avait laissé présager. La richesse de l'assuré semble peu jouer, même si les personnes les plus riches ont l'air de racheter davantage (peut-être conseillées par un agent qui gère leur fortune).
- *effets conjoncturels* : une très grande majorité d'assurés (sauf ceux de la composante 4) rachète plus lorsque l'indice boursier plonge, mais il existe un petit groupe de personnes pour qui ce n'est pas le cas du tout (valeur positive du coefficient élevée). D'un trimestre à l'autre, la probabilité de rachat individuelle des assurés appartenant à un groupe (composante) donné change en fonction de la valeur de l'Ibex 35, créant ainsi une corrélation positive entre les personnes de ce groupe.

Les mêmes constatations peuvent être formulées concernant l'irrationnalité et l'hétérogénéité des réactions des assurés, même si la proportion d'agents irrationels paraît ici très limitée. Il semblerait que la rationalité des agents soit plus forte pour ce type de produit.

3.3.3 Les contrats liés au indices boursiers (Index-Link)

Ce type de contrat est très semblable aux contrats en UC. La différence réside dans le support d'investissement qui est ici plus ciblé car il s'agit uniquement d'indices boursiers. Les variables assurés et contrat dont nous disposons sont identiques aux types de produit précédent (issu de la même base données) et nous étudions donc ces contrats sur la période 1/1/2000 - 31/12/2007. Nous devrions logiquement obtenir des résultats en ligne avec l'étude des produits en UC.

Modélisation et prévisions par mélange de GLM

Comme pour les produits en UC, le modèle statistique de régression logistique dynamique est inefficace tant sur la période d'échantillonnage où il ne reflète pas les pics en début de période, que sur la période validation où le niveau de rachat n'est pas bien ajusté (cf graphe 3.14). La conclusion de ce constat est que certes les effets économiques sont modélisés, mais la calibration de ces effets n'est visiblement pas adéquate. A l'inverse, le graphique 3.15 est très satisfaisant, le taux de rachat observé aussi bien sur la période d'apprentissage que sur la période validation reste toujours dans l'intervalle de confiance du taux prédit. La prise en compte spécifique des variables explicatives dans le mélange permet d'arriver à ces résultats, avec toujours en idée de modéliser l'hétérogénéité par des coefficients de régression variables entre composantes pour les effets conjoncturels.

Impact des variables explicatives par les mélanges de Logit

L'estimation des coefficients de régression du modèle mélange en annexe C.4.3 dévoile moins d'hétérogénéité que pour les produits en UC, peut-être à cause du fait que le sup-

port des produits soit relativement simple à interpréter (seul l'ibex 35 sert de valorisation au contrat). Cela rend la compréhension du produit et l'interprétation des résultats du contrat plus simples pour l'assuré, contrairement à un produit qui serait indexé sur plusieurs supports et dont l'assuré aurait du mal à savoir de manière globale la valeur. Nous adoptons toujours la même méthode de calibration des coefficients de régression pour les variables (structurelles → coefficients identiques, conjoncturelles → coefficients variables entre composantes). Les impacts des facteurs de risque sont les suivants :

- effets *structurels* : identiques à toutes les composantes. Pas d'effet saisonnalité introduit sur ce type de produit, l'ancienneté du contrat joue toujours dans le même sens (les assurés rachètent globalement rapidement). Les hommes rachètent plus que les femmes (effet ajouté car visible dans les statistiques descriptives mais non retranscrit par les arbres), et les personnes âgées semblent racheter moins souvent que les autres.
- effets *conjoncturels* : le critère BIC sélectionne un mélange à seulement deux composantes. L'hétérogénéité est donc moins grande a priori, d'ailleurs l'ensemble des assurés réagit aux mouvements de l'Ibex 35 dans le même sens (seule la sensibilité à ces mouvements est plus ou moins grande).
  La probabilité de rachat individuelle des assurés augmente exponentiellement en fonction de l'évolution de l'Ibex 35, avec la même amplitude pour tous les individus appartenant à la même composante (corrélation positive entre les comportements).

Le calibrage du modèle pour cette famille de produit ne nécessite que deux composantes (annexe C.4.3), et l'estimation des proportions du mélange semble robuste.

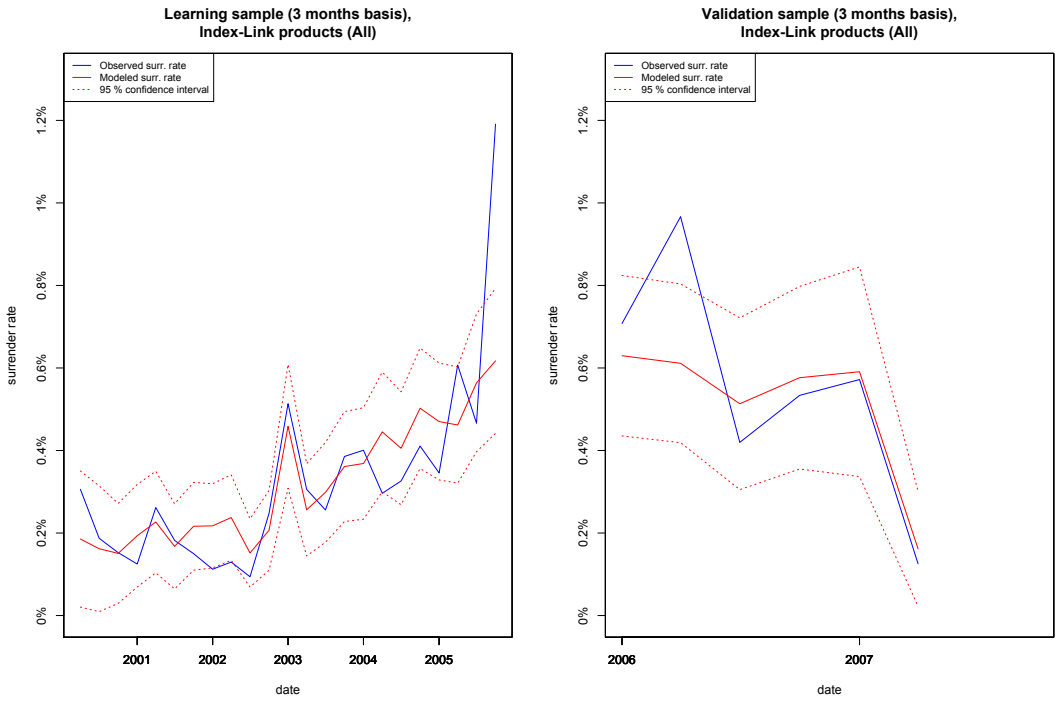
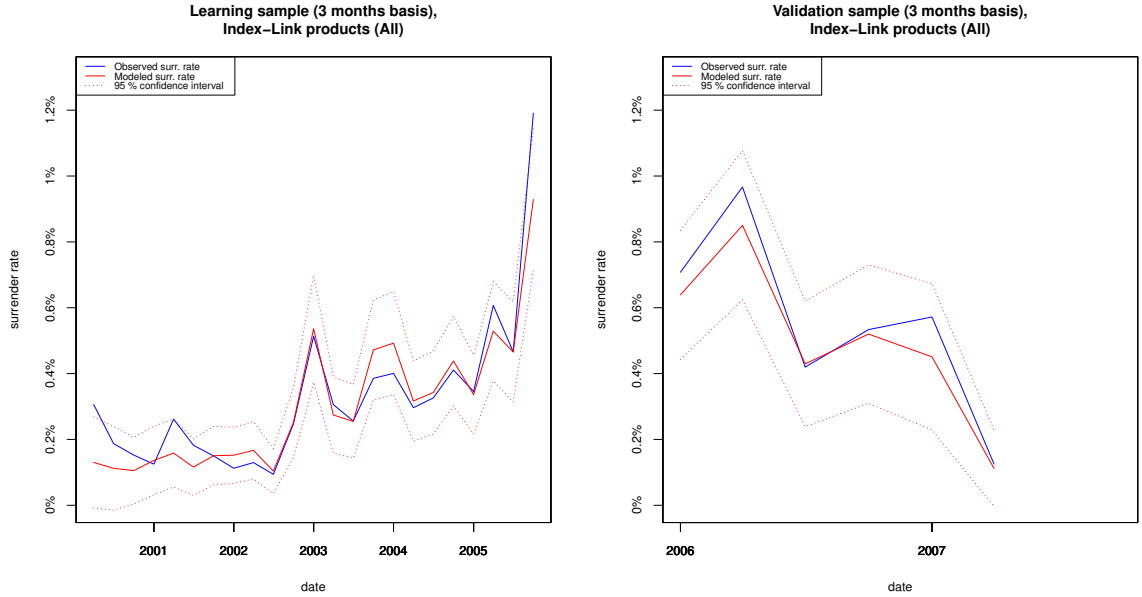


FIGURE 3.14 – Modélisation et prévision du taux de rachat des produits Index-Link par régression logistique dynamique.

FIGURE 3.15 – Modélisation et prévision du taux de rachat par mélange de Logit, produits Index-Link.



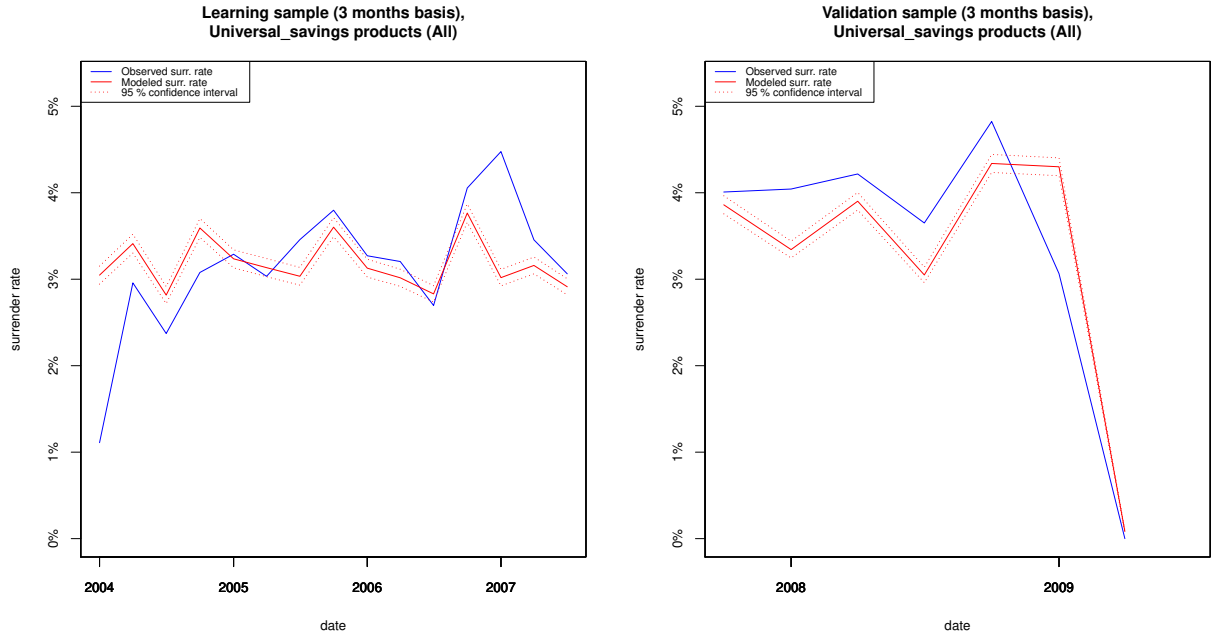
3.3.4 Famille "Universal savings"

Ces contrats d'épargne à taux garanti offrent des garanties prévoyance supplémentaires. En général, il s'agit de garantie classique contre le décès de l'assuré, mais certains assureurs proposent des options supplémentaires (appelées "rider") comme par exemple des garanties d'incapacité ou invalidité. Dans leur fonctionnement les Universal Savings se rapprochent des Ahorro, mais la composante Prévoyance vient sûrement modifier l'usage qui en est fait par l'assuré. La période d'observation va de 1985 à fin 2009, et les variables explicatives dont nous disposons sont le numéro du produit, la date d'émission, la date de sortie et sa raison (si sortie il y a), la date de naissance de l'assuré, son sexe, sa richesse, la fréquence de la prime, la prime de risque, la prime d'épargne et le réseau de distribution. Dommage que l'on ne dispose pas de l'information sur le type de prime (nivelée ou non) qui doit être un facteur explicatif important (effet psychologique).

Modélisation et prévisions par mélange de GLM

La chute du taux de rachat fin 2008 est relativement bien capté dans le modèle simple de régression logistique dynamique, ce qui veut dire que l'impact du facteur provoquant cette chute a été bien quantifié dans la procédure. Par contre, les niveaux et variations du taux de rachat sur la période d'apprentissage sont mal ajustés par le modèle. Les effets semblent pourtant être modélisés dans le bons sens (les variations se font dans la même direction sur la courbe 3.16), au détriment de l'amplitude qui manque de précision. Le mélange de régressions logistiques permet non seulement de capter les bons effets mais surtout de rendre compte de l'hétérogénéité présente dans les données, en capturant parfaitement le comportement des groupes de personnes et en restituant une modélisation très précise du taux de rachat, agré-

FIGURE 3.16 – Modélisation et prévision du taux de rachat des produits Universal Savings par régression logistique dynamique.



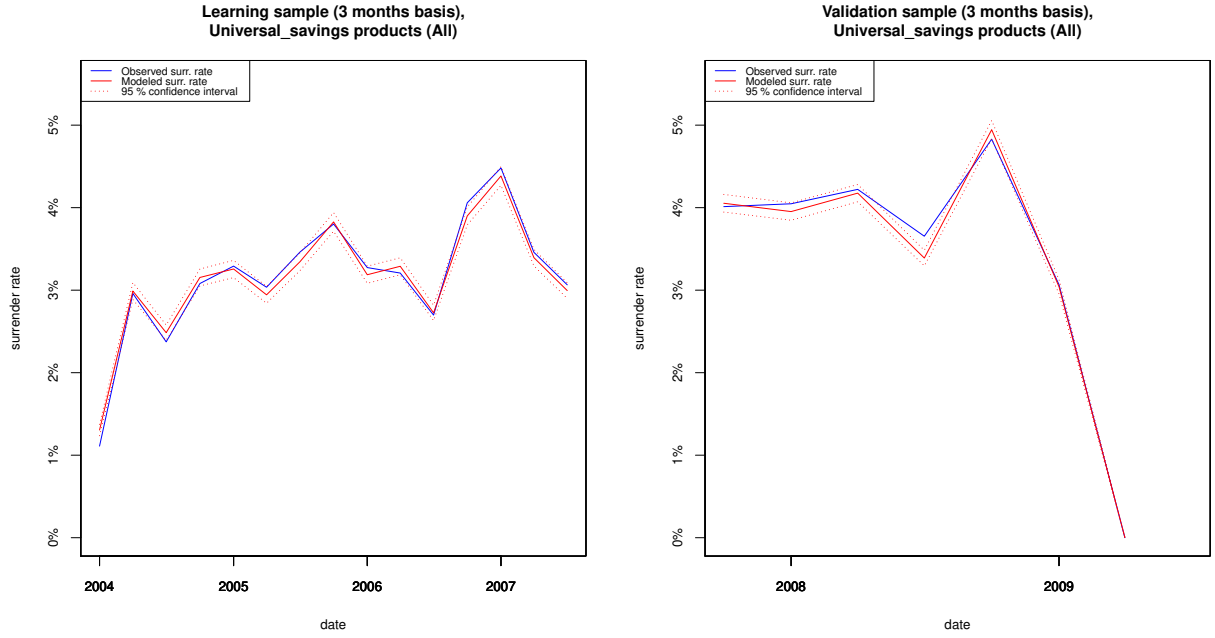


FIGURE 3.17 – Modélisation et prévision du taux de rachat par mélange de Logit, produits Universal Savings.

gation de l'ensemble des décisions individuelles (graphe 3.17). De plus, le taux observé restant dans l'intervalle de confiance des prévisions (malgré son étroitesse) tout au long de la période d'étude confirme la parfaite adéquation du modèle.

Impact des variables explicatives par les mélanges de Logit

La découverte que nous avons faite pour cette famille de produit est étonnante : il ne semble pas nécessaire de prendre en compte le contexte économique pour avoir une modélisation précise. Les graphiques ci-dessus sont issus de modélisations logistiques à une ou plusieurs composantes, mais pour lesquelles aucune variable de type taux long-terme ou Ibex 35 n'ont été introduites. La caractéristique du type de produit, qui accorde une plus grande importance aux garanties de prévoyance (liées aux risques de la vie), change visiblement la manière qu'ont les assurés de prendre leur décision. La couverture prévoyance semble prendre le dessus sur l'environnement économique, prouvant que le produit est perçu davantage comme un produit de prévoyance que comme un produit d'épargne.

L'estimation des coefficients de régression du modèle mélange en annexe C.5.3 confirme l'importance de la variable "richesse" dans le processus de décision. L'hétérogénéité des comportements est capté principalement grâce à cette variable (le risque de base donné par l'intercept est presque le même pour toutes les composantes). Les impacts des facteurs de risque sont :

- effets *structurels* : faible effet de saisonnalité avec une augmentation des rachats en fin d'année civile, l'effet de l'ancienneté du contrat est sensiblement différent. Les assurés semblent racheter leur contrat en moyenne plus tard (ancienneté moyenne). Les personnes âgées rachètent moins (besoin accrue de se couvrir contre les risques vie), et le réseau de distribution est discriminant : un suivi du contrat plus personnalisé conduit à une diminution de rachats.
- effets *conjoncturels* : ce sont les assurés de richesse intermédiaire qui rachètent le plus. Vu les poids des composantes (annexe C.24), la plupart des assurés adopte ce comportement (composantes 3 et 4). Cependant, d'autres (une faible minorité) ont un comportement différent et rachète davantage malgré leur grande richesse.

Le mélange comporte cinq composantes, en proportion respective de la composante 1 à 5 : 13 %, 10 %, 30 %, 35 % et 12 %. L'estimation robuste de ces poids confirment que chaque composante joue un rôle de capture d'un type de comportement, permettant au modèle de reéquilibrer les effets par période en fonction de la composition du portefeuille.

3.3.5 Les contrats à taux garanti : les "Pure savings".

Cette gamme de contrat peut s'apparenter complètement aux contrats de type "Ahorro". La dénomination diffère car la base de données que nous utilisons pour les étudier est différente de celle des "Ahorro", plus complète et couvrant une plus large période : de 1967 à fin 2009 (mais comme précédemment, les rachats n'ont été répertoriés qu'à partir du 1/1/2004). Les informations disponibles sont identiques à celles concernant les produits Universal Savings car la base de données d'origine est la même.

Modélisation et prévisions par mélange de GLM

Nous avons vu que l'hétérogénéité des données n'était pas si grande que d'habitude sur cette ligne de produit. Cette remarque est quelque part validée par le modèle de régression logistique dynamique, qui ne s'en sort pas si mal en termes de modélisation et de prévision (graphe 3.18).

Quelques ajustements (notamment sur la période de validation) seraient préférables mais la dynamique du taux de rachat est grosso modo reproduite par le modèle. Qu'en est-il par l'usage des mélanges ?

Les produits de type Pure Savings sont l'unique cas dans toutes nos données où l'apport de la modélisation mélange n'est pas forcément évident. Le graphique 3.19 relate l'évolution du taux de rachat observé et du taux de rachat prédit et semble montrer un meilleur ajustement du modèle, mais au prix d'une certaine complexification. En effet, nous pouvons remarquer que le niveau est mal ajusté en milieu d'année 2008 (nous sommes d'ailleurs assez loin de la réalité), ce qui laisse supposer que certains effets non-observables ont joué à ce moment là, mais ne sont pas captés par la modélisation.

Impact des variables explicatives par les mélanges de Logit

Le "boxplot" de l'estimation des coefficients de régression du modèle mélange disponible en annexe C.6.3 implique quatre composantes pour le mélange. Les impacts des facteurs de risque sont les suivants :

- effets *structurels* : un effet de saisonnalité prononcé (forte baisse des rachats en été, de juillet à septembre), l'effet de l'ancienneté du contrat est un peu différent : les assurés dont l'ancienneté appartient à la tranche la plus basse sont plus susceptibles (fortement) de racheter alors que ceux de la deuxième tranche rachète très sensiblement plus. Plus leur âge augmente et moins les assurés rachètent leur contrat.
- effets *conjoncturels* : l'Ibex 35 et le taux 10Y jouent toujours dans le même sens pour ce type de produit. Leurs effets sont presque de même amplitude (avec un petit plus pour l'Ibex 35). Lorsque l'Ibex et le taux 10Y baissent, la probabilité individuelle de rachat augmente pour une grande majorité des assurés avec des groupes très sensibles

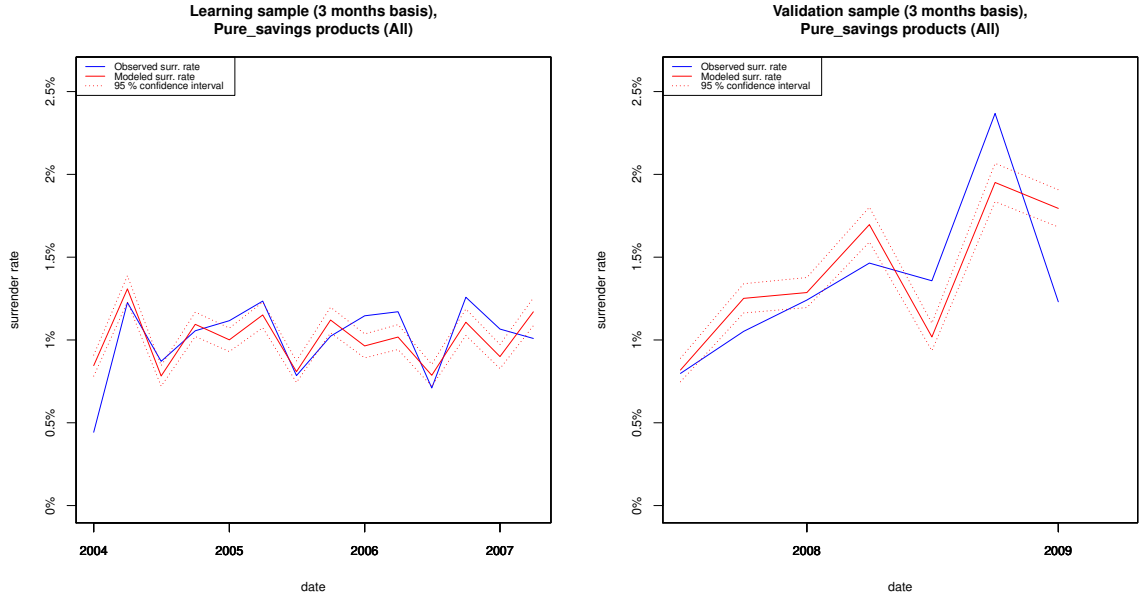
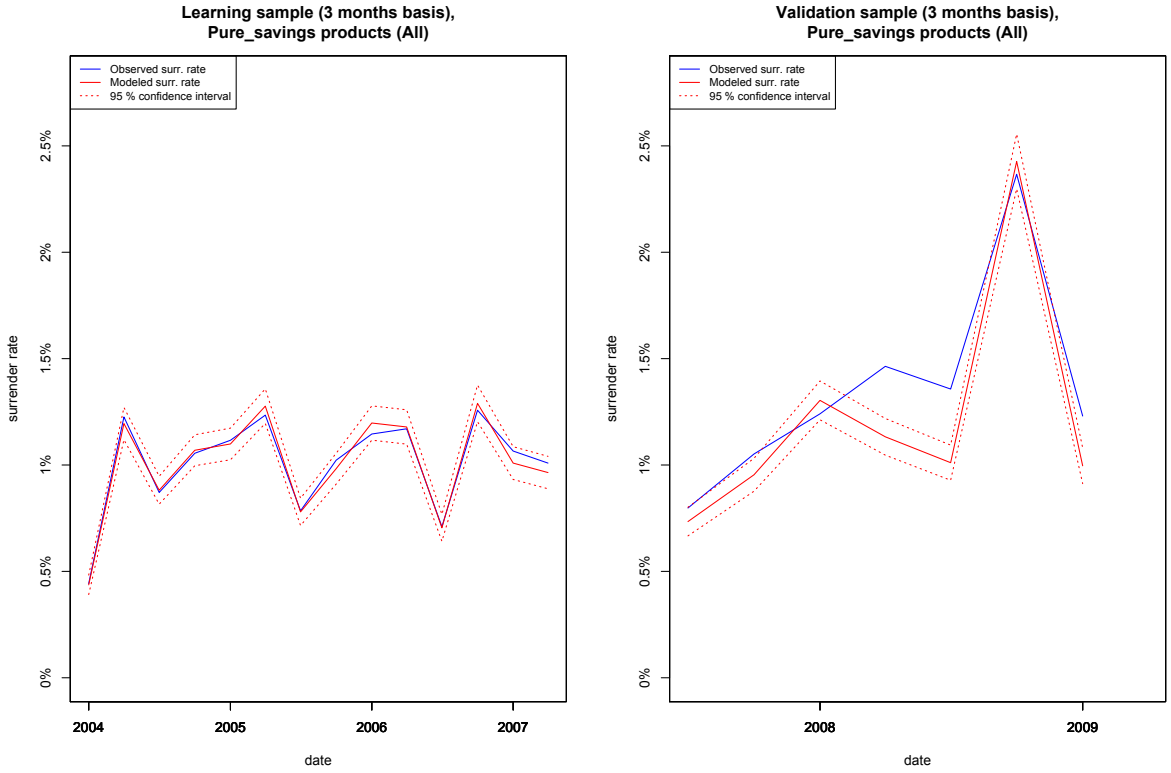


FIGURE 3.18 – Modélisation et prévision du taux de rachat des produits Pure Savings par régression logistique dynamique.

FIGURE 3.19 – Modélisation et prévision des rachats par mélange de Logit, Pure Savings.



(composantes 2 et 4) et d'autre moins sensible (composante 1). Au vu des poids des composantes (annexe, figure C.30), un gros tiers des assurés (36 %) se comportent de manière contraire.

– *corrélation* : introduite via le contexte économique, toutes les personnes d'un groupe donné vont voir leur probabilité de rachat individuelle grimper ou chuter simultanément, faisant ainsi varier les prévisions de rachat au niveau global.

Le fait que les estimations des poids des composantes aient des écart-types importants (0 appartient à l'intervalle de confiance pour chaque estimation) doit partiellement être à l'origine du résultat mitigé que nous obtenons. Il faut toutefois avoir l'honnêteté de dire que nous avons cherché une meilleure modélisation sans pour autant la trouver, ceci est le signe que l'usage des mélanges ici n'est pas forcément pertinent.

3.3.6 Les produits structurés ou "Structured Products".

Il est relativement difficile de décrire les produits stucturés puisque par définition ils sont très variés. Ce sont en général des produits qui dépendent fortement de la performance des marchés financiers, dans le sillage des produits en UC ou des produits indexés sur les indices boursiers. Chaque produit a sa caractéristique, et ce sont en général les proportions d'investissement sur tel ou tel support qui varient entre les produits. La période d'observation s'étend de début 2000 à fin 2009 (avec toujours les rachats enregistrés seulement depuis début 2004) et les informations disponibles sur le contrat et l'assuré sont identiques à celles des Pure Savings

que nous venons d'étudier.

Modélisation et prévisions par mélange de GLM

Les prévisions par régression logistique dynamique du graphique 3.20 laissent à désirer. Autant le modèle performe bien sur la période d'apprentissage, autant il donne des résultats très mauvais sur la période de validation où presque toutes les observations sont en dehors de l'intervalle de confiance (lors du *back-testing*). Le changement de contexte économique qui impacte les comportements de rachat n'est donc pas bien capté par le modèle de régression logistique à une composante. Avec un mélange de régressions logistiques, les prévisions sont nettement meilleures : quatre composantes suffisent à décrire précisément les comportements variables des assurés. Nous avons utilisé une méthode légèrement différente des approches considérées jusqu'ici, qui permet de mieux modéliser la corrélation entre individus. De par la nature du type de produit, la logique voudrait en effet que cette corrélation entre les comportements soit susceptible d'augmenter plus rapidement et plus intensément.

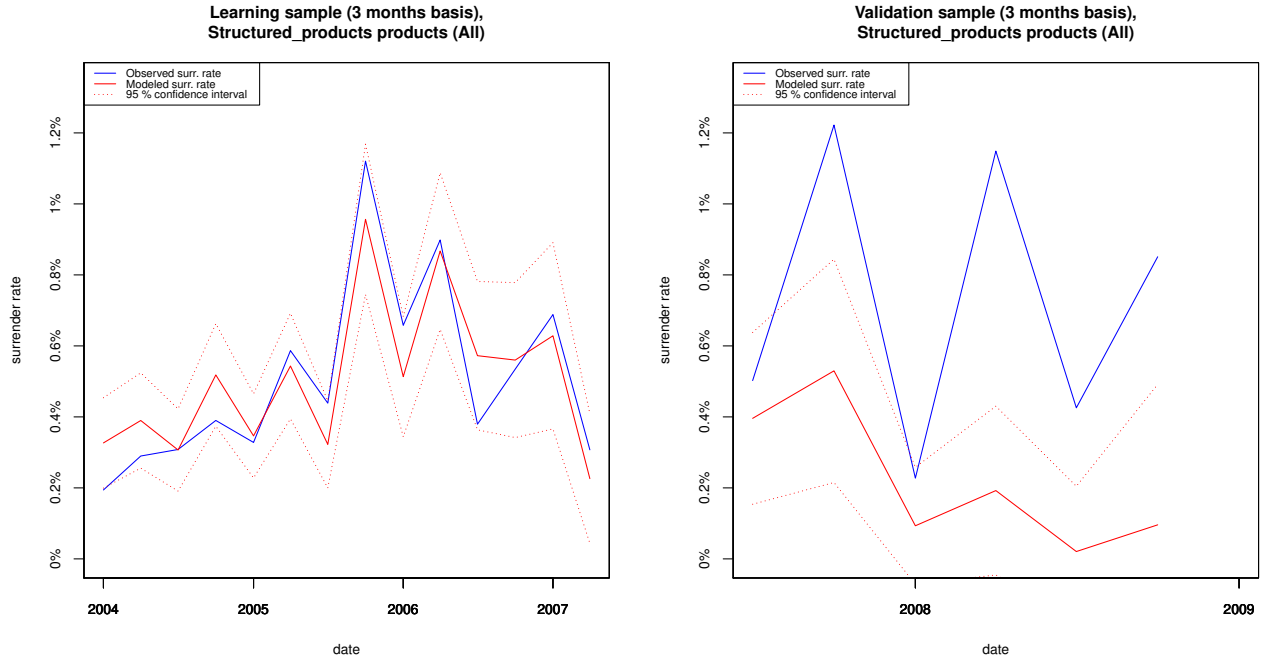
Impact des variables explicatives par les mélanges de Logit

L'usage du modèle mélange est un peu spécifique pour les produits structurés. Nous n'avons introduit aucune variable explicative d'effet structurel (de type ancienneté de contrat, etc) car nous avons remarqué que les prévisions de comportement ne sont finalement pas du tout dictées par ces caractéristiques. Nos seuls facteurs de risque sont ceux liés à l'environnement économique et financier (taux 10Y et ibex 35), mais pris de manière spéciale comme le montrent les "boxplot" des effets de ces variables disponibles en annexe C.7.3 :

- effets *structurels* : aucun input.
- effets *conjoncturels* : introduits via l'Ibex 35 et le taux 10Y. Seul l'Ibex 35 joue sur la probabilité de rachat individuelle des assurés de chaque composante, avec un risque de base ("intercept") comparable entre composantes. Certains rachètent davantage lorsque l'indice croît, reflétant un comportement rationel (composantes 2, 3 et 4) ; alors que les individus appartenant à la composante 1 ont tendance à moins racheter dans un contexte haussier de l'indice.
- *corrélation* : l'originalité de l'approche pour les produits structurés consiste à pouvoir ajuster la taille (proportion) des composantes en fonction du contexte économique. Ceci est réalisé par l'introduction du taux 10 ans en variable explicative des poids des composantes (cf annexe C.37) : si le taux 10 ans augmente alors la probabilité d'appartenir à la composante 4 sera celle qui diminuera le plus, suivie de la composante 2 puis de la composante 1 (et inversement). En revanche plus d'individus adopteront le comportement représenté par la composante 3. Ainsi les assurés sont virtuellement autorisés à changer de composante chaque trimestre suivant l'économie.

Finalement, l'arbitrage du modèle dans la balance de la proportion de personnes adoptant une attitude plutôt rationnelle ou non permet d'obtenir un modèle dont les prévisions sont excellentes. L'aspect dynamique de la taille des composantes est ce qui nous a permis ici de trouver un résultat honorable sur des données relativement spécifiques.

FIGURE 3.20 – Modélisation et prévision des rachats des produits Structurés par régression logistique dynamique.



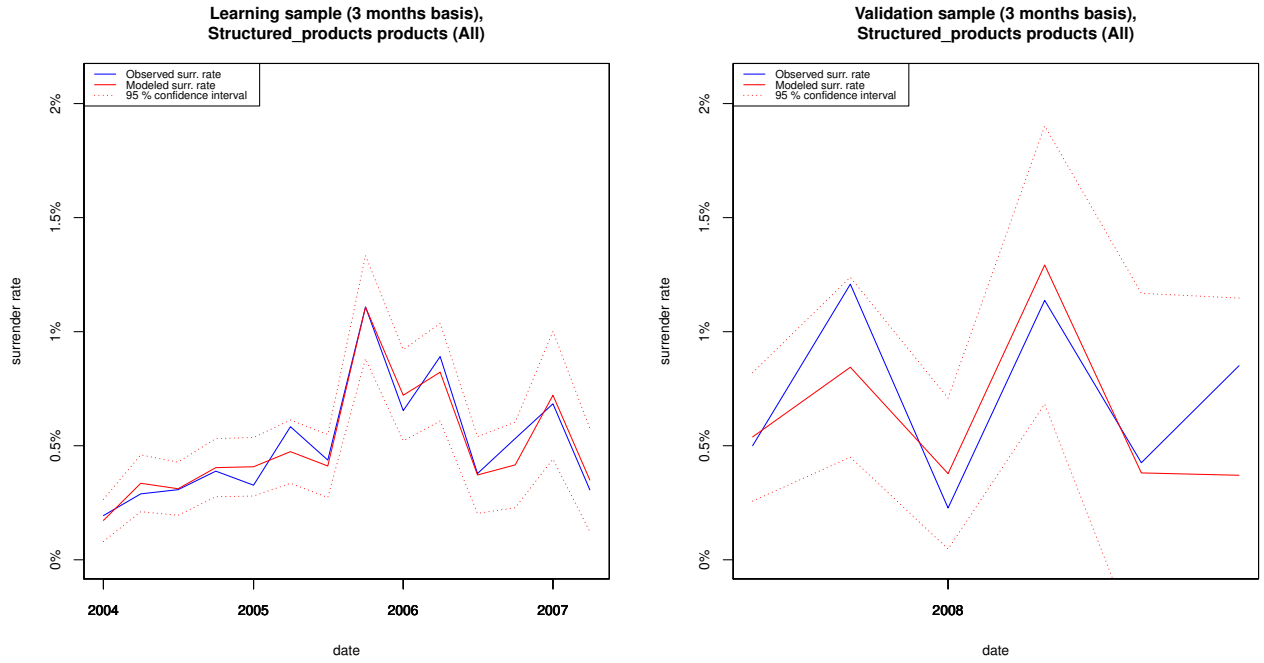


FIGURE 3.21 – Modélisation et prévision des rachats par mélange de Logit, produits Structurés.

3.3.7 Bilan global

Le tableau 3.2 récapitule l'ensemble des résultats des modélisations ci-dessus. Cette représentation permet de faire ressortir des effets similaires sur des types de famille ressemblantes. La confiance que nous pouvons avoir en les estimations est résumée en colonne $Conf.$ (sur une échelle simplifiée allant de 1 à 3), et est basée sur la valeur et l'écart-type du coefficient calibré. Une confiance de 1 signifie que la calibration n'apparait pas robuste ; pour une estimation relativement satisfaisante une confiance de 2 est attribuée et enfin la valeur 3 représente une estimation excellente. La colonne $Nb.Comp.$ concerne le nombre de composantes retenu dans la modélisation, tandis que la qualité globale (toujours sur une échelle de 1 à 3) est évaluée suivant les graphiques de prévision en période de validation et les résultats des tests de Pearson et de Wilcoxon. Nous spécifions en dernière colonne quelques informations supplémentaires, en l'occurence la taille de l'échantillon d'apprentissage, la durée de la période de retour en arrière *delta* et la date de début pour la modélisation. Nous aurions évidemment pu jouer sur la valeur de ces différentes options à des fins d'optimisation des résultats (précision), mais notre but était de montrer que notre modélisation fonctionnait globalement sans avoir à adapter ces paramètres suivant les produits.

Nous tirons de ce tableau quelques conclusions intéressantes, notamment que nous pourrions encore davantage regrouper (si tel était le besoin) les familles qui ont des modélisations proches :

Famille	Covariables poids		Covariables composantes				Nb. comp.	Qualité globale	Remarques (taille apprentissage, lookback period "delta")
			β fixés		β variables				
	Nom	Conf.	Nom	Conf.	Nom	Conf.			
Ahorro	intercept	2	saisonnalité ancienneté fréquence prime âge souscription option PB	3 3 3 3 3	intercept taux 10Y	3 3	5	3	apprentissage : 2/3 *delta* : 1trimestre date début : 1/1/2000
Index-Link	intercept	3	ancienneté âge souscription sexe	3 2 3	intercept ibex 35	3 2	2	3^+	apprentissage : 2/3 *delta* : 1trimestre date début : 1/1/2000
Mixtos	intercept	1	saisonnalité ancienneté option PB	3 3 3	intercept prime risque taux 10Y	3 3 3	5	3^+	apprentissage : 2/3 *delta* : 1trimestre date début : 1/1/2000
Pure Savings	intercept	2	saisonnalité ancienneté âge souscription	2 3 3	intercept ibex 35 taux 10Y	2 3 3	4	2^+	apprentissage : 2/3 *delta* : 1trimestre date début : 1/1/2004
Unit-Link	intercept	2	saisonnalité ancienneté prime de risque	2 3 3	intercept ibex 35	3 3	5	3	apprentissage : 2/3 *delta* : 1 trimestre date début : 1/1/2000
Universal Savings	intercept	3	saisonnalité ancienneté âge souscription réseau distribution	3 3 2 3	intercept richesse	3 3	5	3^+	apprentissage : 2/3 *delta* : 1trimestre date début : 1/1/2004
Structured products	intercept Taux 10Y	1 2			intercept Ibex 35	3 3	4	3	apprentissage : 2/3 *delta* : 1, début : 1/1/2004

TABLE 3.2 – Tableau récapitulatif des modélisations retenues pour chaque famille de produits.

- les contrats à **taux garanti** : grosso modo les “Ahorro”, “Mixtos”, “Pure Savings” et “Universal Savings” admettent le même type de modèle avec :
  - effets *structurels* (fixes entre composantes) dont toujours la saisonnalité et l'ancienneté du contrat, plus une (ou deux) variable(s) dépendante de la famille ;
  - effets *conjoncturels* : guidés par le taux long-terme ;
  - effets de *corrélation* : potentiels si les variables de richesse sont discriminantes.
- les contrats à rendement **non garanti** avec les “Index-Link” et “Unit-Link” :
  - effets *structurels* : pas ou peu de saisonnalité, l'ancienneté du contrat et une variable additionnelle dépendante de la famille ;
  - effets *conjoncturels* : plus intenses, guidés par les marchés financiers (Ibex 35) ;
  - effets de *corrélation* : potentiels si un scénario hyper stressé se réalise.
- les **produits structurés** : la complexité des produits peut expliquer ce comportement plus extrême :
  - effets *structurels* : pas d'effet clair donc inexistant dans la modélisation ;
  - effets *conjoncturels* : intense et dictés par le marché financier ;
  - effets de *corrélation* : introduits via le comportement du marché long-terme avec la taille de la composante risquée qui augmente si le marché se dégrade.

3.4 Conclusion

Nous proposons dans ce chapitre une méthodologie de prise en compte des facteurs de risque. La clef réside dans la distinction entre les effets structurels supposés constants entre groupes homogènes (d'un point de vue comportemental) d'assurés, et les effets conjoncturels qui sont autorisés à varier entre groupes. Cette suggestion provient d'une intuition logique et donne des résultats plus qu'acceptables dans un contexte de contrats d'épargne à supports variés. L'introduction des modèles mélange nous a permis non seulement d'élargir notre champs de connaissance mais aussi d'améliorer la flexibilité de la modélisation en permettant la représentation d'une éventuelle multimodalité de la densité des comportements de rachat, caractéristique d'une forte hétérogénéité. A la vue des résultats numériques, il reste néanmoins un point à élucider sur la question de sélection de modèle : il semble que le critère BIC sélectionne parfois un nombre de composantes dans le mélange qui soit trop important. Nous discutons de ce sujet dans le prochain chapitre, et étendons la problématique à l'ensemble de la famille GLM afin de généraliser notre étude.

Bibliographie

Bohning, D., Dietz, E., Schaub, R., Schlattmann, P. and Lindsay, B. (1994), 'The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family', *Annals of the Institute of Statistical Mathematics* **46**, 373–388. 75

Bohning, D. and Seidel, W. (2003), 'Editorial : recent developments in mixture models', *Computational Statistics and Data Analysis* **41**, 349–357. 76

Box, G. and Cox, D. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society B* **26**, 211–252. 69

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton. 75

Dempster, A., N.M., L. and D.B., R. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society* **39**, 1–38. 73

Follmann, D. and Lambert, D. (1989), 'Generalizing logistic regression by non-parametric mixing', *Journal of the American Statistical Association* **84**, 295–300. 76

Garel, B. (2007), 'Recent asymptotic results in testing for mixtures', *Computational Statistics and Data Analysis* **51**, 5295–5304. 76

Ghosh, J. and Sen, P. (1985), 'On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results', **2**, 789–806. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer. 75

Lindsay, B. and Lesperance, M. (1995), 'A review of semiparametric mixture models', *Journal of Statistical Planning and Inference* **47**, 29–99. 76

Lindstrom, M. and Bates, D. (1988), 'Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data', *Journal of the American Statistical Association* **83**, 1014–1022. 75

Margolin, B., Kim, B. and Risko, K. (1989), 'The ames salmonella/microsome mutagenicity assay : issues of inference and validation', *Journal of the American Statistical Association* **84**, 651–661. 77

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, Wiley Series In Porbability and Statistics. 70, 71, 75

Pearson, K. (1894), 'Contributions to the theory of mathematical evolution', *Philosophical Transactions of the Royal Society of London A* **185**, 71–110. 69

Teicher, H. (1963), 'Identifiability of finite mixtures', *Annals of Mathematical Statistics* **34**, 1265–1269. 77

Wang, P. (1994), Mixed Regression Models for Discrete Data, PhD thesis, University of British Columbia, Vancouver. 76

Wolfe, J. (1971), A monte carlo study of sampling distribution of the likelihood ratio for mixtures of multinormal distributions, Technical Report STB-72-2, San Diego : U.S. Naval Personnel and Training Research Laboratory. 75

Chapitre 4

Sélection de mélange de GLMs

Ce chapitre correspond aux derniers développements et aspire à une future publication dans une revue de statistiques appliquées.

Nous avons vu dans le chapitre précédent un moyen de modéliser précisément les comportements de rachat. Il découle de cette modélisation des questions légitimes quant à l'interprétation des résultats finaux. En effet, une problématique classique en modélisation comportementale est de pouvoir remonter au groupe d'appartenance de chaque individu à partir de la partition finale établie. Ce n'est pas tant l'affectation d'une observation à un groupe qui pose problème, mais plutôt l'interprétation qui s'ensuit : en quoi tel groupe diffère-t-il de son voisin, qu'est-ce qui le caractérise ? L'affectation de l'individu j à un groupe est réalisée par la règle de Bayes ou du maximum-a-posteriori (MAP), donnée en reprenant les notations des chapitres 1 et 3 par :

$$\hat{z}_j^{MAP}(\hat{\psi}^{MLE}) = \underset{i=1,\dots,G}{\arg\max}\, \tau_i(y_j; \hat{\psi}^{MLE}).$$

Ainsi, l'individu j est logiquement affecté au groupe dont il maximise la probabilité a posteriori d'appartenance, calculée à partir de l'estimateur du maximum de vraisemblance. Dans notre cas, les groupes (classes) formés sont représentés par les composantes des mélanges de régressions logistiques. Autrement dit, nous effectuons une classification non supervisée de nos assurés à partir de notre modèle : derrière cette technique couramment appelée "model-based clustering" se cache l'idée que chaque assuré est issu d'une classe, et que l'estimation du mélange permet de remonter à ces différentes classes que nous n'observons pas. Lorsque nous regardons les valeurs numériques associées à l'estimation de chaque composante de ces mélanges, il arrive fréquemment que :
i) certaines composantes se ressemblent fortement,
ii) la variance des coefficients estimés soit grande.
La ressemblance implique que les assurés ayant été classés dans des composantes presque similaires aient des réactions quasi-identiques. Nous pouvons dès lors nous poser la question de la pertinence de la modélisation trouvée : est-ce bien la "meilleure" représentation de la réalité ? Avoir moins de groupes nous permettrait-il d'améliorer la robustesse de nos estimations ?

Nous verrons par la suite que "meilleure" doit s'interpréter selon un certain critère de choix, lequel doit permettre de répondre au mieux à la problématique originelle. En l'occurrence nos problématiques sont d'être capable de bien approcher la loi des données observées (afin d'effectuer des prévisions robustes) et de correctement segmenter notre portefeuille d'assurance. Nous supposons dans toute la suite que les données correspondent à l'échantillonnage indépendant

d'une loi théorique de densité f^0 (inconnue). Le critère BIC, que nous utilisons au chapitre 3 comme critère de sélection de modèle, est connu pour ses propriétés de convergence vers cette loi théorique. Cependant seul un de nos deux objectifs est visé par ce critère, notre étude suggérant d'autre part une éventuelle surestimation du nombre de composantes des mélanges dans certains cas (ex. : les composantes 2, 4 et 5 pour les produits Ahorro, annexe C.2.4). Dans l'optique où nous voudrions comprendre quels sont les grands types de réaction des assurés face à un changement de contexte économique ou une modification contractuelle, nous devons les classer selon des comportements bien différenciés. L'idée d'utiliser les modèles mélanges comme outil pour la classification non-supervisée n'est pas nouvelle, et certains des avantages de cette méthode sont notamment résumés dans Biernacki (2009). Afin de tirer au mieux parti de ce type de modélisation, il est nécessaire de s'attarder sur la question de sélection de modèle et donc indirectement sur le choix du nombre de composantes du mélange.

Après quelques rappels sur les notions essentielles de la théorie de l'information et du maximum de vraisemblance, ce chapitre développe l'étude théorique et pratique du critère de sélection ICL (Integrated Classification Likelihood) dans le cadre des mélanges de modèles linéaires généralisés (GLM). Ce critère, qui semble avoir été introduit à la fin des années 1990 dans l'article de Biernacki and Govaert (1997), se révèle être particulièrement bien adapté à la classification par modèle mélange. En particulier, nous démontrons les propriétés de convergence du critère ICL pour des mélanges de GLM sous certaines conditions. Pour ce faire nous présentons un nouvel estimateur, le maximum de vraisemblance classifiante conditionnelle, défini pour la première fois dans Baudry et al. (2008). La régression logistique faisant partie des modèles GLM, nos résultats théoriques seront directement applicables dans le contexte de notre étude opérationnelle.

4.1 Théorie de l'information et sélection de modèle

La question de sélection de modèle est un problème classique de statistique qui a été largement étudié par la communauté scientifique. Ce problème vient du fait que l'inférence pour l'estimation d'un modèle paramètrique nous amène très fréquemment à considérer non pas un modèle en particulier mais un ensemble de modèles. Naturellement, l'étape suivante consiste à faire un choix parmi ces modèles sur la base d'une argumentation rigoureuse, basée dans la littérature sur la théorie de l'information. Les statisticiens ont principalement développé des méthodes de minimisation de critère d'information pénalisé : c'est ainsi qu'apparaissent parmi bien d'autres les plus célèbres AIC (Akaike Information Criterion, Akaike (1973)), C_p de Mallows dans le contexte de la régression par moindres carrés (Mallows (1974)), et BIC (Bayesian Information Criterion, Schwarz (1978)) dans un contexte bayésien. Ces critères, très largement diffusés, font toutefois appel à des hypothèses et des justifications théoriques bien souvent omises dans la plupart des applications. Nous nous proposons donc de reformaliser leur construction et le contexte de leur utilisation, basée sur l'estimation par maximum de vraisemblance. La compréhension des notions phares qu'ils sous-tendent sera un élément-clef de la définition et de l'étude théorique du nouveau critère développé en section 4.2.3.

4.1.1 Distance de Kullback-Leibler

Il y a un bon demi siècle, Kullback and Leibler (1951) introduisent la distance de Kullback-Leibler (notée distance KL dans la suite) comme mesure de la proximité entre deux distributions de probabilité. Ils s'intéressent à la question de discrimination statistique entre deux

échantillons, et la motivation de leur travail est de définir une "distance" ou "divergence" entre deux populations statistiques en termes de mesure d'information. Pour cela, les auteurs supposent donnés deux espaces probabilisés $(Y, \mathcal{Y}, \nu_i)$, i = 1, 2 ; tels qu'il existe une mesure de probabilité dominante notée λ. D'après le théorème de Radom-Nikodym, il existe des densités uniques f_i, i = 1, 2, λ-mesurables avec $0 < f_i(y) < \infty$ $[\lambda]$, telles que pour i = 1, 2,

$$\nu_i(E) = \int_E f_i(y) d\lambda(y), \quad \forall E \in \mathcal{Y}.$$

En notant H_i (i = 1, 2) l'hypothèse selon laquelle l'observation y est issue de la population dont la mesure de probabilité est ν_i, ils définissent "$\ln \dfrac{f_1(y)}{f_2(y)}$" comme l'*information* de y pour la discrimination entre H_1 et H_2. C'est ainsi qu'est créée la "distance" KL, résumé de l'information moyenne d'une observation pour la discrimination entre H_1 et H_2 :

$$d_{KL} = I(1:2) = I_{1:2}(Y) = \int d\nu_1(y) \ln \frac{f_1(y)}{f_2(y)} = \int f_1(y) \ln \frac{f_1(y)}{f_2(y)} d\lambda(y).$$

Kullback and Leibler (1951) soulignent également le lien entre leur mesure et l'information de Fisher par l'introduction de la "divergence". En réalité cette mesure de Kullback-Leibler n'est pas une distance (elle ne satisfait pas toutes les propriétés d'une distance), mais elle permet de mesurer la différence d'information entre deux populations et a d'intéressantes propriétés. Notamment,

Lemme 1. *$I(1:2)$ est presque partout définie positive, donc $I(1:2) \geq 0$ avec égalité si et seulement si $f_1(y) = f_2(y)$ $[\lambda]$.*

Démonstration. La preuve est consultable dans l'article d'origine, Kullback and Leibler (1951). □

Dans les paragraphes à venir, nous montrons l'importance du lien entre la distance KL et la théorie du maximum de vraisemblance, qui repose finalement sur cette mesure de l'information. Beaucoup d'autres travaux utilisent également cette théorie, dont le critère de sélection de modèle AIC que nous détailllons au 4.1.3.

4.1.2 Estimateur du maximum de vraisemblance (MLE)

L'estimation par maximum de vraisemblance repose sur certaines justifications dont il est indispensable d'avoir conscience. Ces justifications, souvent écartées par ses très nombreux utilisateurs, s'imposent pourtant comme la base théorique à de nouveaux développements potentiels. Nous proposons ici de revenir sur les conditions qui garantissent dans un cadre paramétrique la convergence de l'estimateur du maximum de vraisemblance vers le paramètre théorique à estimer. Notre objectif est de se familiariser avec ces notions, avant de présenter la définition d'un nouvel estimateur qui se révéle plus adapté dans notre contexte, et dont les propriétés de convergence sont prouvées dans le champ de notre étude.

Notations

Nous notons dans le reste du chapitre $f(y;\psi)$ la densité de la famille paramétrique considérée pour modéliser les données Y. Notons $f^0(y)$ la densité théorique de Y, inconnue et

correspondant au modèle paramétrique sous-jacent à Y. Nous y associons ψ^0 son paramètre théorique. Lorsque le modèle est correctement spécifié (la densité théorique appartient à la famille paramétrique étudiée), $f^0(y) = f(y;\psi^0)$. Dans le cas contraire, on dira que $f(y;\psi^0)$ est le quasi-vrai modèle.

Convergence de l'estimateur MLE

A la suite des travaux de Doob (1934) et Cramér (1946), Wald (1949) étudie la convergence de l'estimateur du maximum de vraisemblance pour des distributions de probabilité dépendant d'un unique paramètre. Il formule huit hypothèses (en fait 7 car la dernière est une implication d'une des sept premières) qui permettent d'aboutir aux propriétés de convergence de cet estimateur vers le paramètre théorique de la distribution des données. Néanmoins l'ensemble de ses résultats reposent sur l'hypothèse que le modèle a été correctement spécifié. C'est le point de départ d'une période de recherche intense dans ce domaine, qui aboutit notamment aux travaux de Redner (1981) et Nishii (1988) dans le contexte des mélanges. Le premier cité étend les résultats de Wald (1949) au cas où la distribution théorique des données est représentée par plus d'un paramètre, incluant ainsi les distributions qui souffrent du problème d'identifiabilité. Il prend l'exemple des modèles mélange et démontre la convergence de l'estimateur du maximum de vraisemblance lorsque l'espace des paramètres est supposé compact, ce qui permet de garantir l'existence d'un tel estimateur. Redner suppose à quelques différences près les mêmes hypothèses que Wald (1949) (sans les hypothèses 1 et 8 mais en en introduisant une nouvelle).
Nishii (1988) se concentre sur le fait que le modèle théorique est inconnu de l'observateur, ce qui augmente considérablement les chances de mauvaise spécification du modèle. Dans un tel cadre, a-t-on toujours les mêmes propriétés ? Il montre la convergence forte de l'estimateur du maximum de vraisemblance vers le paramètre théorique en montrant que maximiser la vraisemblance revient à minimiser la distance KL entre la quasi-vraie distribution et la famille paramétrique considérée (voir son exemple p. 393-394). Nishii (1988) suppose pour cela que l'espace des paramètres est convexe et adopte une approche différente de Wald (1949). Dans sa preuve, il formule essentiellement des hypothèses sur la dérivabilité et l'intégrabilité de la fonction de vraisemblance, ce qui lui permet de prouver cette convergence par un développement de Taylor. Nous verrons qu'un parallèle évident pourra être fait entre les hypothèses que nous formulerons pour nos résultats et celles de Nishii (1988) (section 2).

4.1.3 Critères de sélection pénalisés

Rappelons que $Y = (Y_1, Y_2, ..., Y_n)$ est un échantillon de n variables aléatoires continues indépendantes, de densité inconnue f^0. Nous désirons estimer f^0, et disposons d'un ensemble fini de m modèles au choix $\{M_1, M_2, ..., M_m\}$. L'objectif d'un critère de sélection est donc de trouver le meilleur modèle parmi cet ensemble. Nous nous plaçons dans un cadre paramétrique (comme c'est le cas tout au long de la thèse) où chaque modèle M_g ($g \in [\![1, m]\!]$) de dimension K_g (nombre de paramètres libres) correspond à une densité f_{M_g}, avec pour paramètre ψ_g. Soit Ψ_g l'espace de dimension K_g, où $\psi_g \in \Psi_g$.

Un critère de vraisemblance pénalisée empêche la sélection d'un modèle sur la seule valeur de la vraisemblance obtenue. En effet, celle-ci s'accroît logiquement lorsque le modèle se complexifie, notamment lorsque les modèles considérés sont emboîtés (un mélange à deux composantes est "emboîté" dans un mélange à trois composantes, une régression à p covariables

est "emboîtée" dans une régression à q covariables si $p < q$, ...). Il n'est donc pas optimal de choisir son modèle sur la seule valeur de la vraisemblance puisque le modèle sélectionné sera systématiquement le plus complexe. Par conséquent, un critère de vraisemblance pénalisée classique est toujours de la forme

$$IC(K_g) = -\ln L(\hat{\psi}_g^{MLE}) + pen(K_g),$$

où K_g est la dimension du modèle (sa complexité), $\hat{\psi}_g^{MLE}$ est l'estimateur du maximum de vraisemblance, et $pen(K_g)$ la pénalité. Dans ce contexte, nous cherchons le modèle donné par

$$\hat{K} = \underset{K_g \in \{K_1,...,K_m\}}{\arg\min} IC(K_g).$$

Conditions suffisantes de convergence

Nishii (1988), toujours dans le même article, déroule des propriétés de convergence du modèle $\hat{M}$ sélectionné (par ce type de critère d'information pénalisé) vers le modèle théorique M^0 de Y. Pour cela, il suffit que la pénalité du critère satisfasse les deux conditions énoncées dans le théorème qui suit.

Théorème 3. *(Nishii). Soit $\hat{M}$ un modèle sélectionné via un critère d'information de type $IC(K_g) = -2\ln L(\hat{\psi}_g) + c_n K_g$, où $\hat{\psi}_g$, $L(\hat{\psi}_g)$ et K_g sont respectivement l'estimateur du quasi-maximum de vraisemblance, la quasi-vraisemblance et la dimension du modèle M_g. Ainsi $\hat{M}$ minimise*

$$IC(K_g) = -\ln L(\hat{\psi}_g) + c_n K_g,$$

pour n observations et des modèles emboîtés $M_g = \{M_1, ..., M_m\}$ de dimension $K_g = \{K_1, ..., K_m\}$ (par définition, un modèle est dit emboîté dans un autre s'il est un cas particulier de cet autre modèle plus général). Si c_n satisfait les deux conditions

$$\lim_{n\to\infty} \frac{1}{n} c_n = 0 \quad et \quad \lim_{n\to\infty} \frac{c_n}{\ln\ln n} = +\infty,$$

Alors $\hat{M}$ converge fortement vers le quasi-vrai modèle M^0, donc $\lim_{n\to\infty} \hat{M} = M^0$ p.s.

La première condition garantit que le modèle sélectionné ne sous-estime pas le nombre théorique de composantes, tandis que la deuxième sert à limiter la surestimation de ce même nombre de composantes. En effet "c_n négligeable devant n" permet de ne pas trop pénaliser la vraisemblance quand n grandit, alors que "$\ln\ln n$ négligeable devant c_n" provoque l'effet inverse : il faut donc trouver un juste milieu dans le poids de la pénalité. Nishii montre également des résultats de convergence faible en relaxant la deuxième hypothèse sur c_n.
Nishii déduit de son étude que le critère AIC n'est pas consistant, alors que le critère BIC l'est fortement. Nous présentons ci-dessous la construction de ces deux critères de vraisemblance pénalisée afin de visualiser concrètement les raisons de leur convergence (ou non), ainsi que leurs spécificités.

Le critère AIC

Nous nous appuyons dans ce paragraphe sur l'article d'origine, Akaike (1973), tout en reprenant les notations introduites depuis le début du chapitre. Dans son papier Akaike relève

la nécessité de développer de nouvelles méthodes de sélection de modèle dans un contexte global, par opposition aux récents développements de l'époque qui ne s'appliquaient qu'à une certaine classe de modèles. Il introduit la notion de critère d'information, en ce sens que sa démarche est complètement liée à l'étude de la distance KL. Le succès d'Akaike (1973) vient notamment du pont établi entre la théorie du maximum de vraisemblance, base théorique largement reconnue par les statisticiens pour l'estimation paramétrique, et le critère AIC. De plus ce critère ne nécessite pas de calcul supplémentaire en dehors des calculs inhérents à la méthode du maximum de vraisemblance, ce qui est un avantage non-négligeable au vu des performances calculatoires des outils informatiques de l'époque.
L'idée novatrice d'Akaike est de choisir comme estimateur final parmi un ensemble d'estimateurs $\hat{\psi}$ (pour une densité de probabilité $f(y;\psi)$) celui qui maximise la log-vraisemblance **espérée**. Autrement dit, il cherche

$$\begin{aligned} M_{AIC} &= \underset{M_g\in\{M_1,\dots,M_m\}}{\arg\max} \left(\mathbb{E}_{(Y,\hat{\psi_g})}\left[\ln f_{M_g}(Y;\hat{\psi_g})\right]\right) \\ &= \underset{M_g\in\{M_1,\dots,M_m\}}{\arg\max} \left(\mathbb{E}_{\hat{\psi_g}}\left[\int_{\mathcal{Y}} \ln f_{M_g}(Y;\hat{\psi_g}) f(Y;\psi^0)dy\right]\right). \end{aligned} \tag{4.1}$$

L'égalité ci-dessus vient de l'hypothèse d'indépendance entre les lois de Y et de $\hat{\psi_g}$, ce qui permet d'avoir

$$f_{(Y,\hat{\psi_g})}(y,\hat{\psi_g}) = f_Y(y;\psi^0) f_{\hat{\psi_g}}(\hat{\psi_g}).$$

L'auteur différencie sa méthode de celle du maximum de vraisemblance en justifiant du fait que cette dernière ne s'intéresse à l'estimation du paramètre ψ_g du modèle M_g que pour une réalisation donnée des observations : ainsi $\hat{\psi_g}(Z)$ maximise $\ln f_{M_g}(z;\psi_g)$ pour **une seule** réalisation de Z (clairement $\hat{\psi_g}$ est une statistique de Z). Ainsi, la méthode du maximum de vraisemblance ne nécessite aucune connaissance sur ψ^0, le paramètre théorique de la densité de la loi de Y.

Pour comparer sans perte d'efficacité un modèle général donné par $f_{M_g}(.;\psi_g)$ avec le modèle théorique $f(.;\psi^0)$, il utilise le célèbre ratio de vraisemblance $\tau(y) = f_{M_g}(y;\psi_g)/f(y;\psi^0)$ qui détermine par l'introduction d'une fonction Φ la discrimination en y entre ψ_g et ψ^0. Nous en déduisons immédiatement la discrimination moyenne dans le cas où ψ^0 est "vrai" (donc Y a effectivement pour densité $f(y;\psi^0)$) :

$$D(\psi_g,\psi^0,\Phi) = \int_{\mathcal{Y}} f(y;\psi^0)\Phi(\tau(y))dy = \mathbb{E}_Y[\Phi(\tau(Y))].$$

Comment choisir Φ pour définir cette discrimination moyenne ? En effectuant un développement de Taylor à l'ordre 2 de la fonction composée $\Phi(\tau(y))$ pour ψ_g au voisinage de ψ^0, nous obtenons sous certaines conditions de régularité de Φ et en remarquant que le terme d'ordre 1 s'annule (la vraisemblance est maximisée en ψ^0 donc sa dérivée en ce point vaut 0) :

$$\begin{aligned} D(\psi_g,\psi^0,\Phi) &= \mathbb{E}_Y\left[\Phi(1) + \frac{1}{2}\Phi''(1)(\psi_g-\psi^0)^T I(\psi^0)(\psi_g-\psi^0) + o(\|\psi_g-\psi^0\|^2)\right], \\ &= \Phi(1) + \frac{1}{2}\Phi''(1)(\psi_g-\psi^0)^T J(\psi^0)(\psi_g-\psi^0) + o(\|\psi_g-\psi^0\|^2). \end{aligned} \tag{4.2}$$

où

$$J(\psi^0) = \int_{\mathcal{Y}} \left[\left(\frac{\partial \ln f_{M_g}(y;\psi_g)}{\partial\psi_g}\right)_{\psi_g=\psi^0} \left(\frac{\partial \ln f_{M_g}(y;\psi_g)}{\partial\psi_g}\right)^T_{\psi_g=\psi^0}\right] f(y;\psi^0)dy.$$

En effet pour ψ_g au voisinage de ψ^0, $\tau(y)$ se trouve au voisinage de 1 ; et le terme d'ordre 1 disparait. $J(\psi^0)$ n'est rien d'autre que la matrice d'information de Fisher en ψ^0, obtenue par passage à l'espérance (rappelons qu'elle s'exprime également $\mathbb{E}_Y\left[\left(\frac{\partial}{\partial\psi}\ln L(\psi;Y)\right)^2_{\psi=\psi^0}\right]$).
Pour que la discrimination se comporte comme une distance entre ψ_g et ψ^0, il faut que $\Phi(1)=0$ et que $\Phi''(1)>0$. Cette remarque amène l'auteur à choisir de manière arbitraire $\Phi(x) = -2\ln(x)$. Par ce choix il retombe (à un facteur 2 près) sur la distance KL, appelée aussi *négentropie* :

$$\begin{aligned} D(\psi_g,\psi^0) &= 2\int_{\mathcal{Y}} f(y;\psi^0)\ln\frac{f(y;\psi^0)}{f_{M_g}(y;\psi_g)}dy \\ &= 2\left(\mathbb{E}_Y[\ln f(Y;\psi^0)] - \mathbb{E}_Y[\ln f_{M_g}(Y;\psi_g)]\right) = 2\,d_{KL}(f^0, f_{M_g}). \end{aligned} \tag{4.3}$$

De plus, nous retrouvons l'objectif initial dans le deuxième terme en passant à l'espérance dans (4.3) et en considérant l'estimateur du maximum de vraisemblance pour ψ_g, dans le cas où Y et Z sont indépendantes :

$$\begin{aligned} \mathbb{E}_Z[D(\hat{\psi}_g(Z),\psi^0)] &= \mathbb{E}_Z\left[2\left(\mathbb{E}_Y[\ln f(Y;\psi^0)] - \mathbb{E}_Y[\ln f_{M_g}(Y;\hat{\psi}_g(Z))]\right)\right] \\ &= 2\,\mathbb{E}_Y[\ln f(Y;\psi^0)] - 2\underbrace{\mathbb{E}_{(Y,Z)}[\ln f_{M_g}(Y;\hat{\psi}_g(Z))]}_{(4.1)}, \qquad (4.4) \\ &= \mathbb{E}_Z\left[2\,d_{KL}\left(f(y;\psi^0), f_{M_g}(y;\hat{\psi}_g(Z))\right)\right]. \qquad (4.5) \end{aligned}$$

Nous appelons cette quantité la *négentropie probabilisée.*
Par conséquent, Akaike en déduit que **maximiser la log-vraisemblance espérée n'est donc rien d'autre que minimiser l'espérance de la distance KL** entre la densité estimée $f_{M_g}(.;\hat{\psi}_g(Z))$ et la densité théorique $f(.;\psi^0)$. De plus il remarque que dans le cas de n observations indépendantes, la fonction Φ choisie conserve la propriété d'additivité :

$$D_n(\psi_g,\psi^0) = nD(\psi_g,\psi^0).$$

Afin d'évaluer l'adéquation du modèle, Akaike se base sur le principe de maximisation de l'entropie provenant de la théorie des grands échantillons. Celui-ci préconise d'étudier la *negentropie probabilisée* en le paramètre théorique :

$$R(\psi^0) = \mathbb{E}_Z[D(\hat{\psi}^0(Z),\psi^0)] = 2\mathbb{E}_Y[\ln f(Y;\psi^0)] - 2\mathbb{E}_{(Y,Z)}[\ln f_{M_g}(Y;\hat{\psi}^0(Z))].$$

Remarquer que $\hat{\psi}^0(Z)$ remplace $\hat{\psi}_g(Z)$ dans (4.4). Finalement, le modèle sélectionné sera donc celui dont la valeur $R(\psi^0)$ sera la plus petite. Cependant quelques problèmes subsistent : ψ^0 est inconnu, de même que l'espérance sur Z ; cela nous empêche de calculer la valeur minimale de $R(\psi^0)$. Pour résoudre ce problème, Akaike utilise alors la loi faible des grands nombres sur n observations indépendantes :

$$\hat{D}_n(\psi_g,\psi^0) = \frac{2}{n}\sum_{k=1}^{n}\ln\frac{f(y_k;\psi^0)}{f_{M_g}(y_k;\psi_g)} \xrightarrow{\mathbb{P}} D(\psi_g,\psi^0) = 2(\mathbb{E}_Y[\ln f(Y;\psi^0)] - \mathbb{E}_Y[\ln f_{M_g}(Y;\psi_g)]).$$

Sous certaines conditions de régularité de la densité f, cette convergence simple devient uniforme (donc convergence du $\sup_{\psi\in\Psi}$). Ceci garantit que l'estimation du maximum du rapport

moyen des log-vraisemblances (autrement dit l'estimation du maximum de vraisemblance) converge en probabilité vers l'estimation du maximum d'entropie (ou minimum de négentropie, donc minimum en distance KL).

Ne disposant que de $\hat{\psi}_g(Z)$ comme estimation du maximum de vraisemblance dans le modèle M_g, Akaike choisit d'approcher

$$R(\psi^0) = \mathbb{E}_Z[D(\hat{\psi}^0(Z), \psi^0)] \quad \text{par} \quad \mathbb{E}_Z[D(\hat{\psi}_g(Z), \psi^0)].$$

Il reste à déterminer le biais introduit par cette approximation, et le fait d'utiliser les mêmes données pour estimer le maximum de vraisemblance qui sert ensuite à évaluer la "distance" $D(\hat{\psi}_g(Z), \psi^0)$. D'après (4.4) et (4.5), nous avons

$$\mathbb{E}_Z\left[2\, d_{KL}\left(f(y;\psi^0), f_{M_g}(y;\hat{\psi}_g(Z))\right)\right] = 2\,\mathbb{E}_Y[\ln f(Y;\psi^0)] - 2\,\mathbb{E}_{(Y,Z)}[\ln f_{M_g}(Y;\hat{\psi}_g(Z))]$$

Donc

$$\mathbb{E}_{(Y,Z)}[\ln f_{M_g}(Y;\hat{\psi}_g(Z))] = \overbrace{\mathbb{E}_Y[\ln f(Y;\psi^0)]}^{\text{indép. du modèle}} - \mathbb{E}_Z\left[d_{KL}\left(f(y;\psi^0), f_{M_g}(y;\hat{\psi}_g(Z))\right)\right]$$

D'où

$$\frac{1}{n}\ln L(\hat{\psi}_g(Z)) = \mathbb{E}_Y[\ln f(Y;\psi^0)] - \overbrace{\mathbb{E}_Z\left[d_{KL}\left(f(y;\psi^0), f_{M_g}(y;\hat{\psi}_g(Z))\right)\right]}^{1/2\,\mathbb{E}_Z[D(\hat{\psi}_g(Z),\psi^0)]}.$$

Il faut donc évaluer le biais donné par

$$B(K_g) = \mathbb{E}_Z\left[\frac{1}{n}\ln L(\hat{\psi}_g(Z);Y) - \int_{\mathcal{Y}} f(y;\psi^0)\ln f_{M_g}(y;\hat{\psi}_g(Z))dy\right]. \tag{4.6}$$

Pour estimer ce biais, il définit une norme ($\|.\|_0$) et un produit scalaire ($<.,.>_0$) à partir de l'information de Fisher $J(\psi^0)$, via l'approximation quadratique de $D(\psi_g, \psi^0, \Phi)$ par $W(\psi_g, \psi^0)$ obtenue en utilisant la formule de Taylor (avec $\Phi(1) = -2\ln(1) = 0$) :

$$W(\psi_g, \psi^0) = (\psi_g - \psi^0)^T J(\psi^0)(\psi_g - \psi^0).$$

Ainsi pour un modèle M_g, il obtient par Pythagore :

$$\begin{aligned} D(\hat{\psi}_g(Z), \psi^0) \simeq W(\hat{\psi}_g(Z), \psi^0) &= \|\hat{\psi}_g(Z) - \psi^0\|_0^2 \\ &= \|\psi^{0|K_g} - \psi^0\|_0^2 + \|\hat{\psi}_g(Z) - \psi^{0|K_g}\|_0^2; \end{aligned}$$

avec $\psi^{0|K_g}$ la projection de ψ^0 sur Ψ_{K_g}, la métrique de l'information.
Pour estimer $R(\psi^0)$, il utilise donc $\mathbb{E}_Z[W(\hat{\psi}_g(Z), \psi^0)]$. Finalement après évaluation des différents termes, il trouve

$$n\mathbb{E}_Z[W(\hat{\psi}_g(Z), \psi^0)] \;\simeq\; n\hat{D}_n(\hat{\psi}_g(Z), \hat{\psi}^0(Z)) + 2\underbrace{K_g}_{\simeq B(K_g)} - K_m,$$

où K_g est la dimension du modèle M_g étudié, K_m la dimension maximale. Cette expression ayant K_m identique pour tous les sous-modèles considérés, la pénalité retenue est $2K_g$. Finalement, le critère AIC est plus connu sous la forme analogue

$$AIC_g = -2\ln(f_{M_g}(Y, \hat{\psi}_g)) + 2K_g,$$

et le modèle sélectionné satisfait :

$$M_{AIC} = \underset{M_g \in \{M_1,\ldots,M_m\}}{\arg\min} AIC_g.$$

Le critère BIC

La plupart des publications utilisent le critère BIC comme critère de sélection de modèle. Dans le domaine de la médecine par exemple, Mun et al. (2008) sélectionnent par BIC un mélange gaussien multivarié pour modéliser le risque d'une prise abusive d'alcool en fonction de certains facteurs environnementaux. Nous suivons dans ce paragraphe les excellentes présentations de Raftery (1994) et Lebarbier and Mary-Huard (2004) de l'article à l'origine du critère BIC (Schwarz (1978)).

Dans un contexte bayésien nous considérons les modèles M_g et les paramètres ψ_g comme des variables aléatoires. Ils admettent donc des distributions *a priori*, respectivement $P(M_g)$ et $P(\psi_g|M_g)$, ce qui serait utile pour intégrer des informations particulières que nous connaîtrions au préalable (bien que souvent $P(M_g)$ soit non-informative, c'est à dire uniforme). De toute façon cette information n'apparaît pas dans la composition finale du critère BIC, pour des raisons d'approximation asymptotique.
Le modèle M_g sélectionné par BIC maximise la probabilité *a posteriori* $P(M_g|Y)$, d'où :

$$M_{BIC} = \underset{M_g \in \{M_1,\dots,M_m\}}{\arg\max} P(M_g|Y).$$

BIC cherche donc à sélectionner le modèle le plus probable au vu des données. D'après la formule de Bayes, nous avons :

$$P(M_g|Y) = \frac{P(Y|M_g)P(M_g)}{P(Y)}.$$

La loi *a priori* des modèles M_g est supposée non-informative : $P(M_1) = P(M_2) = ... = P(M_m)$. Nous réalisons donc qu'il suffit de calculer $P(Y|M_g)$ pour effectuer notre choix. Ainsi par la formule des probabilités totales, il vient

$$\begin{aligned} P(Y|M_g) &= \int_{\psi_g} P(Y,\psi_g|M_g)d\psi_g \\ &= \int_{\psi_g} P(Y|\psi_g,M_g)P(\psi_g|M_g)d\psi_g \quad \text{(Bayes)} \\ &= \int_{\psi_g} f_{M_g}(Y,\psi_g)P(\psi_g|M_g)d\psi_g, \end{aligned}$$

où $f_{M_g}(Y,\theta_g)$ est la vraisemblance du modèle M_g de paramètre ψ_g. Cette intégrale peut s'exprimer sous la forme

$$P(Y|M_g) = \int_{\psi_g} e^{f(\psi_g)}d\psi_g \quad , \quad \text{où } f(\psi_g) = \ln(f_{M_g}(Y,\psi_g)P(\psi_g|M_g)).$$

Cette formule nous fait naturellement penser à celle de la transformée de Laplace, d'où l'utilisation de la méthode d'approximation de Laplace pour calculer cette probabilité.

Proposition 6. *(Approximation de Laplace). Soit une fonction $L : \mathbb{R}^d \to \mathbb{R}$ telle que L est C^2 sur $\mathbb{R}^d$ et atteint un unique maximum sur $\mathbb{R}^d$ en $u^\star$. Alors*

$$\int_{\mathbb{R}} e^{nL(u)}du = e^{nL(u^\star)}\left(\frac{2\pi}{n}\right)^{\frac{d}{2}} \|-L''(u^\star)\|^{-\frac{1}{2}} + O(n^{-1})$$

Cette proposition reste valable sous certaines conditions (satisfaites chez nous) pour des fonctions L dépendantes de n. La fonction L vaut sur l'ensemble des observations Y_j indépendantes

$$L(u) = L_n(\psi_g) = \frac{f(\psi_g)}{n} = \frac{1}{n}\sum_{j=1}^{n} \ln(f_{M_g}(Y_j, \psi_g)) + \frac{\ln(P(\psi_g|M_g))}{n}$$

Notons $\psi_g^\star = \underset{\psi_g \in \psi_g}{\arg\max}\, L_n(\psi_g)$, et $H_{\psi_g^\star}$ l'opposé de la matrice hessienne des dérivées partielles d'ordre 2 de $L_n(\psi_g)$ en ψ_g,

$$H_{\psi_g^\star} = -\left[\frac{\partial^2 L_n(\psi_g)}{\partial\psi_g^i \partial\psi_g^l}\right]_{i,l} |_{\psi_g=\psi_g^\star},$$

Alors nous avons

$$P(Y|M_g) = e^{f(\psi_g^\star)} \left(\frac{2\pi}{n}\right)^{\frac{K_g}{2}} \|H_{\psi_g^\star}\|^{-\frac{1}{2}} + O\left(\frac{1}{n}\right), \quad \text{d'où}$$

$$\ln(P(Y|M_g)) = \ln(f_{M_g}(Y, \psi_g^\star)) + \ln(P(\psi_g^\star|M_g)) + \frac{K_g}{2}(\ln 2\pi - \ln n) - \frac{1}{2}\ln(\|H_{\psi_g^\star}\|) + O\left(\frac{1}{n}\right).$$

Reste donc à calculer $\psi_g^\star$ et $H_{\psi_g^\star}$. Quand $n \to \infty$, $\ln(f_{M_g}(Y, \psi_g)P(\psi_g|M_g))$ croît alors que $\ln(P(\psi_g|M_g))$ reste constant ; donc ce dernier terme a tendance à disparaître.
Asymptotiquement, $\psi_g^\star$ peut être remplacé par l'estimateur du maximum de vraisemblance $\hat{\psi}_g$ défini par $\hat{\psi}_g = \underset{\psi_g \in \Psi_g}{\arg\max}\, \frac{1}{n} f_{M_g}(Y, \psi_g)$. Nous procédons de même pour le calcul de $H_{\psi_g^\star}$, ce qui nous renvoie au calcul de la matrice d'information de Fisher que nous noterons $J_{\hat{\psi}_g}$. Ces approximations introduisent un terme d'erreur en $n^{-1/2}$, ce qui donne au final quand $n \to \infty$:

$$\ln(P(Y|M_g)) = \overbrace{\ln(f_{M_g}(Y, \hat{\psi}_g)) - \frac{K_g}{2}\ln n}^{\text{tend vers } -\infty \text{ avec } n} + \underbrace{\ln(P(\hat{\psi}_g|M_g)) + \frac{K_g}{2}\ln 2\pi - \frac{1}{2}\ln(\|J_{\hat{\psi}_g}\|)}_{\text{borné}} + O\left(\frac{1}{\sqrt{n}}\right)$$

C'est à partir de cette équation que nous retrouvons la forme du critère BIC, par des considérations asymptotiques et en négligeant le terme borné et le terme d'erreur :

$$\ln(P(Y|M_g)) \simeq \ln(f_{M_g}(Y, \hat{\psi}_g)) - \frac{K_g}{2}\ln n.$$

Le terme de pénalité en $\ln n$ est ainsi issu de l'approximation de Laplace.
Afin d'uniformiser avec les critères déjà existants, le critère BIC est donné par

$$BIC_g = -2\ln(f_{M_g}(X, \hat{\psi}_g)) + K_g \ln n,$$

et le modèle sélectionné satisfait :

$$M_{BIC} = \underset{M_g \in \{M_1, \ldots, M_m\}}{\arg\min} BIC_g.$$

Nous avons considéré que la loi *a priori* des modèles était uniforme. Dans le cas contraire, un terme supplémentaire apparait mais cette configuration est relativement rare. Par contre le premier des deux termes que nous négligeons (terme borné) s'apparente à une erreur systématique de l'approximation, qui pourrait se révéler préoccupante car non-négligeable dans certains cas.

4.1.4 Notion de consistance pour la dimension

Le critère BIC Nous pouvons montrer de manière rigoureuse que le critère BIC associé à l'estimateur du maximum de vraisemblance est consistant, dans le sens où il sélectionne parmi un ensemble de modèles le modèle qui tend à être le modèle théorique. Notons au passage que la pénalité du critère BIC satisfait les conditions de Nishii (1988) évoquées en section 4.1.3.

Il est également intéressant de donner une interprétation intuitive de cette "consistance pour la dimension", grâce à la notion de quasi-vrai modèle. Supposons que les modèles M_1, M_2, ..., M_m sont emboîtés ; et notons $\hat{d}_{KL}(f^0, M_g)$ la plus petite distance KL de f^0 au modèle M_g :

$$\hat{d}_{KL}(f^0, M_g) = \inf_{\psi_g \in \Psi_g} d_{KL}(f^0(.), f_{M_g}(., \psi_g)).$$

$\hat{d}_{KL}$ est logiquement décroissante en fonction de la dimension K_g associée au modèle M_g. Soit M_t le modèle à partir duquel cette distance ne diminue plus (il y a toujours existence de ce modèle). Selon le critère de distance KL, M_t est préférable à tous les sous-modèles M_g ($g \in [\![1, t-1]\!]$) puisqu'il est plus fidèle à f^0. De la même façon, M_t est préférable à tous les sur-modèles M_g ($g \in [\![t+1, m]\!]$) car ils sont plus complexes sans pour autant apporter davantage de précision (risque "d'overfitting"). On dit que le critère BIC est consistant pour ce modèle M_t particulier, appelé quasi-vrai modèle. Nous nous intéressons à l'étude de la différence

$$BIC_g - BIC_t, \quad g \neq t \text{ et } n \to \infty.$$

Cas où $g < t$:

$$\begin{aligned}
BIC_g - BIC_t &= -2\ln(f_{M_g}(Y, \hat{\psi}_g)) + K_g \ln n - \left[-2\ln(f_{M_t}(Y, \hat{\psi}_t)) + K_t \ln n\right] \\
&= -2\ln(f_{M_g}(Y, \hat{\psi}_g)) + 2\ln(f_{M_t}(Y, \hat{\psi}_t)) + (K_g - K_t)\ln n \\
&= 2n\left[-\frac{1}{n}\sum_{j=1}^{n}\ln(f_{M_g}(y_j, \hat{\psi}_g)) + \frac{1}{n}\sum_{j=1}^{n}\ln(f_{M_t}(y_j, \hat{\psi}_t))\right] + (K_g - K_t)\ln n \\
&= 2n\left[\frac{1}{n}\sum_{j=1}^{n}\ln\left(\frac{f^0(y_j)}{f_{M_g}(y_j, \hat{\psi}_g)}\right) - \frac{1}{n}\sum_{j=1}^{n}\ln\left(\frac{f^0(y_j)}{f_{M_t}(y_j, \hat{\psi}_t)}\right)\right] + (K_g - K_t)\ln n.
\end{aligned}$$

D'après Ripley (1995), les deux sommes sont des estimateurs consistants de $\hat{d}_{KL}(f^0, M_g)$ et $\hat{d}_{KL}(f^0, M_t)$. Nous obtenons

$$BIC_g - BIC_t \simeq 2n\left[\hat{d}_{KL}(f^0, M_g) - \hat{d}_{KL}(f^0, M_t)\right] + (K_g - K_t)\ln n.$$

Asymptotiquement, le premier terme en n domine par rapport au deuxième terme en $\ln n$, et tend vers $+\infty$ lorsque $n \to \infty$. Cela signifie que les modèles M_g sont asymptotiquement disqualifiés car le BIC doit être minimisé, or $BIC_g >> BIC_t$. Nous ne tendons donc pas à sous-estimer la dimension réelle du modèle.

Cas où $g > t$: le terme $2\ln(f_{M_g}(Y, \hat{\psi}_g)) - 2\ln(f_{M_t}(Y, \hat{\psi}_t))$ correspond à la statistique du rapport de vraisemblance pour des modèles emboîtés, qui sous l'hypothèse H_0 (selon laquelle $\psi = \hat{\psi}_t$) suit asymptotiquement une loi du χ^2 à $(K_g - K_t)$ degrés de liberté. D'où

$$BIC_g - BIC_t \simeq -\chi^2_{(K_g - K_t)} + (K_g - K_t)\ln n.$$

Ici le second terme domine et tend vers $+\infty$ lorsque $n \to \infty$, donc les modèles M_g sont encore une fois disqualifiés.

Pour résumer, c'est donc le terme en $\ln n$ obtenu par l'approximation de Laplace qui permet au critère BIC de converger ! Que se passe-t-il si la famille des modèles considérée est mal spécifiée (i.e. que le modèle théorique n'appartient pas à cette famille) ? L'hypothèse liée à cette question n'apparait nulle part dans la construction du critère BIC, pourtant certains auteurs l'ont posé sans justifier son utilité. Ce que nous savons, c'est que le BIC converge en probabilité vers le quasi-vrai modèle lorsqu'il est unique. Cependant, le quasi-vrai modèle peut être très éloigné en distance KL du modèle théorique.

Le critère AIC De nombreux papiers comparent les performances obtenues par AIC et BIC en termes de sélection de modèle, dans le but de désigner un "meilleur" critère. En réalité, ces deux critères ne sont pas comparables car ils poursuivent deux objectifs bien différents. BIC cherche à maximiser la probabilité a posteriori que le modèle sélectionné soit le modèle théorique, alors que AIC essaie d'atteindre le meilleur compromis biais-variance.
Dans la pratique, BIC sélectionne rapidement des modèles de dimension plus petite que AIC (dès que $n > 7$ car $\ln(7) \simeq 2$ dans le terme de pénalité). Il est alors logique de se poser la question de la consistance pour la dimension du critère AIC. Comme nous l'avons vu dans la construction du critère, le modèle retenu est :

$$M_{AIC} = \underset{M_g \in \{M_1,\dots,M_m\}}{\arg\min} (-2 \ln f_{M_g}(Y, \hat{\psi}_g) + 2K_g).$$

Avec le même raisonnement asymptotique que celui utilisé dans le cas du BIC, nous pouvons montrer que AIC n'est pas consistant pour la dimension. En effet,

$$\begin{array}{lrcl} g < t : & AIC_g - AIC_t & \simeq & 2n[\hat{d}_{KL}(f^0, M_g) - \hat{d}_{KL}(f^0, M_t)] + 2(K_g - K_t) \\ g > t : & AIC_g - AIC_t & \simeq & -\chi^2_{K_g - K_t} + 2(K_g - K_t) \end{array}$$

Dans le premier cas, les modèles M_g sont asymptotiquement disqualifiés pour les mêmes raisons que précédemment. Par contre, la probabilité de disqualifier les modèles "surdimensionnés" ne tend pas vers 0 dans le deuxième cas, puisque le terme de pénalité ne diverge pas. AIC n'est donc pas consistant pour la dimension. Cependant, AIC a d'autres propriétés intéressantes. Rappelons que ce critère a pour objectif de minimiser l'espérance de la distance KL :

$$\begin{aligned} M_{AIC} &= \underset{M_g \in \{M_1,\dots,M_m\}}{\arg\min} \mathbb{E}\left[\int_{\mathcal{Y}} \ln\left(\frac{f^0(y)}{f_{M_g}(y, \hat{\psi}_g)}\right) f^0(y) dy\right] \\ &= \underset{M_g \in \{M_1,\dots,M_m\}}{\arg\min} \left(\hat{d}_{KL}(f^0, M_g) + \mathbb{E}\left[\int_{\mathcal{Y}} \ln\left(\frac{f_{M_g}(y, \bar{\psi}_g)}{f_{M_g}(y, \hat{\psi}_g)}\right) f^0(y) dy\right]\right), \end{aligned}$$

avec $\bar{\psi}_g$ est la valeur de ψ_g qui minimise la distance KL entre f^0 et $f_{M_g}(., \psi_g)$.
Dans cette dernière expression, le premier terme désigne le biais (distance du modèle M_g à f^0) alors que le deuxième terme mesure la variance (difficulté d'estimer $f_{M_g}(y, \bar{\psi}_g)$). Le modèle sélectionné par AIC réalise donc le meilleur compromis biais-variance parmi l'ensemble des modèles, et *est dit à ce titre efficace*. Contrairement à ce qui est souvent dit, AIC dépend de la taille d'échantillon n car il somme sur les échantillons la distance KL entre f^0 et $f_{M_g}(., \bar{\psi}_g)$.

Le bilan de l'étude de ces deux critères est qu'il n'existe pas de critère universellement meilleur. La connaissance des données et le but de l'expérimentateur (modèle explicatif ou prédictif) doivent conditionner le choix du critère de sélection. En pratique et dans le contexte des mélanges, il est bien connu que ces deux critères ont tendance à surestimer le nombre théorique de composantes (voir le cas de modèle mal spécifié dans Baudry (2009)). Cela semble également être le cas dans nos applications, où nous avons noté de fortes ressemblances entre composantes d'un même mélange.

4.2 Sélection de modèle mélange

Comme nous l'avons vu précédemment, l'utilisation de la modélisation mélange a explosé depuis la parution de l'article de Dempster et al. (1977). Par conséquent savoir pourquoi choisir tel ou tel modèle mélange a suscité l'intérêt de beaucoup de chercheurs, sans pour autant qu'il n'émerge une solution universelle pour répondre à la question du choix du nombre de composantes. De nombreux articles sont consacrés au développement de méthodes de calibration d'un mélange, mais celles-ci souffrent régulièrement d'un manque de justification théorique. Bien souvent, les propriétés de convergence des critères de sélection et des algorithmes proposés ne sont pas garanties : c'est le cas par exemple de l'algorithme SSMEM proposé par Hai Xan et al. (2004). Pour pallier le fait que l'algorithme EM ne puisse estimer directement le nombre G de composantes (G doit être spécifié a priori), Hai Xan et al. (2004) introduisent des nouveaux critères de distance qui vont servir à décider itérativement d'un regroupement ou d'une division des composantes une fois les paramètres de ces composantes estimés via l'EM. Dans le même esprit, Wenbin (2006) définit une distance entre la densité du mélange obtenue par l'EM et la densité des observations via la méthode de Parzen (estimation par noyau gaussien). Cette distance sert de socle au choix du nombre de composantes a posteriori (après estimation par EM), en minimisant la pénalité qui y est liée. L'algorithme semble bien se comporter mais les données de test sont originellement bien séparées, ce qui ne nous permet pas d'être convaincu de sa pertinence.
Une revue sur la question de l'évaluation du nombre de composantes d'un mélange est proposée dans Oliviera-Brochado and Vitorino Martins (2005). Les auteurs rappellent que le nombre de composantes du mélange n'est évidemment pas observable dans la majorité des cas, et que cinq grandes approches ont vu le jour : les tests d'hypothèses présentés en section 3.1.4 (et donc du bootstrap sur le ratio de vraisemblance), les critères d'information (log-vraisemblance pénalisée), les critères de classification (liés à la statistique d'entropie), le ratio d'information minimum (information ratio matrix) et enfin les outils graphiques. Garel (2007) souligne la difficulté d'établir la multimodalité avec le test du ratio de vraisemblance généralisé. En effet, le résultat classique selon lequel la distribution de la statistique de ce test suit une loi du χ^2 n'est en général pas applicable dans le cas des mélanges. Son papier donne un aperçu des récents développements liés à l'utilisation de cette technique pour détecter l'hétérogénéité des données. Une méthode consiste à utiliser le bootstrap afin de pallier à cette difficulté : en guise d'exemple, Schlattmann (2003) étudie par des mélanges de lois de Poisson l'homogénéité des SMR (Standard Mortality Ratio) dus à la leucémie infantile en Allemagne dans les années 1980.

Etant donné nos objectifs, nous nous focalisons sur les critères de classification. Toutefois Oliviera-Brochado and Vitorino Martins (2005) effectuent des comparaisons intéressantes entre les méthodes proposées dans la littérature suivant le type d'étude menée : la conclusion

est qu'il n'existe pour le moment pas de meilleure méthode de sélection de mélange quelle que soit la nature de celui-ci. Leur étude sur la détermination du nombre de composantes dans le cas de mélange de régressions (Oliviera-Brochado and Vitorino Martins (2008)) par diverses techniques illustre très bien ce propos, de même que le papier de Sarstedt et al. (2011).
Dans le cadre de modèles mélanges où l'on souhaite explicitement décrire la structure de la population et où l'objectif est de trouver le nombre de composantes, la plupart des auteurs (McLachlan and Peel (2000), Fraley and Raftery (1998)) s'accordent à dire que le critère BIC donne de meilleurs résultats que le critère AIC puisqu'il recherche le quasi-vrai modèle. La convergence du BIC pour estimer l'ordre d'un mélange gaussien a notamment été démontrée dans Keribin (1999). Néanmoins Celeux and Soromenho (1996) précisent qu'il existe de meilleurs critères de sélection que les critères classiques majoritairement utilisés dans la littérature pour réaliser une classification. Le critère ICL, que nous avons choisi d'étudier dans la suite de cette thèse, en est un. La très grande majorité des travaux se sont concentrés sur les propriétés de ce critère dans le cadre gaussien (Baudry (2009)), mais n'ont pas abordé le contexte des mélanges de GLMs. Une panoplie d'études sur simulation et données réelles dans Baudry (2009) suggèrent un meilleur comportement d'ICL par rapport au BIC dans une optique de clustering. Nous présentons dans la section suivante quelques notions phares grâce à l'usage des mélanges de lois normales. La question de la convergence du critère de sélection ICL pour des mélanges gaussiens sera également abordée.

4.2.1 Introduction avec les mélanges gaussiens

Avant de présenter l'étude sur les mélanges de GLMs, nous proposons d'introduire la problématique avec des mélanges gaussiens. Ce choix nous parait pertinent dans la mesure où il est plus facile de comprendre les nouveaux concepts sur des objets que nous avons l'habitude de manipuler, les mélanges gaussiens étant somme toute des mélanges relativement classiques. La densité d'une loi normale multivariée $\mathcal{N}(\mu, \Sigma)$ par rapport à la mesure de Lebesgue λ sur $\mathbb{R}^d$ est donnée par

$$\forall y \in \mathbb{R}^d, \forall \mu \in \mathbb{R}^d, \forall \Sigma \in \mathbb{S}_+^d, \quad f_{\mathcal{N}}(y; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)},$$

où $\mathbb{S}_+^d$ est l'ensemble des matrices symétriques définies positives sur $\mathbb{R}^d$. On note θ le vecteur de paramètres $\theta = (\mu, \Sigma)$. Comme nous l'avons vu au chapitre 3, nous pouvons exprimer la densité d'un mélange gaussien dans le cas général (distribution mélangeante discrète ou continue) par rapport à la mesure de Lebesgue comme

$$f(y) = \int f_{\mathcal{N}}(y; \theta) d\nu(\theta) \quad d\lambda\text{-p.s.},$$

où ν est la distribution de probabilité sur l'espace des paramètres.
Nous nous restreignons à l'étude des mélanges finis, donc avec support fini. Autrement dit nous avons $\nu = \sum_{i=1}^G \pi_i \delta_{\theta_i}$; avec δ_{θ_i} la mesure de Dirac en θ, $\{\theta_1, ..., \theta_G\}$ le support de ν, et les $\pi_i \in [0, 1]$ telles que $\sum_{i=1}^G \pi_i = 1$. L'ensemble des G-uplets $(\pi_1, ..., \pi_G)$ qui satisfont cette condition est noté par la suite Π_G.
Les composantes du mélange sont les $f_{\mathcal{N}}(.; \theta_i)$, et les poids sont les π_i. Comme dans le chapitre 3, l'ensemble des paramètres est représenté par le vecteur $\psi = (\pi_1, ..., \pi_G, \xi^T) \in \Psi$ avec

$\xi^T = (\theta_1^T, ..., \theta_G^T)$. Nous avons donc finalement pour un mélange gaussien discret

$$\forall G \in \mathbb{N}^*, \forall \psi_G \in \left(\Pi_G \times (\mathbb{R}^d)^G \times (\mathbb{S}_+^d)^G\right), \quad f(.;\psi_G) = \sum_{i=1}^{G} \pi_i f_{\mathcal{N}}(.;\theta_i). \tag{4.7}$$

Nous pouvons ainsi définir l'ensemble des mélanges gaussiens à G composantes comme l'ensemble des densités qui appartiennent à

$$M_G = \left\{ f(.;\psi_G) = \sum_{i=1}^{G} \pi_i f_{\mathcal{N}}(.;\theta_i) \mid \psi_G = (\pi_1, ..., \pi_G, \theta_1, ..., \theta_G) \in \Psi_G \right\},$$

avec $\Psi_G \subset \Pi_G \times (\mathbb{R}^d \times \mathbb{S}_+^d)^G$. Notons par la suite $\Theta_G = (\mathbb{R}^d \times \mathbb{S}_+^d)^G$.
En fait, Baudry (2009) considère un sous-ensemble de M_G pour pouvoir mener à bien l'étude d'un nouvel estimateur ainsi que ses propriétés de convergence. Il propose le nouvel ensemble

$$\tilde{M}_G = \left\{ (\pi_1 f_{\mathcal{N}}(.;\theta_1), ..., \pi_G f_{\mathcal{N}}(.;\theta_G)) \mid (\pi_1, ..., \pi_G) \in \Pi_G, (\theta_1, ..., \theta_G) \in (\mathbb{R}^d \times \mathbb{S}_+^d)^G \right\}$$

afin d'éviter les problèmes de non-identifiabilité (dus aux possibles permutations) qui empêchent un "mapping" unique entre l'espace des paramètres et l'ensemble des modèles mélanges qui en sont issus. Comme discuté en section 3.1.2, la non-identifiabilité peut s'éviter en imposant des contraintes sur les paramètres ($\pi_1 < ... < \pi_G$ par exemple). Dans la pratique, Keribin (1999) se satisfait de l'identifiabilité "faible" (autorisation du "label switching") pour obtenir ses résultats de convergence du maximum de vraisemblance : imposer $\pi_i > 0$ ($\forall i$) et $\theta_i \neq \theta_k$ ($i \neq k$) dans le contexte des mélanges gaussiens suffit à la garantir. Baudry (2009) travaille dans l'ensemble $\tilde{M}_G$ et ne suppose donc pas d'hypothèse garantissant l'identifiabilité faible ou forte des mélanges.

Comme vu dans les parties 3.1.1 et 3.1.3, les données complètes pour l'individu j sont les paires (Y_j, Z_j) dont la densité est donnée ici par

$$f(y_j, z_j; \psi_G) = \prod_{i=1}^{G} \left[\pi_i f_{\mathcal{N}}(y_j;\theta_i)\right]^{z_{ij}}.$$

En effet, nous pouvons vérifier que la loi de ce couple définit bien la loi mélange de l'équation (4.7) pour Y, sachant que Z est une loi multinomiale (les $Y_1, ..., Y_n$, de même que les $Z_1, ..., Z_n$, sont des échantillons i.i.d.). Cette remarque est la clef de voûte pour l'implémentation de l'algorithme EM qui maximise la log-vraisemblance du vecteur Y tout en considèrant le problème aux données complètes. Redner and Walker (1984) prouvent la monotonie et la convergence de cet algorithme, qui en font aujourd'hui l'outil le plus utilisé pour ce type de problème. Les auteurs avertissent toutefois les utilisateurs du comportement parfois étrange de l'algorithme EM pour calibrer des mélanges, à cause de problèmes bien connus (valeurs initiales de l'algorithme, non-convexité de la vraisemblance, convergence vers des maxima locaux, bornitude) liés à la complexité de la fonction de vraisemblance. Ces difficultés sont d'autant plus flagrantes que les composantes du mélange ne sont pas bien séparées, donc que la multimodalité des données n'est pas évidente. Malheureusement ce sera bien souvent le cas dans les applications.
D'un point de vue pratique, la convergence vers un maximum local s'explique soit par des valeurs initiales de l'algorithme mal choisies ; soit par un petit groupe d'observations très

proches les unes des autres qui vont constituer une composante dont la covariance va tendre vers 0, provoquant l'explosion de la vraisemblance et privilégiant cette solution à celle du maximum global. Il faut alors par exemple fixer une borne inférieure sur la variance des composantes. Concrètement, des résultats asymptotiques montrent que ce problème tend à disparaitre lorsque $n \to +\infty$. Théoriquement, la difficulté majeure dans le cas de mélanges gaussiens provient de la non-bornitude de la vraisemblance (ou sa dérivée) lorsque nous nous plaçons aux frontières de notre espace des paramètres. D'où l'idée de considérer l'espace des paramètres dans un **ouvert**, mais nous y reviendrons plus tard. La vraisemblance des densités appartenant à l'ensemble M_G vaut

$$\forall \psi_G \in \Psi_G, \quad L(\psi_G; y_1, ..., y_n) = L(\psi_G) = \prod_{j=1}^{n} \sum_{i=1}^{G} \pi_i f_{\mathcal{N}}(y_j; \theta_i).$$

Sachant que le maximum de vraisemblance préconise de choisir $\hat{\psi}_G \in \underset{\psi_G \in \Psi_G}{\arg\max}\, L(\psi_G)$ comme meilleur estimateur des paramètres, nous concevons aisément qu'un problème se pose si $L(\psi_G) \to +\infty$... Il faut donc restreindre l'espace des paramètres en supposant par exemple qu'il est compact (Redner (1981)), ce qui permet de garantir a priori l'existence d'un tel estimateur. La difficulté de considérer un tel espace est d'en choisir les nouvelles frontières, en prenant le risque que la distribution théorique des données en devienne exclue (si tant est qu'elle appartienne effectivement à la famille considérée !). Plusieurs propositions émergent alors, dépendant essentiellement de la paramétrisation du modèle mélange en question (Baudry (2009), p.29).

Revenons maintenant à l'algorithme EM de calibration du mélange : l'idée de cet algorithme est de transformer le problème (en passant aux données complètes) dans le but de simplifier l'étape d'optimisation. En effet, il est bien plus facile de maximiser une somme de logarithmes que de maximiser le logarithme d'une somme. Mathématiquement, nous avions $\forall \psi_G \in \Psi_G$,

$$L(\psi_G; y) = \prod_{j=1}^{n} \sum_{i=1}^{G} \pi_i f_{\mathcal{N}}(y_j; \theta_i) \quad \text{qui devient} \quad L_c(\psi_G; y, z) = \prod_{j=1}^{n} \prod_{i=1}^{G} \left(\pi_i f_{\mathcal{N}}(y_j; \theta_i)\right)^{z_{ij}}.$$

Ainsi la log-vraisemblance à maximiser est $\ln L_c(\psi_G) = \sum_{j=1}^{n} \sum_{i=1}^{G} z_{ij} \ln\left(\pi_i f_{\mathcal{N}}(y_j; \theta_i)\right)$.
Nous appelons cette quantité la log-vraisemblance complète : maximiser la log-vraisemblance complète est équivalent (à un terme près) à maximiser la log-vraisemblance des données observées y. Concrètement nous maximisons séparément et un à un le logarithme de chaque densité gaussienne par rapport aux observations qui y sont assignées, ce qui est numériquement très simple. Comme Z n'est pas observée, nous maximisons plus exactement $\mathbb{E}_Z[\ln L_c(\psi; Y, Z)|Y]$; d'où l'étape E de l'algorithme EM (voir section 3.1.3 pour la description détaillée de l'algorithme).

Notion de classe : cluster ou composante ?

Nous voulons effectuer une classification non-supervisée à partir de notre modèle mélange. A partir de cette observation, il est important de préciser ce que nous entendons par "classe" : une "classe" est-elle une composante ? Un regroupement de composantes ? Qu'est-ce qu'un cluster ? Généralement, un cluster est un regroupement visuel : il s'agit d'individus proches les uns des autres d'un point de vue géométrique si nous les projetons dans un plan. Un cluster

peut donc être un regroupement de composantes du mélange, typiquement si ces composantes se ressemblent. Dans notre étude nous assimilerons les classes (clusters) aux composantes du mélange, dans la mesure où notre objectif est justement d'obtenir un mélange final qui permette de bien distinguer les groupes (composantes) entre eux. Les contraintes imposées sur l'espace des paramètres jouent un rôle prépondérant dans la panoplie des formes que peut prendre une composante. Pensons par exemple dans le cas gaussien à une contrainte sur la matrice de covariance de type $\Sigma_i = \sigma_i^2 I$, alors les composantes ne peuvent avoir qu'une forme d'ellipsoide parallèle aux axes définissant l'espace des observations. Il sera donc important de visualiser la contrainte imposée en fonction de la forme des clusters que nous souhaitons considérer.

4.2.2 Le maximum de vraisemblance classifiante conditionnelle

L'estimation par maximum de vraisemblance classifiante conditionnelle fait intervenir une nouvelle quantité : la vraisemblance classifiante conditionnelle. Cette fonction est issue de la vision donnée par l'algorithme EM, et se rapproche de la vraisemblance des données complètes dont nous avons parlé dans la section précédente. Dans cette partie nous exhibons dans un premier temps le lien entre vraisemblance des données observées et vraisemblance des données complètes, afin d'en avoir une interprétation et une représentation plus précises. Puis nous développons un exemple qui non seulement explicite la différence majeure avec l'estimation par maximum de vraisemblance, mais expose également les nouveaux problèmes auxquels nous sommes confrontés avec l'utilisation de cette fonction. L'accent est ensuite mis sur l'étude des propriétés de convergence de l'estimateur découlant de cette quantité, grâce aux théorèmes résultant de la théorie asymptotique classique que nous adaptons à notre contexte d'étude.

De la vraisemblance à la vraisemblance classifiante conditionnelle (L_{cc})

Plusieurs auteurs ont tenté d'exploiter le lien entre la vraisemblance des données observées et la vraisemblance des données complètes. De ces études ont émergé un bon nombre d'algorithmes divers et variés, dont le plus connu est le CEM (Classification EM). Cet algorithme, proposé par Celeux and Govaert (1992), consiste à ajouter une étape d'affectation des observations par la règle MAP entre les étapes Espérance et Maximisation de l'algorithme EM. L'optimisation est facilitée puisque les probabilités conditionnelles a posteriori d'appartenir à telle ou telle composante disparaissent dans l'expression à maximiser. Quelques années auparavant, Hathaway (1986) avait déjà remarqué qu'un terme spécifique apparaissait dans l'écriture de la vraisemblance aux données complètes, ou **vraisemblance classifiante** : $\forall \psi_G \in \Psi_G$,

$$
\begin{aligned}
\ln L_c(\psi_G; y, z) &= \sum_{j=1}^{n}\sum_{i=1}^{G} z_{ij} \ln\left(\pi_i f_{\mathcal{N}}(y_j;\theta_i)\right) \\
&= \sum_{j=1}^{n}\sum_{i=1}^{G} z_{ij} \ln\Big(\underbrace{\frac{\pi_i f_{\mathcal{N}}(y_j;\theta_i)}{\sum_{k=1}^{G}\pi_k f_{\mathcal{N}}(y_j;\theta_k)}}_{\tau_i(y_j;\psi_G)}\Big) + \underbrace{\sum_{j=1}^{n}\overbrace{\sum_{i=1}^{G} z_{ij}}^{=1} \ln\left(\sum_{k=1}^{G}\pi_k f_{\mathcal{N}}(y_j;\theta_k)\right)}_{\ln L(\psi_G;y)} \\
&= \ln L(\psi_G; y) + \sum_{j=1}^{n}\sum_{i=1}^{G} z_{ij} \ln \tau_i(y_j;\psi_G) \qquad (4.8)
\end{aligned}
$$

Evidemment, cette relation reste vraie en toute généralité : nous sommes ici dans le cas de mélanges gaussiens car nous présenterons les idées sous cet angle pour en faciliter la compréhension, mais nous travaillerons ensuite avec des mélanges de GLMs.

Le terme d'entropie

Le terme qui lie les deux vraisemblances est très proche de ce que l'on appelle couramment l'entropie. Initialement, la fonction d'entropie est définie comme suit :

$$\forall \psi_G \in \Psi_G, \forall y_j \in \mathbb{R}^d, Ent(\psi_G; y_j) = -\sum_{i=1}^{G} \tau_i(y_j; \psi_G) \ln \tau_i(y_j; \psi_G).$$

Ainsi nous réalisons que cette fonction résulte de l'espérance de la variable aléatoire Z (non observée) prise dans le deuxième membre de (4.8). D'où le nom et la définition de la **vraisemblance classifiante conditionnelle** :

$$\begin{aligned} \ln L_{cc}(\psi_G; Y) &= \mathbb{E}_Z \left[\ln L_c(\psi_G; Y, Z)\right] \\ &= \ln L(\psi_G; Y) + \sum_{j=1}^{n}\sum_{i=1}^{G} \mathbb{E}_Z[Z_{ij}|Y_j] \ln \tau_i(Y_j; \psi_G) \\ &= \ln L(\psi_G; Y) - Ent(\psi_G; Y), \end{aligned} \tag{4.9}$$

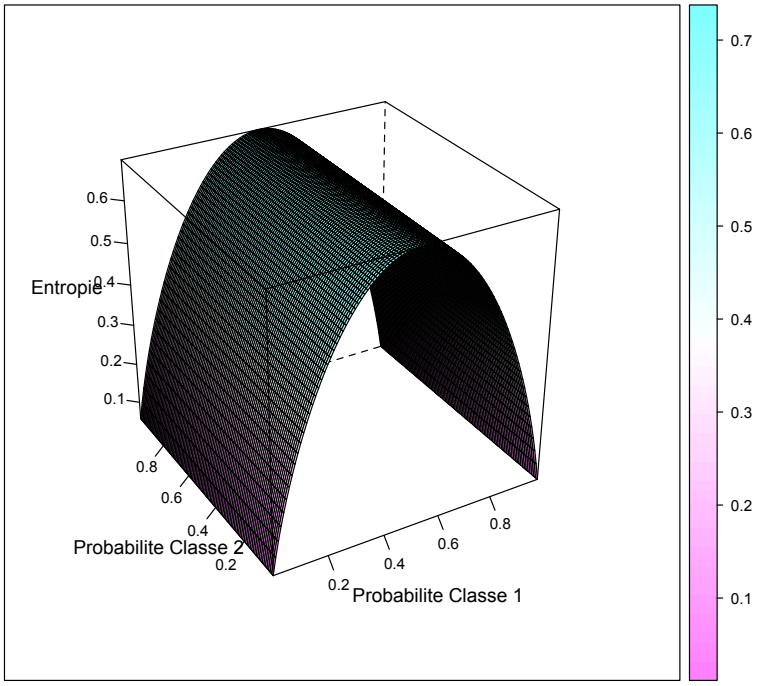
avec $Ent(\psi_G; Y) = \sum_{j=1}^{n} Ent(\psi_G; Y_j)$.

Voyons maintenant le comportement de la fonction d'entropie avec des antécédents τ_i qui ne sont rien d'autre que des probabilités. L'objectif est d'étudier

$$\begin{aligned} Ent : \quad [0,1]^G &\to \mathbb{R} \\ \tau = (\tau_1, ..., \tau_G) &\to Ent(\tau) = -\textstyle\sum_{i=1}^{G} \tau_i \ln \tau_i \end{aligned}$$

Si nous traçons le graphe de cette fonction en supposant qu'il y a deux composantes dans le mélange (toujours sous la contrainte sur Π_G), nous obtenons la figure 4.1. L'interprétation en est relativement simple : l'entropie est maximale en cas d'équiprobabilité, et minimale lorsque la probabilité d'être dans l'une ou l'autre des classes vaut 1. Ce terme peut être vu comme une pénalisation de la vraisemblance observée, comme le montre l'équation (4.9). Ainsi, plus la probabilité de classer une observation dans l'un ou l'autre des groupes est équirépartie, plus la pénalité est grande. Nous pénalisons donc fortement un manque de confiance lors de la classification via la règle MAP (après calibration du mélange). Au contraire, si les probabilités a posteriori de classer les observations dans telle ou telle composante tendent vers 0 ou 1, alors la pénalisation est minime. La vraisemblance classifiante conditionnelle est donc quasiment équivalente à la vraisemblance des données observées lorsque l'information disponible dans l'échantillon permet de construire un mélange dans lequel les observations sont clairement affectées aux composantes. Notons d'ailleurs que dans le cas d'homogénéité (1 composante), ces deux vraisemblances sont identiques.
Remarquez l'analogie avec la méthode CART développée au chapitre 1 : il est très intéressant de constater que la fonction d'entropie correspond précisément à la mesure d'hétérogénéité de l'échantillon définie par la fonction d'impureté dans CART (l'index de Gini) ! Nous ne sommes donc pas étonnés d'une telle interprétation, puisque ces idées sont totalement concordantes.

FIGURE 4.1 – Fonction d'entropie avec $G = 2$ (2 groupes possibles).



D'autre part, l'entropie a une limite nulle lorsque l'une des probabilités τ_i tend vers 0. En revanche, elle n'est pas dérivable en 0 car en considérant la fonction $f(\tau_i) = \tau_i \ln \tau_i$:

$$f^{'}(\tau_i) = \ln(\tau_i) + 1 \quad \Rightarrow \quad \lim_{\tau_i \to 0^+} f^{'}(\tau_i) = -\infty \quad \Rightarrow \quad \lim_{\tau_i \to 0^+} (Ent\,\tau)^{'} = +\infty.$$

Ceci constituera un point clef dans la définition de l'espace des paramètres acceptable pour assurer la convergence de l'estimateur basé sur la vraisemblance classifiante conditionnelle. Il faudra donc éviter que les proportions du mélange *a posteriori* tendent vers la valeur nulle. En effet une des hypothèses requiert que la dérivée de la vraisemblance classifiante conditionnelle reste bornée, mais nous reviendrons sur ce point en section 4.2.2.

L'estimateur $ML_{cc}E$ dans le cas de mélange gaussien

Rappelons que l'ensemble des mélanges gaussiens est défini ici par

$$M_G = \left\{ \sum_{i=1}^{G} \pi_i f_{\mathcal{N}}(.; \theta_i) \mid (\pi_1, ..., \pi_G, \theta_1, ..., \theta_G) \in \Psi_G \right\},$$

avec $\Psi_G \subset \Pi_G \times \Theta_G$ où $\Theta_G \subset (\mathbb{R}^d \times \mathbb{S}_+^d)^G$. D'habitude, les contraintes sur l'espace des paramètres sont essentiellement formulées sur l'espace Θ_G, puisque seuls les paramètres sur Θ_G interviennent dans la maximisation de la vraisemblance complète. K_G est la dimension "optimale" du modèle M_G : à titre d'exemple, il n'est pas nécessaire de calculer les G proportions du mélange puisque la dernière peut être déduite des autres grâce à la contrainte sur Π_G.

Exemple 1. *Considérons un mélange gaussien d-multivarié M_G à G composantes, avec la paramétrisation donnée par la représentation spectrale des mélanges : nous pouvons écrire la matrice de covariance de chaque composante comme $\Sigma_i = \lambda_i D_i A_i D_i^{'}$ où D_i est l'orientation de la composante, et $\lambda_i A_i$ est sa forme (Biernacki (2009) et McLachlan and Peel (2000) p.110). Nous imposons la contrainte que $\Psi_G \subset \Pi_G \times \Theta_G$ soit un ensemble compact, par exemple : $\forall i \in [\![1, ..., G]\!]$,*

$$\begin{gathered}\pi_i \geq \pi_{\min}, \\ \lambda_{\min} \leq \lambda_i \leq \lambda_{\max}, \\ \forall k \in [\![1, d]\!],\ \mu_{\min} \leq \mu_i^k \leq \mu_{\max}, \\ \forall k \in [\![1, d]\!],\ a_{\min} \leq A_i^k \leq a_{\max},\end{gathered}$$

où A_i est la matrice de diagonale $(A_i^1, ..., A_i^d)$. Ce modèle a pour dimension

$$K_G = \underbrace{(G-1)}_{poids} + \underbrace{Gd}_{moyennes} + \underbrace{G\frac{d(d+1)}{2}}_{covariances}.$$

Nous avons vu qu'il est indispensable d'avoir un mélange identifiable (section 3.1.2), ce qui nous oblige en fait à considérer l'espace

$$\tilde{M}_G = \left\{ (\pi_1 f_{\mathcal{N}}(.;\theta_1), ..., \pi_G f_{\mathcal{N}}(.;\theta_G)) \mid (\pi_1, ..., \pi_G) \in \Pi_G, (\theta_1, ..., \theta_G) \in \Theta_G \subset (\mathbb{R}^d \times \mathbb{S}_+^d)^G \right\}.$$

C'est d'autant plus vrai que nous allons dorénavant travailler avec la vraisemblance classifiante conditionnelle, qui nécessite de connaître les distributions de chaque composante pour correctement définir l'entropie. Les contraintes imposées à l'espace des paramètres doivent donc notamment garantir cette identifiabilité.

Par analogie avec la méthode du maximum de vraisemblance, nous pouvons définir un nouvel estimateur à partir de la vraisemblance L_{cc}. Afin de garder une certaine logique, cet estimateur est appelé "estimateur du maximum de vraisemblance classifiante conditionnelle" et est noté $ML_{cc}E$. De la même manière que l'estimateur du maximum de vraisemblance mais en adaptant le raisonnement, l'estimateur du maximum de vraisemblance classifiante $ML_{cc}E$ pour un modèle M_G satisfait

$$\psi_G^{ML_{cc}E} = \underset{\psi_G \in \Psi_G}{\arg\max}\, \mathbb{E}_{f^0}[\ln L_{cc}(\psi_G, Y)],$$

estimé naturellement de manière empirique par la loi des grands nombres :

$$\hat{\psi}_G^{ML_{cc}E} = \underset{\psi_G \in \Psi_G}{\arg\max}\, \frac{1}{n} \sum_{j=1}^{n} \ln L_{cc}(\psi_G; y_j).$$

En développant l'expression de la log-vraisemblance classifiante conditionnelle pour des mélanges gaussiens, les contraintes que nous devons imposer sur l'espace des paramètres (pour que celle-ci ne diverge pas) deviennent quasiment évidentes. En effet, $\ln L_{cc}(\psi_G; y_j)$ vaut pour une observation y_j

$$\underbrace{\ln\left(\sum_{i=1}^{G} \pi_i\, f_{\mathcal{N}}(y_j;\theta_i)\right)}_{\ln L(\psi_G;y_j)} + \underbrace{\sum_{i=1}^{G} \frac{\pi_i f_{\mathcal{N}}(y_j;\theta_i)}{\sum_{k=1}^{G} \pi_k f_{\mathcal{N}}(y_j;\theta_k)} \ln\left(\frac{\pi_i f_{\mathcal{N}}(y_j;\theta_i)}{\sum_{k=1}^{G} \pi_k f_{\mathcal{N}}(y_j;\theta_k)}\right)}_{-Ent(\psi_G;y_j)}. \tag{4.10}$$

L'annexe D détaille l'étude des limites de ces deux termes dans les différentes configurations possibles, sachant que

$$f_{\mathcal{N}}(y_j;\theta_i) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det\Sigma_i}}\, e^{-\frac{1}{2}(y_j-\mu_i)^T\Sigma_i^{-1}(y_j-\mu_i)}.$$

En prenant le cas unidimensionnel pour simplifier ($\det\Sigma_i$ devient σ_i^2), nous avons :

$$\left.\begin{array}{lcl} \sigma_i^2 & \to & 0 \\ \sigma_i^2 & \to & +\infty \\ \mu_i & \to & -\infty \\ \mu_i & \to & +\infty \\ \pi_i & \to & 0 \end{array}\right\} \Rightarrow \ln L_{cc}(\psi_G;y_j) \text{ et/ou } \frac{\partial \ln L_{cc}(\psi_G;y_j)}{\partial\theta_i} \text{ diverge(nt).}$$

Ces limites, obtenues astucieusement ou par développements limités dans le but de lever les formes indéterminées rencontrées, suggèrent que les situations critiques correspondent majoritairement à des paramètres qui ne seraient pas bornés.

Objectif du $ML_{cc}E$

L'exemple suivant permet de se rendre compte de la différence fondamentale entre l'estimateur $ML_{cc}E$ et l'estimateur MLE, ce dernier minimisant la distance KL entre la distribution à estimer $f(.;\psi_G)$ et la distribution théorique $f^0(.)$.

Exemple 2. *(Baudry). La densité théorique f^0 est celle d'une loi normale centrée réduite unidimensionnelle $\mathcal{N}(0,1)$ $(d=1)$. Considérons le modèle*

$$M = \left\{ \frac{1}{2} f_{\mathcal{N}}(.;-\mu,\sigma^2) + \frac{1}{2} f_{\mathcal{N}}(.;\mu,\sigma^2);\ \mu\in\mathbb{R},\ \sigma^2\in\mathbb{R}^{+*} \right\}.$$

Aucune contrainte supplémentaire n'est imposée.
Même dans un modèle extrêmement simple où σ^2 serait fixée, il est impossible de trouver l'expression du $ML_{cc}E$! Par contre, nous pouvons le calculer numériquement, et nous obtenons

$$(\mu^{ML_{cc}E}, \sigma^{2,ML_{cc}E}) = (0.83, 0.31).$$

Ce résultat signifie que l'estimateur $ML_{cc}E$ construit un mélange à deux composantes tel que $\mathbb{E}_{f^0}[\ln L_{cc}(\mu,\sigma^2)]$ est maximisée en un point unique (à un "label switch" près). Cependant, cet estimateur ne correspond en rien à l'estimateur MLE car celui-ci aurait donné l'estimation $(\mu^{MLE}, \sigma^{2,MLE}) = (0,1)$. En effet cette estimation minimise la distance KL à la densité théorique, et cette densité obtenue par MLE ne serait rien d'autre que la densité théorique elle-même !

Cet exemple illustre parfaitement le but de l'estimateur $ML_{cc}E$, qui n'est pas de retrouver la distribution théorique des données même lorsqu'elle est contenue dans le modèle considéré (ce qui est le cas ici). L'estimateur MLE n'ayant pas de règle pour désigner deux classes (composantes) adéquates pour ce modèle, il construirait les deux mêmes composantes $\frac{1}{2}f_{\mathcal{N}}(.;0,1)$ exactement superposées, et l'affectation des observations à l'une ou l'autre de ces composantes serait complètement arbitraire (avec probabilité de 1/2, d'où une entropie maximale). En revanche le compromis recherché par le $ML_{cc}E$, qui pénalise cette trop grande entropie, amène

à trouver une autre estimation des paramètres : en découle une plus grande confiance dans l'assignation des observations à l'une ou l'autre des composantes. Nous distinguons également le bémol de l'utilisation de cet estimateur : si les données ne sont pas issues d'un mélange, il est clairement moins bon que l'estimateur MLE par définition (car ce dernier minimise la distance KL a la quasi-vraie distribution).

Convergence de l'estimateur $MLccE$

Avant toute chose rappelons que nous considérons un modèle paramétrique M_G de dimension K_G, de paramètre $\psi_G \in \Psi_G$ tel que $\Psi_G \subset \mathbb{R}^{K_G}$. Oublions également le contexte des mélanges et introduisons les notations suivantes :

- $\forall \psi_G \in \mathbb{R}^{K_G},\ \|\psi_G\|_\infty = \max\limits_{1\leq k\leq K_G} |\psi_G^k|$, où ψ_G^k est la k^e coordonnée de ψ_G dans la base canonique de $\mathbb{R}^{K_G}$;
- $\forall \psi_G \in \Psi_G,\ \forall \tilde{\Psi}_G \subset \Psi_G$, notons d la distance $d(\psi_G, \tilde{\Psi}_G) = \inf\limits_{\tilde{\psi}_G \in \tilde{\Psi}_G} \|\psi_G - \tilde{\psi}_G\|_\infty$.

Nous avons vu que l'estimateur $MLccE$ peut être approché par le M-estimateur suivant :

$$\hat{\psi}_G^{MLccE} = \underset{\psi_G \in \Psi_G}{\arg\max} \underbrace{\frac{1}{n}\sum_{j=1}^{n} \ln L_{cc}(\psi_G; y_j)}_{L_n(\psi_G; Y)}.$$

Le théorème de convergence de l'estimateur $MLccE$ se place dans un cadre très général et provient d'une adaptation des résultats de van der Vaart (1998) et Baudry (2009). van der Vaart (1998) donne les hypothèses de convergence faible d'un M-estimateur pourvu que celui-ci existe ! Notre version de ce théorème permet d'expliciter les conditions suffisantes pour la **convergence forte** d'un M-estimateur vers le meilleur paramètre dans un problème d'optimisation de log-vraisemblance classifiante conditionnelle.

Théorème 4. *Soit* $\Psi_G \subset \mathbb{R}^{K_G}$ *et* $\ln L_{cc} : \Psi_G \times \mathbb{R}^d \to \mathbb{R}$.

Si nous avons les trois hypothèses suivantes :

- *(H1-A) :* $\exists \psi_G^b \in \Psi_G$ *tel que* $\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G^b; Y)\right] = \max\limits_{\psi_G \in \Psi_G} \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G; Y)\right]$ *;*
- *(H2-A) :* $\forall \epsilon > 0, \quad \sup\limits_{\{\psi_G;\, d(\psi_G, \Psi_G^b) > \epsilon\}} \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G; Y)\right] < \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G^b; Y)\right]$,

  où $\Psi_G^b = \left\{ \psi_G^b : \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G^b; Y)\right] = \max\limits_{\psi_G \in \Psi_G} \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G; Y)\right] \right\}$.
- *(H3-A) :* $\forall \psi_G \in \Psi_G, \quad \sup\limits_{\psi_G \in \Psi_G} \left| L_n(\psi_G; Y) - \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G; Y)\right] \right| \underset{n\to\infty}{\longrightarrow} 0$ *p.s. ;*

Alors,

en définissant $\hat{\psi}_G = \hat{\psi}_G(Y_1, ..., Y_n) \in \Psi_G$ *tel que* $\exists n_0 \in \mathbb{N},\ \forall n \geq n_0$,

$$L_n(\hat{\psi}_G; Y) \geq L_n(\psi_G^b; Y) - \xi_n$$

$$\text{avec } \begin{cases} \xi_n \geq 0 \quad p.s. \\ \xi_n \underset{n\to\infty}{\longrightarrow} 0 \ p.s. \end{cases} \text{, nous avons} \qquad d(\hat{\psi}_G, \Psi_G^b) \underset{n\to\infty}{\longrightarrow} 0 \quad p.s.$$

Autrement dit, le M-estimateur $ML_{cc}E$ tend presque sûrement vers le meilleur estimateur possible (inconnu). La preuve de ce théorème se décompose comme suit :

Démonstration. Soit $\epsilon > 0$ fixé. Posons

$$\eta = \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G^b; Y)\right] \quad - \sup_{d(\psi_G, \Psi_G^b) > \epsilon} \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G; Y)\right].$$

Nous savons d'après (H1-A) et (H2-A) que η existe et que $\eta > 0$.
D'après l'hypothèse (H3-A), nous obtenons :

$$\left[\exists n_0 \in \mathbb{N},\, \forall n \geq n_0,\, \sup_{\psi_G \in \Psi_G} |L_n(\psi_G) - \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G)\right]| < \frac{\eta}{3}\right] \text{p.s.},$$

et par définition de l'estimateur $\hat{\psi}_G$,

$$\left[\exists n_0 \in \mathbb{N},\, \forall n \geq n_0, L_n(\hat{\psi}_G) \geq L_n(\psi_G^b) - \frac{\eta}{3}\right] \text{p.s.}$$

Enfin en décomposant l'erreur entre le M-estimateur et le meilleur paramètre, on obtient

$$\begin{aligned}\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G^b)\right] - \mathbb{E}_{f^0}\left[\ln L_{cc}(\hat{\psi}_G)\right] \leq \quad & \overbrace{\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G^b)\right] - L_n(\psi_G^b)}^{<\eta/3} + \overbrace{L_n(\psi_G^b) - L_n(\hat{\psi}_G)}^{\leq \eta/3} \\ + \quad & \underbrace{L_n(\hat{\psi}_G) - \mathbb{E}_{f^0}\left[\ln L_{cc}(\hat{\psi}_G)\right]}_{<\eta/3}.\end{aligned}$$

Comme l'intersection de deux événements de probabilité 1 est elle-même de probabilité 1, nous avons

$$\left[\exists n_0 \in \mathbb{N},\, \forall n \geq n_0, \quad \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G^b)\right] - \mathbb{E}_{f^0}\left[\ln L_{cc}(\hat{\psi}_G)\right] < \eta\right] \text{p.s.}$$

D'où $\left[\exists n_0 \in \mathbb{N},\, \forall n \geq n_0,\, d(\hat{\psi}_G, \Psi_G^b) \leq \epsilon\right]$ p.s.

Ce résultat étant vrai pour un ϵ donné, il suffit de prendre une suite dénombrable $(\epsilon_p)_{p\in\mathbb{N}}$ donnée par $\epsilon_p = \frac{1}{2^p}$, et de considérer l'intersection de tous ces événements de probabilité 1. Nous obtenons ainsi le résultat quel que soit ϵ. De là, nous pouvons conclure

$$\left[\forall \epsilon,\, \exists n_0 \in \mathbb{N},\, \forall n \geq n_0,\, d(\hat{\psi}_G, \Psi_G^b) < \epsilon\right] \text{p.s.}$$

□

D'après l'hypothèse (H1-A), les meilleurs estimateurs en termes de maximum de vraisemblance classifiante conditionnelle ne doivent pas constituer un ensemble vide ($\Psi_G^b \neq \{\emptyset\}$). **Cette hypothèse est garantie si l'espérance de la log-vraisemblance L_{cc} est continue d'une part, et si l'espace des paramètres est compact d'autre part**. En effet, une fonction continue sur un ensemble compact atteint toujours son supremum. Dans le cas d'un mélange gaussien par exemple, la vraisemblance L_{cc} est une fonction continue en ψ_G puisque les densités gaussiennes et l'entropie sont continues. Reste donc à prouver la continuité de son espérance (qui est une intégrale paramétrique en ψ_G) ; or d'après le théorème de convergence

dominée, la continuité est assurée tant que la distribution f^0 n'a pas de queue trop épaisse (voir section 4.3.3).

L'hypothèse (H2-A) signifie que l'ensemble Ψ_G^b est discret, et que les paramètres que nous considérons se situent toujours à une distance positive de cet ensemble. **Elle est aussi garantie sous l'hypothèse de compacité de** Ψ_G. En effet, $\psi_G \mapsto \mathbb{E}_{f^0}[\ln L_{cc}(\psi_G; Y)]$ atteint son maximum sur l'espace $\Psi_G \backslash \{\psi_G \in \Psi_G : d(\psi_G, \Psi_G^b) > \epsilon\}$. Cet espace est fermé et borné si Ψ_G est compact, donc le supremum est nécessairement inférieur à $\mathbb{E}_{f^0}[\ln L_{cc}(\psi_G^b)]$.

L'hypothèse (H3-A) est une hypothèse très forte, inspirée du théorème de Glivenko-Cantelli pour les fonctions de répartition. Cette généralisation stipule la convergence uniforme de la moyenne empirique de la log-vraisemblance classifiante conditionnelle vers son espérance mathématique, et **nécessite l'étude approfondie de la classe des fonctions considérées** (L_{cc}). Dans la suite, nous détaillons davantage cette hypothèse qui fait intervenir de nouvelles notions de mesure de complexité.

Résultats auxiliaires fondamentaux Comme nous l'avons vu, la convergence presque sûre du $ML_{cc}E$ vers le meilleur paramètre ne coule pas de source : en effet, elle nécessite de démontrer un résultat de convergence uniforme. Afin de prouver cette convergence uniforme, nous introduisons les notations suivantes :

- Soit $r \in \mathbb{N}^* \cup \{\infty\}$, et $g : \mathbb{R}^d \to \mathbb{R}$. $\|g\|_r$ est la norme L_r de g par rapport à f^0,
  avec $\begin{cases} \text{si } r < \infty : \|g\|_r = \mathbb{E}_{f^0}\left[|g(Y)|^r\right]^{\frac{1}{r}} ; \\ \text{sinon } : \|g\|_\infty = ess\sup_{Y \sim f^0} |g(Y)| \text{ où } ess\sup_{Z \sim \mathbb{P}} Z = \inf\{z : \mathbb{P}(Z \leq z) = 1\}. \end{cases}$
- Soit une application linéaire $t : \mathbb{R}^{K_G} \to \mathbb{R}$.
  $\|t\|_\infty$ est la norme usuelle sur un espace vectoriel normé : $\|t\|_\infty = \max\limits_{\psi_G \in \mathbb{R}^{K_G}} \dfrac{t(\psi_G)}{\|\psi_G\|_\infty}$.
- $\forall \tilde{\Psi}_G$ borné tel que $\tilde{\Psi}_G \subset \mathbb{R}^{K_G}$, $diam\, \tilde{\Psi}_G = \sup\limits_{\psi_1, \psi_2 \in \tilde{\Psi}_G} \|\psi_1 - \psi_2\|_\infty$.

Autant les hypothèses (H1-A) et (H2-A) apparaissent comme presque triviales si l'espace des paramètres est compact, autant (H3-A) aurait besoin d'être explicitée en des termes plus communs. Un certain nombre de résultats vont nous être utiles pour garantir cette hypothèse. Le but de cette section est de présenter un ensemble de définitions et de lemmes qui permettront de substituer (H3-A) par de nouvelles hypothèses, vérifiables de manière plus directe. A ce titre, nous donnons la définition d'une classe de fonctions $\mathbb{P}$-*Glivenko Cantelli* :

Définition. *Une classe $\mathcal{G}$ de fonctions mesurables $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ est $\mathbb{P}$-Glivenko Cantelli si et seulement si*

$$\left\| \left| \frac{1}{n} \sum_{j=1}^{n} g(Y_i) - \mathbb{E}\left[g(Y)\right] \right| \right\|_{\mathcal{G}} := \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{j=1}^{n} g(Y_i) - \mathbb{E}\left[g(Y)\right] \right| \longrightarrow 0 \quad p.s.,$$

où $Y_1, ..., Y_n$ est un échantillon de distribution $\mathbb{P}$ et l'espérance sur Y est prise par rapport à la distribution $\mathbb{P}$.

A la vue de cette définition, l'objectif est évident : prouver que la vraisemblance L_{cc} est une classe de fonctions $\mathbb{P}$-Glivenko Cantelli nous permettrait de retrouver immédiatement (H3-A). Pour démontrer cette propriété, nous introduisons la notion de "bracketing entropy" (Dudley (1999)) :

Définition. *Soient $r \in \mathbb{N}^*$, et $l, u \in L_r(\mathbb{P})$.*
Le crochet $[l,u]$ est l'ensemble des fonctions $g \in \mathcal{G}$ / $\forall y \in \mathbb{R}^d$, $l(y) \leq g(y) \leq u(y)$.
$[l,u]$ est un ϵ-crochet si $\|l-u\|_r = \mathbb{E}\left[|l-u|^r\right]^{\frac{1}{r}} \leq \epsilon$.
On note $N_{[]}(\epsilon, \mathcal{G}, L_r(\mathbb{P}))$ le nombre minimal d'ϵ-crochets pour couvrir $\mathcal{G}$.
L'entropie associée ou "bracketing entropy", notée $\mathcal{E}_{[]}(\epsilon, \mathcal{G}, L_r(\mathbb{P}))$, correspond au logarithme de $N_{[]}(\epsilon, \mathcal{G}, L_r(\mathbb{P}))$.

En fait, la "bracketing entropy" est une mesure L_r de la complexité de la classe de fonctions $\mathcal{G}$. Cette définition permet de se placer dans un cadre où contrôler les bornes du crochet revient à contrôler les fonctions qui y appartiennent. Il existe un résultat de van der Vaart (1998) (dont la preuve est au chapitre 19) qui stipule un lien entre la "bracketing entropy" d'un ensemble lié à une classe de fonction et la propriété d'être $\mathbb{P}$-*Glivenko Cantelli* pour cette même classe de fonctions :

Théorème 5. *Toute classe $\mathcal{G}$ de fonctions mesurables telles que $\mathcal{E}_{[]}(\epsilon, \mathcal{G}, L_1(\mathbb{P})) < \infty$ pour tout $\epsilon > 0$ est $\mathbb{P}$-Glivenko Cantelli.*

Pour obtenir l'hypothèse de convergence uniforme presque sûre, nous allons donc considérer la norme $L_1(\mathbb{P})$. Il suffit maintenant de trouver sous quelles conditions la classe des fonctions définies par la vraisemblance classifiante conditionnelle des mélanges de GLMs a une "bracketing entropy" finie, auquel cas nous détiendrons l'hypothèse (H3-A). Suivant les propriétés de l'espace des paramètres, il existe deux lemmes qui amènent à un tel résultat. Nous les présentons et en discutons ci-après. Dans toute la suite, la notation $(\partial \ln L_{cc} \,/\, \partial\psi)$ désigne le vecteur des dérivées partielles de la log-vraisemblance classifiante conditionnelle par rapport à chacune des composantes du vecteur ψ.

Lemme 2. *(Bracketing entropy, **cas convexe**).*
Soit $r \in \mathbb{N}^*$. Soient $K_G \in \mathbb{N}^*$ et $\Psi_G \subset \mathbb{R}^{K_G}$ un ensemble convexe.
Soit $\Psi_G^{\mathcal{O}}$ un ouvert de $\mathbb{R}^{K_G}$ tel que $\Psi_G \subset \Psi_G^{\mathcal{O}}$ et $\ln L_{cc} : \Psi_G^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
La fonction $\psi \in \Psi_G^{\mathcal{O}} \mapsto \ln L_{cc}(\psi; y)$ est supposée C^1 (f^0-presque partout) sur $\Psi_G^{\mathcal{O}}$.
Supposons que

$$\begin{aligned} L^{'}(y) &= \sup_{\psi \in \Psi_G} \left\| \left(\frac{\partial \ln L_{cc}}{\partial \psi} \right)_{(\psi;y)} \right\|_\infty < \infty \quad f^0 d\lambda - p.s., \\ \|L^{'}\|_r &= \mathbb{E}_{f^0}\left[L^{'}(Y)^r \right]^{\frac{1}{r}} < \infty; \end{aligned}$$

Alors (avec $\tilde{\Psi}_G$ borné)

$$\forall \tilde{\Psi}_G \subset \Psi_G, \forall \epsilon > 0,\ N_{[]}(\epsilon, \{\ln L_{cc}(\psi) : \psi \in \tilde{\Psi}_G\}, \|.\|_r) \leq \max\left(\left(\frac{\|L^{'}\|_r \ diam\, \tilde{\Psi}_G}{\epsilon} \right)^{K_G}, 1 \right).$$

Nous retiendrons que dans le cas d'un espace de paramètres convexe, la "bracketing entropy" reste bornée tant que certaines conditions de régularité sont satisfaites pour la dérivée de la log-vraisemblance classifiante conditionnelle. Ce résultat s'inspire de la propriété de fonction lipschitzienne énoncée dans van der Vaart (1998) (p. 271) et Baudry (2009) dans le cadre général du contraste.

Démonstration. Soit $\epsilon > 0$, $\tilde{\Psi}_G \subset \Psi_G$ convexe et borné. Soit $\tilde{\Psi}_{G,\epsilon}$ une grille de pas ϵ dans Ψ_G qui couvre $\tilde{\Psi}_G$ dans toutes ses dimensions. Par exemple, $\tilde{\Psi}^1_{G,\epsilon}, ..., \tilde{\Psi}^{K_g}_{G,\epsilon}$ avec

$$\forall k \in [\![1, K_g]\!], \quad \tilde{\Psi}^k_{G,\epsilon} = \left\{ \tilde{\psi}^k_{\min}, \tilde{\psi}^k_{\min} + \epsilon, ..., \tilde{\psi}^k_{\max} \right\},$$

où

$$\forall k \in [\![1, K_g]\!], \quad \left\{ \psi^k : \psi \in \tilde{\Psi}_G \right\} \subset \left[\tilde{\psi}^k_{\min} - \frac{\epsilon}{2}, \tilde{\psi}^k_{\max} + \frac{\epsilon}{2} \right].$$

Comme Ψ_G est convexe, cette grille existe toujours. Pour simplifier, nous supposons sans perte de généralité que $\tilde{\Psi}_{G,\epsilon} \subset \tilde{\Psi}_G$. D'après la définition de la norme $\|.\|_\infty$, nous sommes en mesure de borner les écarts entre les points de la grille et n'importe quel point de l'espace (et ce dans toutes les dimensions) :

$$\forall \tilde{\psi}_G \in \tilde{\Psi}_G, \ \exists \tilde{\psi}_{G,\epsilon} \in \tilde{\Psi}_{G,\epsilon} \text{ tel que } \|\tilde{\psi}_G - \tilde{\psi}_{G,\epsilon}\| \leq \frac{\epsilon}{2}.$$

Nous en déduisons immédiatement que le cardinal de $\tilde{\Psi}_{G,\epsilon}$ est au plus de

$$\max \left(\prod_{k=1}^{K_G} \frac{\sup_{\psi \in \tilde{\Psi}_G} \psi^k_G - \inf_{\psi \in \tilde{\Psi}_G} \psi^k_G}{\epsilon}, 1 \right) \leq \max \left(\left(\frac{diam \tilde{\Psi}_G}{\epsilon} \right)^{K_G}, 1 \right).$$

Soient ψ_1 et ψ_2 dans Ψ_G, et $y \in \mathbb{R}^d$. D'après le théorème des accroissements finis,

$$\begin{aligned}
|\ln L_{cc}(\psi_1; y) - \ln L_{cc}(\psi_2; y)| &\leq \sup_{\psi \in [\psi_1, \psi_2]} \left\| \left(\frac{\partial \ln L_{cc}}{\partial \psi} \right)_{(\psi; y)} \right\|_\infty \|\psi_1 - \psi_2\|_\infty \quad f^0 d\lambda - \text{p.p.} \\
&\leq \sup_{\psi \in \Psi_G} \left\| \left(\frac{\partial \ln L_{cc}}{\partial \psi} \right)_{(\psi; y)} \right\|_\infty \|\psi_1 - \psi_2\|_\infty \quad f^0 d\lambda - \text{p.p.} \\
&\leq L^{'}(y) \, \|\psi_1 - \psi_2\|_\infty \quad f^0 d\lambda - \text{p.p.}
\end{aligned}$$

Soit $\tilde{\psi}_G \in \tilde{\Psi}_G$, prenons $\tilde{\psi}_{G,\epsilon} \in \tilde{\Psi}_{G,\epsilon}$ tel que $\|\tilde{\psi}_G - \tilde{\psi}_{G,\epsilon}\|_\infty \leq \frac{\epsilon}{2}$. Alors,

$$\forall y \in \mathbb{R}^d, \quad |\ln L_{cc}(\tilde{\psi}_{G,\epsilon}; y) - \ln L_{cc}(\tilde{\psi}_G; y)| \leq L^{'}(y) \, \frac{\epsilon}{2},$$

et donc

$$\ln L_{cc}(\tilde{\psi}_{G,\epsilon}; y) - \frac{\epsilon}{2} L^{'}(y) \leq \ln L_{cc}(\tilde{\psi}_G; y) \leq \ln L_{cc}(\tilde{\psi}_{G,\epsilon}; y) + \frac{\epsilon}{2} L^{'}(y).$$

En prenant en compte l'hypothèse selon laquelle $\|L^{'}\|_r < \infty$; et sachant que la largeur de cet intervalle est de $\epsilon L^{'}$, nous constatons que l'ensemble des $\epsilon \|L^{'}\|_r$ crochets noté

$$\left\{ \left[\ln L_{cc}(\tilde{\psi}_{G,\epsilon}; y) - \frac{\epsilon}{2} L^{'}(y); \ln L_{cc}(\tilde{\psi}_{G,\epsilon}; y) + \frac{\epsilon}{2} L^{'}(y) \right] : \tilde{\psi}_{G,\epsilon} \in \tilde{\Psi}_{G,\epsilon} \right\}$$

couvre $\left\{ \ln L_{cc}(\tilde{\psi}_G) : \tilde{\psi}_G \in \tilde{\Psi}_G \right\}$ et a un cardinal d'au plus $\max \left(\left(\frac{diam \tilde{\Psi}_G}{\epsilon} \right)^{K_G}, 1 \right)$. □

La démonstration de ce lemme repose sur l'utilisation du théorème des accroissements finis (TAF) qui nous permet d'encadrer la log-vraisemblance L_{cc}, et l'hypothèse de convexité de l'espace des paramètres qui nous permet de construire la grille de pas ϵ. En effet, en nous plaçant dans le plan pour plus de facilité, le TAF garantit qu'il existe un point sur l'intervalle ouvert considéré pour lequel la dérivée est égale à la pente entre les extrémités de ce même intervalle. En prenant le supremum de cette dérivée, l'inégalité est immédiate. Nous utilisons dans la démonstration un ensemble $\tilde{\Psi}_G$ borné avec $\tilde{\Psi}_G \in \Psi_G$: cela ne pose aucun problème car nous utiliserons ce lemme localement autour de ψ_G^b (cf démonstration du théorème 4). D'autre part, il n'y aucune raison pour que l'hypothèse de convexité soit vraie en toute généralité pour l'espace des paramètres. Nous avons d'ailleurs souvent parlé précédemment d'espace de paramètres compact, car la compacité est une propriété très intéressante lors de l'étude de fonctions continues (comme le sont les densités des lois continues que nous manipulons). Il apparait donc utile d'étendre ce résultat en proposant le lemme suivant.

Lemme 3. *(Bracketing entropy,* ***cas compact****).*
Soit $r \in \mathbb{N}^*$. Soient $K_G \in \mathbb{N}^*$ et $\Psi_G \subset \mathbb{R}^{K_G}$ un ensemble compact.
Soit $\Psi_G^{\mathcal{O}}$ un ouvert de $\mathbb{R}^{K_G}$ tel que $\Psi_G \subset \Psi_G^{\mathcal{O}}$ et $\ln L_{cc} : \Psi_G^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
La fonction $\psi \in \Psi_G^{\mathcal{O}} \mapsto \ln L_{cc}(\psi; y)$ est supposée C^1 (f^0-presque partout) sur $\Psi_G^{\mathcal{O}}$.
Supposons que

$$\begin{aligned} L^{'}(y) &= \sup_{\psi \in \Psi_G^{\mathcal{O}}} \left|\left| \left(\frac{\partial \ln L_{cc}}{\partial \psi} \right)_{(\psi;y)} \right|\right|_{\infty} < \infty \quad f^0 d\lambda - p.s., \\ \|L^{'}\|_r &= \mathbb{E}_{f^0}\left[L^{'}(Y)^r \right]^{\frac{1}{r}} < \infty; \end{aligned}$$

Alors $\exists Q \in \mathbb{N}^*$,

$$\forall \tilde{\Psi}_G \subset \Psi_G, \forall \epsilon > 0,\ N_{[]}(\epsilon, \{\ln L_{cc}(\psi) :\ \psi \in \tilde{\Psi}_G\}, \|.\|_r) \leq \max\left(Q\left(\frac{\|L^{'}\|_r\ diam\, \tilde{\Psi}_G}{\epsilon} \right)^{K_G}, 1 \right).$$

Dans cette version, l'espace des paramètres est supposé être compact. Le prix d'une telle hypothèse est l'apparition d'une constante Q qui mesure en fait la non-convexité de cet espace. Q dépend typiquement de la géométrie de Θ ($Q = 1$ si Θ est convexe). Nous remarquons que l'ensemble compact considéré est inclu dans un ouvert sur lequel la propriété sur le supremum de la dérivée de la vraisemblance L_{cc} est toujours valable.

Démonstration. Soient $B_1, ..., B_Q$ un nombre fini de boules ouvertes qui permettent de couvrir l'ensemble Ψ_G, telles que $\bigcup_{q=1}^{Q} B_q \subset \Psi_G^{\mathcal{O}}$. En supposant que Ψ_G est compact, nous avons toujours ce résultat (propriété de Borel-Lebesgue). Ainsi (et en notant *conv* l'enveloppe convexe),

$$\Psi_G = \bigcup_{q=1}^{Q} (B_q \cap \Psi_G) \subset \bigcup_{q=1}^{Q} conv(B_q \cap \Psi_G).$$

$\forall q$, $conv(B_q \cap \Psi_G)$ est convexe, et $\sup_{\psi_G \in conv(B_q \cap \Psi_G)} \left|\left| \left(\frac{\partial \ln L_{cc}}{\partial \psi} \right)_{(\psi;y)} \right|\right|_{\infty} \leq L^{'}(y)$ car $conv(B_q \cap \Psi_G) \subset B_q \subset \Psi_G^{\mathcal{O}}$. Nous appliquons alors le lemme 2 à $(B_q \cap \tilde{\Psi}_G) \subset conv(B_q \cap \Psi_G)$ et nous

obtenons :

$$N_{[]}(\epsilon, \{\ln L_{cc} : \psi_G \in \tilde{\Psi}_G \cap B_q\}, \|.\|_r) \leq \max\left(\left(\frac{\|L^{'}\|_r \, diam\tilde{\Psi}_G}{\epsilon}\right)^{K_G}, 1\right).$$

Puis, nous en concluons le résultat final sachant que

$$N_{[]}(\epsilon, \{\ln L_{cc} : \psi_G \in \tilde{\Psi}_G\}, \|.\|_r) \leq N_{[]}(\epsilon, \cup_{q=1}^{Q}\{\ln L_{cc} : \psi_G \in \tilde{\Psi}_G \cap B_q\}, \|.\|_r).$$

□

Nous utilisons dans cette démonstration le fait que Ψ_G soit toujours localement convexe : en effet, il est couvert par un ensemble de boules ouvertes, qui sont elle-mêmes convexes. Il ne reste ensuite plus qu'à appliquer le lemme 2 à l'enveloppe convexe de l'intersection de Ψ_G avec chacune de ces boules ouvertes. Par cette technique, nous obtenons le résultat sur l'espace compact entier.
Si nous résumons, l'hypothèse (H3-A) du théorème 4 est satisfaite dès lors que certaines propriétés de régularité sont respectées par la classe de fonction considérée (ici la vraisemblance L_{cc}). Nous énonçons un dernier lemme qui ne nous servira pas directement dans cette section, mais dont nous aurons besoin afin de prouver les propriétés de consistance du critère ICL.

Lemme 4. *Soit $r \geq 2$. Soient $K_G \in \mathbb{N}^*$ et $\Psi_G \subset \mathbb{R}^{K_G}$ un ensemble **convexe**.*
Soit $\Psi_G^{\mathcal{O}}$ un ouvert de $\mathbb{R}^{K_G}$ tel que $\Psi_G \subset \Psi_G^{\mathcal{O}}$ et $\ln L_{cc} : \Psi_G^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
$\psi \in \Psi_G^{\mathcal{O}} \mapsto \ln L_{cc}(\psi; y)$ est supposée C^1 (f^0-presque partout) sur $\Psi_G^{\mathcal{O}}$.
Supposons que

$$L(y) = \sup_{\psi \in \Psi_G} |\ln L_{cc}(\psi_G; y)| < \infty \quad f^0 d\lambda - p.s.,$$
$$\|L\|_\infty = ess\sup_{Y \sim f^0} L(Y) < \infty.$$

Et

$$L^{'}(y) = \sup_{\psi \in \Psi_G} \left|\left|\left(\frac{\partial \ln L_{cc}}{\partial \psi}\right)_{(\psi;y)}\right|\right|_\infty < \infty \quad f^0 d\lambda - p.s.,$$
$$\|L^{'}\|_2 = \mathbb{E}_{f^0}\left[L^{'}(Y)^2\right]^{\frac{1}{2}} < \infty.$$

Alors $\forall \tilde{\Psi}_G \subset \Psi_G, \forall \epsilon > 0$,

$$N_{[]}(\epsilon, \{\ln L_{cc}(\psi) : \psi \in \tilde{\Psi}_G\}, \|.\|_r) \leq \max\left(\left(\frac{2^{r-2}\|L\|_\infty^{\frac{r-2}{r}}\|L^{'}\|_2 \; diam\, \tilde{\Psi}_G}{\epsilon^{\frac{r}{2}}}\right)^{K_G}, 1\right).$$

Démonstration. Même raisonnement que la démonstration du lemme 2, mais avec un remaniement de la formule du TAF qui fait apparaître la norme infinie de L. Nous utilisons cette fois l'inégalité suivante :

$$|\ln L_{cc}(\psi_1; y) - \ln L_{cc}(\psi_2; y)|^r \leq \sup_{\psi \in [\psi_1, \psi_2]} \left|\left|\left(\frac{\partial \ln L_{cc}}{\partial \psi}\right)\right|\right|_\infty^2 \|\psi_1 - \psi_2\|_\infty^2 \left(2 \sup_{\{\psi_1, \psi_2\}} |\ln L_{cc}(\psi; y)|\right)^{r-2}$$

□

La nouvelle hypothèse sur L peut poser problème : dans la plupart des cas, elle n'est d'ailleurs pas vérifiée. En général, une condition suffisante qui permet de la garantir est que le support de la densité f^0 soit borné. Pour les lois classiques auxquelles nous pensons, ce n'est évidemment pas le cas en théorie. Néanmoins, il ne semble pas que cette contrainte soit rédhibitoire dans la pratique : en effet les observations des phénomènes modélisés sont bel et bien bornées dans les applications, comme le soulignent Bickel and Doksum (2001) (chapitre 1). Les développements et précisions apportés nous amènent donc à reformuler différemment le théorème 4 dans le cas où **l'espace des paramètres est compact**.

Théorème 6. *(Convergence forte de l'estimateur $ML_{cc}E$, espace de paramètre Ψ_G compact). Soit l'espace des paramètres Ψ_G de dimension K_G, tel que $\Psi_G \subset \mathbb{R}^{K_G}$. Soit $\ln L_{cc} : \Psi_G \times \mathbb{R}^d \to \mathbb{R}$. Soit $\Psi_G^{\mathcal{O}}$ un ouvert de $\mathbb{R}^{K_G}$ sur lequel $\ln L_{cc}$ est bien définie, et tel que $\Psi_G \subset \Psi_G^{\mathcal{O}}$.*

Si nous avons les trois hypothèses suivantes :

(H1-B) : Supposons Ψ_G compact. Alors $\Psi_G^b = \left\{ \psi_G^b : \psi_G^b = \underset{\psi_G \in \Psi_G}{\arg\max}\, \mathbb{E}_{f^0} \left[\ln L_{cc}(\psi_G; Y) \right] \right\}$.

(H2-B) : Supposons que $L^{'}(y) = \underset{\psi_G \in \Psi_G^{\mathcal{O}}}{\sup} \left|\left| \left(\frac{\partial \ln L_{cc}}{\partial \psi_G} \right)_{(\psi_G; y)} \right|\right|_{\infty} < \infty$.

(H3-B) : Supposons également que $\|L^{'}\|_1 < \infty$.

Alors,

$\exists n_0 \in \mathbb{N}, \forall n \geq n_0, \forall \psi_G^b \in \Psi_G^b$, *en définissant* $\hat{\psi}_G^{ML_{cc}E} = \hat{\psi}_G^{ML_{cc}E}(Y_1, ..., Y_n) \in \Psi_G$ *un estimateur qui maximise presque la vraisemblance classifiante conditionnelle tel que*

$$\frac{1}{n} \ln L_{cc}(\hat{\psi}_G^{ML_{cc}E}; Y) \geq \frac{1}{n} \ln L_{cc}(\psi_G^b; Y) - \xi_n$$

$$\text{avec } \begin{cases} \xi_n \geq 0 \quad p.s. \\ \xi_n \underset{n \to \infty}{\longrightarrow} 0 \ p.s. \end{cases} \text{, nous avons } \quad d(\hat{\psi}_G^{ML_{cc}E}, \Psi_G^b) \underset{n \to \infty}{\longrightarrow} 0 \ p.s.$$

Démonstration. Similaire à celle du théorème 4, à ceci près que nous utilisons les Lemmes 2 et 3 pour retomber sur les hypothèses de base puis suivre le même raisonnement. Nous considérons un ouvert $\Psi_G^{\mathcal{O}}$ pour éviter les problèmes de dérivabilité aux bornes de l'espace des paramètres. □

En effet, l'hypothèse (H1-B) permet de retrouver les hypothèses (H1-A) et (H2-A) du théorème 4. Les hypothèses (H2-B) et (H3-B) sont des conditions suffisantes pour retrouver l'hypothèse (H3-A), et la définition de l'estimateur est conforme dans les deux théorèmes. En fait les hypothèses (H2-B) et (H3-B) sont utilisées pour prouver que la "bracketing entropy" est finie, par l'intermédiaire du théorème des accroissements finis.
Sous les hypothèses de compacité de l'ensemble de paramètres Ψ_G et de régularité de la fonction $\ln L_{cc}$, nous savons donc que l'estimateur $ML_{cc}E$ **converge fortement** vers l'ensemble de paramètres Ψ_G^b qui maximise $\mathbb{E}_{f^0} \left[\ln L_{cc}(\psi_G; Y) \right]$. Dans le cas de mélange gaussien, Baudry (2009) montre que l'estimateur $ML_{cc}E$ est convergent **en probabilité** vers le paramètre maximisant $\mathbb{E} \left[\ln L_{cc}(\psi_G, Y) \right]$. D'ailleurs (H3-A) est également vérifiée s'il s'agit de

composantes gaussiennes car la "bracketing entropy" reste finie sous certaines conditions. Les contraintes sur les frontières de l'espace de paramètres compact sont fixées grâce aux limites de la fonction de vraisemblance classifiante conditionnelle et de sa dérivée, un sujet sur lequel nous reviendrons lors de l'étude des mélanges de régressions linéaires.

De manière plus générale, la convergence forte de l'estimateur $ML_{cc}E$ vers l'ensemble Ψ_G^b va nous permettre de poursuivre sur l'étude de la convergence du critère ICL_c appliqué à cet estimateur. Comme précédemment, cette étude s'effectuera d'abord dans un cadre général, puis nous verrons comment étendre ces résultats au contexte des mélanges de GLMs.

4.2.3 Le critère de sélection ICL

Biernacki (2000) essaie de contourner la difficulté du BIC à sélectionner le bon nombre de classes, particulièrement dans le cas d'un modèle mélange mal spécifié. Il veut imiter l'approche du BIC en remplaçant la vraisemblance observée par la vraisemblance classifiante. Il s'attend donc non seulement à trouver un critère qui permette de prendre en compte la qualité de la classification ; mais aussi à éliminer le problème de surestimation du nombre de composantes du mélange, un écueil souvent observé en pratique avec l'utilisation des critères AIC et BIC. Ce problème de surestimation s'interprète facilement dans le cadre de mélange gaussien : si certaines observations sont proches les unes des autres mais n'ont pas la forme d'une ellipsoïde (en se plaçant dans le plan), AIC et BIC sélectionneront un modèle qui consacrera plusieurs composantes à ces observations puisque l'objectif est de "coller" au mieux à la densité des données. Pourtant ces données ne semblent représenter qu'un groupe homogène, et **seule une** composante du mélange devrait y être affectée dans un objectif de clustering.
Le critère ICL fait partie de la classe des critères d'information pénalisés (comme AIC et BIC), ou plus exactement appartient aux critères de classification. Il a été spécialement conçu dans ce but précis, ce qui en fait un critère particulièrement adapté aux questions de clustering d'une population. A l'inverse des critères AIC et BIC qui sélectionnent le modèle qui estime au mieux la densité des observations (via la distance KL), ICL recherche plutôt le "vrai" nombre de groupes (clusters) dans une population donnée. Il vise à établir le meilleur compromis entre qualité d'estimation de cette densité et confiance dans l'affectation des observations aux différentes composantes du mélange. Il s'agit donc d'un estimateur "sur mesure", qui s'inscrit exactement dans le sens que nous voulons donner à l'utilisation des mélanges dans notre problématique opérationnelle. Une attention particulière doit être apportée à la définition de la pénalité car des nuances existent : les premiers travaux considéraient l'entropie comme partie intégrante de la pénalité mais aucun résultat théorique n'a pu être démontré sous cet angle de vue (malgré des résultats prometteurs dans les applications pratiques, Biernacki et al. (2006)). Baudry (2009) propose alors d'étudier la vraisemblance classifiante conditionnelle qui intègre l'entropie à la source, et redéfinit le critère ICL en l'associant à cette nouvelle vraisemblance. La pénalité devient du même coup identique à celle du BIC, et l'estimateur utilisé dans l'expression du critère ICL diffère de l'estimateur classique du maximum de vraisemblance.

Historique

Le choix du modèle mélange revient dans le contexte de notre étude au choix du nombre de composantes de ce mélange. En reprenant nos notations, cette remarque nous amène à considérer l'ensemble de modèles $\{M_1, ..., M_m\}$, où M_g ($g \in [\![1, m]\!]$) est un mélange à g composantes,

de dimension K_g. La densité du mélange M_g est

$$\forall g \in \mathbb{N}^*, \forall \psi_g \in \left(\Pi_g \times (\mathbb{R}^d \times \mathbb{S}_+^d)^g\right), \quad f_{M_g}(y;\psi_g) = \sum_{i=1}^{g} \pi_i f(y;\theta_i).$$

La première définition du critère ICL diffère de celle que nous allons considérer : Biernacki (2000) part du même principe que le critère BIC et sélectionne pami les modèles $\{M_1, ..., M_m\}$ le modèle tel que

$$\begin{aligned} M_{ICL} &= \underset{M_g \in \{M_1,\ldots,M_m\}}{\arg\min} \left(- \max_{\psi_g \in \Psi_g} \ln f_{M_g}(Y, Z; \psi_g) + \frac{K_g}{2} \ln n \right) \\ &= \underset{M_g \in \{M_1,\ldots,M_m\}}{\arg\min} \left(- \max_{\psi_g \in \Psi_g} \ln L_c(\psi_g; Y, Z) + \frac{K_g}{2} \ln n \right). \end{aligned}$$

Pour cela, il utilise l'approximation de Laplace sur la vraisemblance classifiante $L_c(\psi_g; y, z)$. En pratique, le vecteur Z n'est pas observé et il choisit donc de le remplacer par les affectations a posteriori $\hat{Z}^{MAP}(\hat{\psi}_g^{MLE})$. De plus, il considère que $\hat{\psi}_g^{MLE} \simeq \arg\max_{\psi_g} L_c(\psi_g; Y, Z)$ lorsque n est grand. Ces considérations, discutables, amènent ainsi au modèle sélectionné suivant

$$\begin{aligned} M_{ICL_a} &= \underset{M_g \in \{M_1,\ldots,M_m\}}{\arg\min} \left(- \ln f_{M_g}(Y, \hat{Z}^{MAP}; \hat{\psi}_g^{MLE}) + \frac{K_g}{2} \ln n \right) \\ &= \underset{M_g \in \{M_1,\ldots,M_m\}}{\arg\min} \left(- \ln L_c(\hat{\psi}_g^{MLE}; Y, \hat{Z}^{MAP}) + \frac{K_g}{2} \ln n \right) \\ &= \underset{M_g \in \{M_1,\ldots,M_m\}}{\arg\min} \left(- \ln L(\hat{\psi}_g^{MLE}; y) \underbrace{- \sum_{j=1}^{n} \sum_{i=1}^{g} \hat{Z}_{ij}^{MAP} \ln \tau_i(y_j; \hat{\psi}_g^{MLE}) + \frac{K_g}{2} \ln n}_{pen_{ICL_a}(K_g)} \right). \end{aligned}$$

De leur côté, McLachlan and Peel (2000) proposent de remplacer Z par $\tau_i(y; \hat{\psi}_g^{MLE})$. Ils obtiennent ainsi le critère

$$\begin{aligned} M_{ICL_b} &= \underset{M_g \in \{M_1,\ldots,M_m\}}{\arg\min} \left(- \ln f_{M_g}(Y, \tau(\hat{\psi}_g^{MLE}); \hat{\psi}_g^{MLE}) + \frac{K_g}{2} \ln n \right) \\ &= \underset{M_g \in \{M_1,\ldots,M_m\}}{\arg\min} \left(- \ln L_{cc}(\hat{\psi}_g^{MLE}; Y, \tau(\hat{\psi}_g^{MLE})) + \frac{K_g}{2} \ln n \right) \\ &= \underset{M_g \in \{M_1,\ldots,M_m\}}{\arg\min} \Big(- \ln L(\hat{\psi}_g^{MLE}) + \underbrace{Ent(\hat{\psi}_g^{MLE}) + \frac{K_g}{2} \ln n}_{pen_{ICL_b}(K_g)} \Big). \end{aligned}$$

En fait ICL_a et ICL_b ne sont vraiment différents en pratique que si les observations ne sont pas affectées à une composante avec une grande confiance ; nous pouvons montrer que dans

une telle configuration, $ICL_a \geq ICL_b$. Effectivement, $\forall \psi_g \in \Psi_g, \forall y \in \mathbb{R}^d$,

$$\begin{aligned} -Ent(\psi_g; y) &= \sum_{i=1}^{g} \tau_i(y; \psi_g) \ln \tau_i(y; \psi_g) \\ &\leq \sum_{i=1}^{g} \max_{k \in \{1,\dots,g\}} (\tau_k(y; \psi_g)) \ln \tau_i(y; \psi_g) \\ &= \sum_{i=1}^{g} \hat{z}_i^{MAP}(y; \psi_g) \ln \tau_i(y; \psi_g). \end{aligned}$$

Ce résultat montre que ICL_a pénalise davantage que ICL_b un modèle dont l'affectation des observations aux composantes est incertaine. Biernacki (2000) et McLachlan and Peel (2000) ont montré à travers divers exemples simulés et réels que le critère ICL est plus robuste que le critère BIC lorsque le modèle est mal spécifié (ce qui est souvent le cas dans la réalité). Certes, BIC et ICL ont un comportement similaire quand les composantes du mélange sont bien séparées ; mais ICL pénalise fortement les modèles mélanges dans le cas inverse (tout en tenant compte de la complexité du modèle) alors que BIC ne pénalise que la complexité des modèles.
Le principal problème dans cette définition d'ICL est qu'il n'existe de fait aucune relation évidente entre la théorie du maximum de vraisemblance et le terme d'entropie. En outre, le critère défini comme tel n'est pas satisfaisant au regard des théoriciens car ses propriétés n'ont pas pu être prouvées : par exemple, il n'est pas consistant au sens où BIC l'est. Même dans le cas où la distribution théorique appartient à la classe de modèles étudiés, ICL ne garantit pas de retrouver le bon nombre de composantes. Dans ces deux premières définitions, la pénalité du critère ICL comprend deux termes : l'entropie et la pénalisation du BIC en $\ln(n)$. Nous pouvons d'ailleurs remarquer que cette pénalité ne satisfait pas les conditions de Nishii (1988) car elle n'est pas négligeable devant n.
En effet, d'après la loi des grands nombres,

$$\frac{1}{n} Ent(\psi_g; y) \xrightarrow{\mathbb{P}} \mathbb{E}_{f^0} \left[Ent(\psi_g; Y) \right].$$

Nous en déduisons que $Ent(\psi_g; y) = O(n\, \mathbb{E}_{f^0} [Ent(\psi_g; Y)])$, et donc que n et $Ent(\psi_g; y)$ sont du même ordre.

Ainsi et jusqu'à très récemment, il existait clairement un gouffre entre l'intérêt pratique que suscitait ce critère et ses justifications théoriques. C'est alors que Baudry (2009) proposa une nouvelle version du critère ICL, dont la définition est liée à l'estimateur du maximum de vraisemblance classifiante $ML_{cc}E$:

$$M_{ICL_c} = \underset{M_g \in \{M_1,\dots,M_m\}}{\arg\min} \Big(-\ln L_{cc}(\hat{\psi}_g^{ML_{cc}E}) + \underbrace{\frac{K_g}{2} \ln n}_{pen_{ICL_c}} \Big).$$

En introduisant cette idée dans le cadre des mélanges gaussiens, Baudry (2009) démontre de manière rigoureuse que le nombre de composantes sélectionné via ce critère converge faiblement vers le nombre théorique de composantes ; dès lors que nous nous intéressons à des problématiques de clustering. Le paragraphe qui suit énonce de manière succinte les conditions de convergence d'un critère de sélection pénalisé dans un cadre général.

Convergence de critère de sélection

Soient g^b le nombre optimal de composantes du mélange, et g le nombre de composantes d'un modèle M_g. Un critère de sélection pénalisé convergent doit normalement satisfaire à la fois

$$\begin{cases} \forall g < g^b, & \sup_{\psi \in \Psi_{g^b}} \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi)\right] > \sup_{\psi \in \Psi_g} \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi)\right], \\ \forall g \geq g^b, & \sup_{\psi \in \Psi_{g^b}} \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi)\right] > \sup_{\psi \in \Psi_g} \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi)\right]. \end{cases}$$

Autrement dit, le biais des autres modèles par rapport au meilleur modèle M_{g^b} est stationnaire. La procédure a un but d'*identification* : elle doit permettre de retrouver le "vrai" nombre de clusters g^b. Les résultats de cette partie sont directement inspirés et adaptés de van der Vaart (1998), Massart (2007) et Baudry (2009). Nous formulons en premier lieu un théorème de convergence général de critère de sélection basé sur l'emploi d'un M-estimateur.

Théorème 7. *(Consistance faible de critère de sélection).*
Soit $\{M_g\}_{1\leq g\leq m}$ une collection de modèles de paramètres $\{\psi_g\}_{1\leq g\leq m} \in \{\Psi_g\}_{1\leq g\leq m}$ et de dimension $\{K_g\}_{1\leq g\leq m}$, avec $\Psi_g \subset \mathbb{R}^{K_g}$. Ces modèles sont classés dans un ordre croissant de complexité, avec $K_1 \leq K_2 \leq ... \leq K_m$.

Quel que soit g, posons $\Psi_g^b = \underset{\psi_g \in \Psi_g}{\arg\max}\, \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g)\right]$. Soit $\psi_g^b \in \Psi_g^b$.

Supposons que

(H1-C) $g^b = \min\left(\underset{1\leq g\leq m}{\arg\max}\, \mathbb{E}_{f^0}\left[\ln L_{cc}(\Psi_g^b)\right]\right)$;

(H2-C) $\forall g \in [\![1, m]\!]$, *soit* $\hat{\psi}_g \in \Psi_g$. *De plus,*

$\hat{\psi}_g$ *est défini tel que* $L_n(\hat{\psi}_g) \geq L_n(\psi_g^b) - \xi_n$ *où* $\begin{cases} \xi_n \geq 0 \ p.s. \\ \xi_n \underset{n\to\infty}{\longrightarrow} 0 \ p.s. \end{cases}$

$\hat{\psi}_g$ *satisfait :* $L_n(\hat{\psi}_g) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g^b)\right]$ *p.s.* ;

(H3-C) $\forall g \in [\![1, m]\!]$, $\begin{cases} pen(K_g) > 0 \ et \ pen(K_g) = o_{\mathbb{P}}(1) \ quand \ n \to +\infty; \\ n\left(pen(K_g) - pen(K_{g'})\right) \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} \infty \ quand \ g > g' ; \end{cases}$

(H4-C) $n\left(L_n(\hat{\psi}_g) - L_n(\hat{\psi}_{g^b})\right) = O_{\mathbb{P}}(1)$ *quel que soit* $g \in \underset{1\leq g\leq m}{\arg\max}\, \mathbb{E}_{f^0}\left[\ln L_{cc}(\Psi_g^b)\right]$;

Alors

en considérant $\hat{g} = \underset{1\leq g\leq m}{\arg\min}\{-L_n(\hat{\psi}_g) + pen(K_g)\}$, *on a*

$$\mathbb{P}(\hat{g} \neq g^b) \underset{n\to\infty}{\longrightarrow} 0.$$

D'un point de vue des hypothèses, (H1-C) permet d'identifier le nombre de composantes du modèle à choisir. Elle recommande de sélectionner un modèle parcimonieux, c'est à dire le modèle de plus petite dimension parmi des modèles de performance équivalente. (H3-C) définit les conditions que doit satisfaire la pénalité du critère de sélection, tandis que (H2-C)

assure la convergence presque sûre de la moyenne empirique de la log-vraisemblance L_{cc} en le M-estimateur vers son espérance en le meilleur paramètre. Nous allons voir que (H2-C) résulte de l'association du théorème 6 et du lemme 5. En ce qui concerne l'hypothèse (H4-C), elle n'est pas intuitive et nécessite de présenter d'autres résultats afin de l'expliciter. Nous présentons ci-après la preuve de ce théorème.

Démonstration. Posons l'ensemble $E = \underset{1\leq g\leq m}{\arg\max}\, \mathbb{E}_{f^0}\left[\ln L_{cc}(\Psi_g^b)\right]$. D'après (H1-C),

$$\Psi_g^b = \underset{\psi_g\in\Psi_g}{\arg\max}\, \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g)\right] \quad \text{et} \quad g^b = \min\left(\underset{1\leq g\leq m}{\arg\max}\, \mathbb{E}_{f^0}\left[\ln L_{cc}(\Psi_g^b)\right]\right).$$

- Soit $g \notin E$: définissons

$$\epsilon = \frac{1}{2}\left(\mathbb{E}_{f^0}\left[\ln L_{cc}(\Psi_{g^b}^b)\right] - \mathbb{E}_{f^0}\left[\ln L_{cc}(\Psi_g^b)\right]\right) > 0. \tag{4.11}$$

D'après (H2-C) et (H3-C), lorsque n grandit nous avons avec grande probabilité

$$|\,\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g^b)\right] - L_n(\hat{\psi}_g)\,| \leq \frac{\epsilon}{3} \Rightarrow \mathbb{E}_{f^0}\left[\ln L_{cc}(\Psi_g^b)\right] - L_n(\hat{\psi}_g) \geq -\frac{\epsilon}{3} \tag{4.12}$$

$$|\,\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_{g^b}^b)\right] - L_n(\hat{\psi}_{g^b})\,| \leq \frac{\epsilon}{3} \Rightarrow \mathbb{E}_{f^0}\left[\ln L_{cc}(\Psi_{g^b}^b)\right] - L_n(\hat{\psi}_{g^b}) \leq \frac{\epsilon}{3} \tag{4.13}$$

$$pen(K_{g^b}) \leq \frac{\epsilon}{3}. \tag{4.14}$$

Alors

$$\begin{aligned}
\overbrace{-L_n(\hat{\psi}_g) + pen(K_g)}^{IC(K_g)} &\overset{(4.12)}{\geq} -\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g^b)\right] - \frac{\epsilon}{3} \qquad (\text{car } pen(K_g) > 0)\\
&\overset{(4.11)}{\geq} -\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_{g^b}^b)\right] + \frac{5\epsilon}{3}\\
&\overset{(4.13)}{\geq} -L_n(\hat{\psi}_{g^b}) + \frac{4\epsilon}{3}\\
&\overset{(4.14)}{\geq} -L_n(\hat{\psi}_{g^b}) + \frac{4\epsilon}{3} + \overbrace{pen(K_{g^b}) - \frac{\epsilon}{3}}^{\leq 0}\\
&\geq \underbrace{-L_n(\hat{\psi}_{g^b}) + pen(K_{g^b})}_{IC(K_{g^b})} + \epsilon.
\end{aligned}$$

Or nous voulons minimiser le critère d'information IC, donc le modèle sélectionné ($\hat{g}$ composantes) ne sera pas celui qui a g composantes : $\hat{g} \neq g$.

- Soit $g \in E$ et $g \neq g^b$: par définition de g^b, on a forcément $g > g^b$. L'hypothèse (H4-C) implique qu'il existe $v > 0$ tel que pour n grand, nous avons avec grande probabilité

$$n\left(L_n(\hat{\psi}_g) - L_n(\hat{\psi}_{g^b})\right) \leq v.$$

On accroît ensuite n pour avoir $n\,(pen(K_g) - pen(K_{g^b})) > v$ avec une grande probabilité (H3-C). Nous obtenons immédiatement

$$\begin{aligned}\overbrace{-L_n(\hat{\psi}_g) + pen(K_g)}^{IC(K_g)} &\geq -L_n(\hat{\psi}_{g^b}) - \frac{v}{n} + pen(K_g)\\ &> -L_n(\hat{\psi}_{g^b}) + pen(K_{g^b})\\ &> IC(K_{g^b}).\end{aligned}$$

Encore une fois, nous avons avec grande probabilité $\hat{g} \neq g$.

Finalement, la formule des probabilités totales (les sommes concernent des ensembles dénombrables, typiquement inclus dans $\mathbb{N}^*$) donne

$$\mathbb{P}(\hat{g} \neq g^b) = \sum_{\substack{g \neq g^b \\ g \in E}} \underbrace{\mathbb{P}(\hat{g} = g)}_{\substack{\longrightarrow 0 \\ n\to\infty}} \quad + \sum_{g \notin E} \underbrace{\mathbb{P}(\hat{g} = g)}_{\substack{\longrightarrow 0 \\ n\to\infty}},$$

et nous en déduisons le résultat : $\mathbb{P}(\hat{g} \neq g^b) \underset{n\to\infty}{\longrightarrow} 0$. □

Il est à noter qu'une convergence en probabilité au lieu de presque sûre pour (H2-C) suffit à obtenir le même résultat final. L'ajout d'une condition supplémentaire pour la forme de la pénalité (de type "Nishii (1988)"), et une convergence presque sûre dans (H4-C) permettraient d'ailleurs d'étendre ce théorème à une version "presque sûre", mais nous n'avons malheureusement pas encore réussi à prouver la convergence presque sûre de (H4-C). Etant donné la définition du critère ICL_c et le fait que l'estimateur $MLccE$ soit fortement convergent vers le paramètre théorique, nous nous doutons que ce critère permettra de sélectionner un modèle dont le nombre de composantes convergera faiblement vers le nombre théorique de clusters des observations. En effet, la pénalité (comme pour le BIC) permet intuitivement de satisfaire (H3-C). De plus, cette pénalité satisfait les conditions de Nishii (1988) pour avoir un critère consistant.

Résultats auxiliaires fondamentaux Il nous reste principalement à discuter de l'hypothèse (H4-C). Les conditions pour satisfaire (H4-C) reposent sur de multiples résultats que nous allons évoquer dans cette partie. Nous commençons cependant par démontrer la proximité entre (H2-C) et certains résultats vus au préalable.

Lemme 5. *Soit* $\Psi_g \subset \mathbb{R}^{K_g}$ *et* $\ln L_{cc} : \Psi_g \times \mathbb{R}^d$. *Soit* $\Psi_g^b = \underset{\psi_g \in \Psi_g}{\arg\max}\, \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; Y)]$.

Supposons que $\begin{cases} L_n(\hat{\psi}_g; y) \geq L_n(\psi_g^b; y) - \xi_n, \ \text{où } \xi_n \geq 0 \ p.s. \text{ et } \xi_n \longrightarrow 0 \ p.s., \\ \sup\limits_{\psi_g \in \Psi_g} \left| L_n(\psi_g; y) - \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; Y)] \right| \longrightarrow 0 \ \ p.s. \end{cases}$

Alors $L_n(\hat{\psi}_g; y) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g^b; Y)] \quad p.s.$

Démonstration. Soit $\nu > 0$.
D'après la première hypothèse, nous avons

$$\left[\exists n_0 \in \mathbb{N},\ \forall n \geq n_0,\ L_n(\psi_g^b; y) - L_n(\hat{\psi}_g; y) < \frac{\nu}{2}\right] \quad p.s.$$

La deuxième hypothèse suggère que

$$\left[\exists n_0 \in \mathbb{N},\ \forall n \geq n_0,\ \mathbb{E}_{f^0}[\ln L_{cc}(\hat{\psi}_g; Y)] - L_n(\hat{\psi}_g; y) < \frac{\nu}{2}\right] \quad p.s.$$

Etudions la quantité $\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g^b; Y)] - L_n(\hat{\psi}_g; y)$ dont nous ne connaissons pas le signe (puisque $\hat{\psi}_g$ est aléatoire). Nous déduisons :

$$\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g^b; Y)] - L_n(\hat{\psi}_g; y) = \underbrace{\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g^b; Y)] - L_n(\psi_g^b; y)}_{< \nu/2 \quad \text{p.s.}} + \underbrace{L_n(\psi_g^b; y) - L_n(\hat{\psi}_g; y)}_{< \nu/2 \quad \text{p.s.}}, \text{ et}$$

$$\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g^b; Y)] - L_n(\hat{\psi}_g; y) = \underbrace{\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g^b; Y)] - \mathbb{E}_{f^0}[\ln L_{cc}(\hat{\psi}_g; Y)]}_{\geq 0 \quad \text{p.s.}} + \underbrace{\mathbb{E}_{f^0}[\ln L_{cc}(\hat{\psi}_g; Y)] - L_n(\hat{\psi}_g; y)}_{> -\nu/2 \quad \text{p.s.}}.$$

La convergence vers 0 s'effectue donc aussi bien du côté positif que du côté négatif. En prenant une suite dénombrable $(\nu_p)_{p\in\mathbb{N}} = (\frac{1}{2^p})_{p\in\mathbb{N}}$, et puisque l'intersection de deux évènements de probabilité 1 est elle-même de probabilité 1, nous obtenons immédiatement le résultat. □

Ce lemme permet de retrouver (H2-C) : il s'appuie sur la définition de l'estimateur $\hat{\psi}_g$ et l'hypothèse (H3-A). Asymptotiquement, maximiser la moyenne empirique de la log-vraisemblance L_{cc} ne doit pas être très loin de maximiser $\psi_g \mapsto \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; Y)]$ puisqu'ils sont uniformément proches l'un de l'autre. Le même type d'hypothèses que pour pour la convergence forte du $MLccE$ seront donc suffisantes pour avoir (H2-C).

Dans l'optique d'étudier (H4-C), nous nous intéressons à la distance suivante :

$$S_n \ln L_{cc}(\psi; y) = n\left[L_n(\psi; y) - \mathbb{E}_{f^0}[\ln L_{cc}(\psi; Y)]\right] = n\, L_n(\psi; y) - n\, \mathbb{E}_{f^0}[\ln L_{cc}(\psi; Y)].$$

Cette distance représente (à une constante n près) l'écart entre la moyenne empirique et la moyenne théorique de la log-vraisemblance L_{cc} prise en un paramètre quelconque. La démarche consiste à s'intéresser au contrôle de cette "erreur" (appliquée entre $\psi_{g^b}^b$ et $\hat{\psi}_g$) en montrant qu'elle peut s'écrire autrement, et notamment sous une forme qui nous permet d'effectuer des simplifications. Nous allons dans la suite présenter un ensemble de résultats qui vont nous permettre de remonter jusqu'à (H4-C). Tout d'abord, nous utilisons trois résultats existants qui sont issus de Massart (2007).
Introduisons la notation suivante pour la suite : $\forall A$ mesurable avec $\mathbb{P}(A) > 0$, $\forall \phi : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}^A[\phi(X)] = \frac{\mathbb{E}[\phi(X) 1_A(X)]}{\mathbb{P}(A)}$$

Lemme 6. *(Lemme 2.4 dans Massart (2007))*
Soit $Z \in L_1(\mathbb{R})$. Soit la fonction strictement croissante $\varphi : \mathbb{R}^+ \mapsto \mathbb{R}$ telle que pour tout ensemble mesurable A avec $\mathbb{P}(A) > 0$, on a

$$\mathbb{E}^A[Z] \leq \varphi\left(\ln \frac{1}{\mathbb{P}(A)}\right),$$

Alors

$$\forall y > 0, \mathbb{P}[Z \geq \varphi(y)] \leq e^{-y}.$$

Ce lemme est une application de l'inégalité de Markov. Il est très utile pour obtenir des résultats en probabilité à partir de résultats en espérance. C'est d'ailleurs ce lemme qui nous permettra d'utiliser le théorème 8 ci-dessous dans une version où la conclusion vaudrait en probabilité (c'est notre objectif final). En l'occurence, nous avons également besoin du lemme suivant pour démontrer le théorème 8 :

Lemme 7. *(Lemme 4.23 dans Massart (2007))*
Soit S un ensemble dénombrable, $u \in S$ et $a : S \to \mathbb{R}^+$ telle que $a(u) = \inf_{t\in S} a(t)$.
Soit Z un processus indexé par S.
Supposons que $\forall \sigma > 0$, $\mathbb{E}[\sup_{t\in\mathcal{B}(\sigma)} Z(t) - Z(u)] < \infty$ avec $\mathcal{B}(\sigma) = \{t \in S;\, a(t) \leq \sigma\}$.
Alors,
pour une quelconque fonction φ sur $\mathbb{R}^+$ telle que $\dfrac{\varphi(y)}{y}$ est décroissante sur $\mathbb{R}^+$ et satisfait

$$\forall \sigma \geq \sigma_0, \qquad \mathbb{E}\left[\sup_{t\in\mathcal{B}(\sigma)} Z(t) - Z(u)\right] < \varphi(\sigma),$$

nous avons

$$\forall y > \sigma_0, \quad \mathbb{E}\left[\sup_{t\in S} \frac{Z(t) - Z(u)}{a^2(t) + y^2}\right] \leq \frac{4}{y^2}\varphi(y).$$

Ce résultat est tout à fait primordial pour passer d'un contrôle local à un contrôle global des accroissements d'un processus. Il permet de lier les accroissements d'un processus entre deux points par rapport à la distance entre ces deux points. Dans notre cadre, nous l'appliquons à la différence entre l'espérance de la log-vraisemblance L_{cc} en ψ_g^b et son estimateur. Cette différence est censée tendre vers 0 avec une vitesse de convergence en $n^{1/2}$ (théorème central limite) ; mais elle y tend en fait plus vite grâce à ce lemme, avec une vitesse de convergence en n.

Théorème 8. *(Théorème 6.8 dans Massart (2007))*
Soit $\mathcal{F}$ une classe dénombrable de fonctions mesurables à valeurs dans $\mathbb{R}$.
Supposons que

$$\exists\, \sigma > 0,\, \exists\, b > 0 \text{ tel que } \forall f \in \mathcal{F},\, \forall k \geq 2,\, \mathbb{E}\left[|f(Y_i)|^k\right] \leq \frac{k!}{2}\sigma^2 b^{k-2},$$

et $\forall \delta > 0$, $\exists C_\delta$ un ensemble de crochets qui couvrent $\mathcal{F}$ tels que,

$$\forall [g_l, g_u] \in C_\delta,\, \forall k \in \mathbb{N}^* - \{1\}, \quad \mathbb{E}\left[(g_u - g_l)^k(Y_i)\right] \leq \frac{k!}{2}\delta^2 b^{k-2}.$$

Soit $e^{H(\delta)}$ le plus petit cardinal d'une telle couverture de $\mathcal{F}$.

Alors

$\exists \kappa$ une constante absolue telle que $\forall \epsilon \in]0, 1]$, $\forall A$ mesurable avec $\mathbb{P}(A) > 0$,

$$\mathbb{E}^A\left[\sup_{f\in\mathcal{F}} S_n(f)\right] \leq E + (1 + 6\epsilon)\sigma\sqrt{2n \ln \frac{1}{\mathbb{P}(A)}} + 2b \ln \frac{1}{\mathbb{P}(A)}$$

avec $E = \dfrac{\kappa}{\epsilon}\sqrt{n}\int_0^{\epsilon\sigma} \sqrt{\min(H(u), n)}du + 2(b + \sigma)H(\sigma)$.

Avec ce théorème, nous contrôlons le supremum d'un processus empirique sur une certaine classe de fonctions f grâce à des bornes supérieures sur les moments des fonctions dans $\mathcal{F}$, une grille qui couvre $\mathcal{F}$, et donc le nombre de crochets nécessaires à une telle couverture. Ce théorème lie grosso modo le supremum du processus empirique avec la complexité de la classe de fonctions considérée. Nous ne démontrons pas ces résultats ici, mais le lecteur intéressé pourra consulter l'ouvrage de Massart (2007). Grâce à ces résultats, nous pouvons maintenant formuler le lemme suivant.

Lemme 8. *Soit $K_g \in \mathbb{N}^*$ et $\Psi_g \subset \mathbb{R}^{K_g}$ supposé convexe.*
Soit $\Psi_g^{\mathcal{O}}$ un ouvert de $\mathbb{R}^{K_g}$ tel que $\Psi_g \subset \Psi_g^{\mathcal{O}}$. Soit $\ln L_{cc} : \Psi_g^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
Supposons que

- *$\psi_g \in \Psi_g^{\mathcal{O}} \mapsto \ln L_{cc}(\psi_g; y)$ est $f^0 d\lambda$ - presque partout C^1 sur $\Psi_g^{\mathcal{O}}$.*

  Soit $\psi_g^b \in \Psi_g$ tel que $\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g^b; Y)] = \max_{\psi_g \in \Psi_g} \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; Y)]$.

- *Supposons que* $\begin{cases} L(y) = \sup_{\psi_g \in \Psi_g} |\ln L_{cc}(\psi_g; y)| < \infty \quad f^0 d\lambda\text{-p.s.}, \\ \|L\|_\infty = ess\sup_{Y \sim f^0} L(Y) < \infty. \end{cases}$

- *Supposons que* $\begin{cases} L^{'}(y) = \sup_{\psi_g \in \Psi_g} \left\| \left(\frac{\partial \ln L_{cc}}{\partial \psi_g} \right)_{(\psi_g; y)} \right\|_\infty < \infty \quad f^0 d\lambda\text{-p.s.}, \\ \|L^{'}\|_2 = \mathbb{E}_{f^0}[L^{'}(Y)^2]^{\frac{1}{2}} < \infty. \end{cases}$

Alors $\exists\, \alpha > 0$ tel que $\forall n, \forall \beta > 0, \forall \eta > 0$,

$$\mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \frac{S_n\left(\ln L_{cc}(\psi_g) - \ln L_{cc}(\psi_g^b)\right)}{\|\psi_g - \psi_g^b\|_\infty^2 + \beta^2} \leq h(L, L^{'}, \alpha, \beta, \eta, K_g) \right) \geq 1 - e^{-\eta}$$

où $h(L, L^{'}, \alpha, \beta, \eta, K_g) = \frac{\alpha}{\beta^2}\left(\|L^{'}\|_2 \beta \sqrt{nK_g} + (\|L\|_\infty + \|L^{'}\|_2 \beta) K_g + \|L^{'}\|_2 \sqrt{n\eta}\beta + \|L\|_\infty \eta \right)$.

En réalité, ce résultat est la clef de l'hypothèse (H4-C) puisqu'il fournit un moyen de contrôler la distance S_n qui nous intéresse. Nous proposons ci-dessous une preuve de ce lemme afin d'expliciter les étapes du raisonnement.

Démonstration. Vu que nous utilisons le lemme 7 et le théorème 8 dans cette démonstration, nous devons normalement nous plier aux hypothèses qu'ils requièrent, à savoir que la classe des fonctions de vraisemblance classifiante conditionnelle soit dénombrable. Néanmoins, nous pourrions vérifier que ces résultats peuvent être appliqués à un sous-ensemble dense de $\{\ln L_{cc}(\psi_g) : \psi_g \in \Psi_g\}$ qui contienne ψ_g^b, ce qui permet ensuite de généraliser à l'espace entier. Soit le processus empirique centré :

$$\begin{aligned} S_n \ln L_{cc}(\psi_g; Y) &= n\, L_n(\psi_g; Y) - n\, \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; Y)] \\ &= \sum_{j=1}^{n} \left(\ln L_{cc}(\psi_g; Y_j) - \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; Y)] \right) \end{aligned}$$

Considérons une variable muette α ; et gardons en tête que toutes les probabilités et espérances sont formulées sous $f^0 d\lambda$. Soit $\psi_g^b \in \Psi_g$ tel que $\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g^b; Y)\right] = \sup_{\psi_g \in \Psi_g} \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g; Y)\right]$.

Définissons

$$\forall \sigma > 0, \quad \Psi_g(\sigma) = \left\{ \psi_g \in \Psi_g : \ \|\psi_g - \psi_g^b\|_\infty \leq \sigma \right\}.$$

Autrement dit, on se place dans une boule fermée de rayon σ au voisinage de la solution optimale.
Etant donné que $\Psi_g(\sigma)$ est convexe si Ψ_g l'est, nous avons d'une part que $\forall r \in \mathbb{N}^* - \{1\}$,

$$\forall \psi_g \in \Psi_g(\sigma), \ |\ln L_{cc}(\psi_g; y) - \ln L_{cc}(\psi_g^b; y)|^r \leq L^{'}(y)^2 \|\psi_g - \psi_g^b\|_\infty^2 \, (2L(y))^{r-2} \quad f^0 d\lambda - p.s.$$

On obtient par suite que $\forall r \in \mathbb{N}^* - \{1\}, \ \forall \psi_g \in \Psi_g(\sigma)$,

$$\begin{aligned} \mathbb{E}_{f^0}\left[|\ln L_{cc}(\psi_g; Y) - \ln L_{cc}(\psi_g^b; Y)|^r \right] &\leq \|L^{'}\|_2^2 \|\psi_g - \psi_g^b\|_\infty^2 \, (2\|L\|_\infty)^{r-2} \\ &\leq \frac{r!}{2} (\|L^{'}\|_2 \sigma)^2 \left(\frac{2\|L\|_\infty}{2} \right)^{r-2}. \end{aligned}$$

D'autre part, nous pouvons appliquer le lemme 4 qui nous assure que $\forall r \in \mathbb{N}^* - \{1\}, \ \forall \delta > 0$, il existe un ensemble de crochets C_δ qui couvrent $\left\{ (\ln L_{cc}(\psi_g; y) - \ln L_{cc}(\psi_g^b; y)) : \ \psi_g \in \Psi_g(\sigma) \right\}$ (déduit d'un ensemble de crochets qui couvriraient $\{(\ln L_{cc}(\psi_g; y) : \ \psi_g \in \Psi_g(\sigma)\}$) tel que :

$$\forall r \in \mathbb{N}^* - \{1\}, \ \forall [g_l, g_u] \in C_\delta, \quad \|g_u - g_l\|_r \leq \left(\frac{r!}{2} \right)^{\frac{1}{r}} \delta^{\frac{2}{r}} \left(\frac{4\|L\|_\infty}{3} \right)^{\frac{r-2}{r}},$$

et en notant $H(\delta, \Psi_g(\sigma))$ l'entropie, nous aurions au minimum le nombre de crochets :

$$e^{H(\delta, \Psi_g(\sigma))} \leq \max \left(\left(\frac{\overbrace{diam\, \Psi_g(\sigma)}^{\leq 2\sigma} \|L^{'}\|_2}{\delta} \right)^{K_g}, 1 \right). \tag{4.15}$$

Nous utilisons alors le théorème 8 : $\exists \alpha$ constante, $\forall \epsilon \in]0, 1]$, $\forall A$ mesurable avec $\mathbb{P}(A) > 0$,

$$\mathbb{E}^A \left[\sup_{\psi_g \in \Psi_g(\sigma)} S_n(\ln L_{cc}(\psi_g) - \ln L_{cc}(\psi_g^b)) \right] \leq E + (1+6\epsilon)\|L^{'}\|_2 \, \sigma \sqrt{2n \ln \frac{1}{\mathbb{P}(A)}} + \frac{8}{3}\|L\|_\infty \ln \frac{1}{\mathbb{P}(A)} \tag{4.16}$$

avec $E = \frac{\alpha}{\epsilon}\sqrt{n} \int_0^{\epsilon \|L^{'}\|_2 \sigma} \sqrt{H(u, \Psi_g(\sigma))} du + 2 \left(\frac{4}{3}\|L\|_\infty + \|L^{'}\|_2 \sigma \right) H(\|L^{'}\|_2 \sigma, \Psi_g(\sigma))$.
En se servant de l'inégalité de Holder (dérivée de celle de Cauchy-Schwarz), on a : $\forall t \in \mathbb{R}^+$,

$$\begin{aligned} \int_0^t \sqrt{\max\left(\ln \frac{1}{u}, \ln(1) \right)} du &= \int_0^{\min(t,1)} \sqrt{\ln \frac{1}{u}} du \\ &\leq \sqrt{\min(t,1)} \sqrt{\int_0^{\min(t,1)} \ln \frac{1}{u} du} \\ &= \min(t,1) \sqrt{\ln \frac{e}{\min(t,1)}} \end{aligned}$$

A partir de cette observation et en se servant de l'équation (4.16) , nous en déduisons après un changement de variable que $\forall t \in \mathbb{R}^+$,

$$\begin{aligned}\int_0^t \sqrt{H(u,\Psi_g(\sigma))}du &\leq \sqrt{K_g}\int_0^t \sqrt{\max\left(\ln\frac{2\|L^{'}\|_2\sigma}{u},0\right)}du \\ &\leq \sqrt{K_g}\min(t,2\|L^{'}\|_2\sigma)\sqrt{\ln\frac{e}{\min(\frac{t}{2\|L^{'}\|_2\sigma},1)}}. \qquad (4.17)\end{aligned}$$

En appliquant le lemme 7 avec les équations (4.15), (4.16) et (4.17), on obtient

$$\forall\sigma>0,\quad \mathbb{E}_{f^0}\left[\sup_{\psi_g\in\Psi_g(\sigma)} S_n(\ln L_{cc}(\psi_g)-\ln L_{cc}(\psi_g^b))\right]\leq\varphi(\sigma),$$

avec

$$\begin{aligned}\varphi(t)=\frac{\alpha}{\epsilon}\sqrt{n}\sqrt{K_g}\epsilon\|L^{'}\|_2 t\sqrt{\ln\frac{2e}{\epsilon}} &+ 2\left(\frac{4}{3}\|L\|_\infty+\|L^{'}\|_2 t\right)K_g\ln 2 \\ &+ (1+6\epsilon)\|L^{'}\|_2\, t\sqrt{2n\ln\frac{1}{\mathbb{P}(A)}}+\frac{8}{3}\|L\|_\infty\ln\frac{1}{\mathbb{P}(A)}.\end{aligned}$$

Nous pouvons vérifier que $\dfrac{\varphi(t)}{t}$ est décroissante, d'où la conclusion : $\forall\beta>0$

$$\mathbb{E}^A\left[\sup_{\psi_g\in\Psi_g}\frac{S_n(\ln L_{cc}(\psi_g)-\ln L_{cc}(\psi_g^b))}{\|\psi_g^b-\psi_g\|_\infty^2+\beta^2}\right]\leq\frac{4}{\beta^2}\varphi(\beta).$$

En choisissant $\epsilon=1$ et en appliquant le lemme 6, nous obtenons : $\forall\eta>0$, $\forall\beta>0$,

$$\mathbb{P}\left(A\leq B\right)>1-e^{-\eta},$$

$$\text{avec}\begin{cases}A=\sup\limits_{\psi_g\in\Psi_g}\dfrac{S_n(\ln L_{cc}(\psi_g)-\ln L_{cc}(\psi_g^b))}{\|\psi_g-\psi_g^b\|_\infty^2+\beta^2},\\ B=\dfrac{\alpha}{\beta^2}\left(\sqrt{nK_g}\|L^{'}\|_2\beta\sqrt{\ln 2e}+(\|L\|_\infty+\|L^{'}\|_2\beta)K_g\ln 2+\|L^{'}\|_2\beta\sqrt{n\eta}+\|L\|_\infty\eta\right).\end{cases}$$

□

Plus concrètement, le corollaire suivant donne les conditions suffisantes pour obtenir l'hypothèse (H4-C). De fait, ce corollaire est issu d'une application directe du lemme 8, où les constantes sont bien choisies. L'utilisation du lemme apparait dans la preuve du corollaire.

Corollaire 1. *Soit $K_g \in \mathbb{N}^*$ et $\Psi_g \subset \mathbb{R}^{K_g}$ convexe.*
Soit $\Psi_g^{\mathcal{O}}$ un ouvert de $\mathbb{R}^{K_g}$ tel que $\Psi_g \subset \Psi_g^{\mathcal{O}}$. Soit $\ln L_{cc} : \Psi_g^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
Supposons que :

- *$\psi_g \in \Psi_g^{\mathcal{O}} \mapsto \ln L_{cc}(\psi_g; y)$ est $f^0 d\lambda$- presque partout C^1 sur $\Psi_g^{\mathcal{O}}$.*
Soit $\psi_g^b \in \Psi_g$ tel que $\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g^b; Y)] = \max_{\psi_g \in \Psi_g} \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; Y)]$.

- *Supposons que* $\begin{cases} L(y) = \sup\limits_{\psi_g \in \Psi_g^{\mathcal{O}}} |\ln L_{cc}(\psi_g; y)| < \infty \quad f^0 d\lambda\text{-p.s.}, \\ \|L\|_\infty = ess\sup\limits_{Y \sim f^0} L(Y) < \infty. \end{cases}$

- *Supposons que* $\begin{cases} L'(y) = \sup\limits_{\psi_g \in \Psi_g^{\mathcal{O}}} \left\| \left(\dfrac{\partial \ln L_{cc}}{\partial \psi_g} \right)_{(\psi_g; y)} \right\|_\infty < \infty \quad f^0 d\lambda\text{-p.s.}, \\ \|L'\|_2 = \mathbb{E}_{f^0}[L'(Y)^2]^{\frac{1}{2}} < \infty. \end{cases}$

- *Supposons de plus que $I_{\psi_g^b} = \dfrac{\partial^2}{\partial \psi_g^2} \Big(\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; Y)] \Big)_{|\psi_g^b}$ est inversible.*

Soit $(\hat{\psi}_n)_{n \geq 1}$ tel que $\hat{\psi}_n \in \Psi_g$, avec $\begin{cases} L_n(\hat{\psi}_n) \geq L_n(\psi_g^b) - O_{\mathbb{P}}\left(\dfrac{1}{n}\right), \\ \hat{\psi}_n \xrightarrow[n \to \infty]{\mathbb{P}} \psi_g^b. \end{cases}$

Alors

$$n \, \|\hat{\psi}_n - \psi_g^b\|_\infty^2 = O_{\mathbb{P}}(1).$$

La constante inclue dans $O_{\mathbb{P}}(1)$ dépend de K_g, $\|L\|_\infty$, $\|L'\|_2$ et $I_{\psi_g^b}$.

Démonstration. Soit $\epsilon > 0$ tel que la boule fermée $B(\psi_g^b, \epsilon) \subset \Psi_g^{\mathcal{O}}$. Si l'on suppose comme dans ce corollaire que $\hat{\psi}_n \xrightarrow[n \to \infty]{\mathbb{P}} \psi_g^b$, il existe un rang $n_0 \in \mathbb{N}^*$ à partir duquel $\hat{\psi}_n \in B(\psi_g^b, \epsilon)$ avec grande probabilité. Puisque $B(\psi_g^b, \epsilon)$ est convexe, et que les hypothèses du corollaire sont conformes avec l'application du lemme 8, nous l'appliquons à $\hat{\psi}_n$: $\forall n \geq n_0, \forall \beta > 0, \forall \eta > 0$,

$$\mathbb{P}(A \leq B) > 1 - e^{-\eta}, \tag{4.18}$$

avec $\begin{cases} A = \dfrac{S_n(\ln L_{cc}(\hat{\psi}_n) - \ln L_{cc}(\psi_g^b))}{\|\hat{\psi}_n - \psi_g^b\|_\infty^2 + \beta^2}, \\ B = \dfrac{\alpha}{\beta^2} \left(\sqrt{n K_g} \|L'\|_2 \beta + (\|L\|_\infty + \|L'\|_2 \beta) K_g + \|L'\|_2 \beta \sqrt{n\eta} + \|L\|_\infty \eta \right). \end{cases}$

De plus, on suppose que $I_{\psi_g^b}$ est inversible, ce qui permet d'écrire que $\forall \psi_g \in B(\psi_g^b, \epsilon)$,

$$\begin{aligned} \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g)] - \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g^b)\right] &= (\psi_g - \psi_g^b)^T I_{\psi_g^b} (\psi_g - \psi_g^b) + r(\|\psi_g - \psi_g^b\|_\infty^2) \|\psi_g - \psi_g^b\|_\infty^2 \\ &\geq \|\psi_g - \psi_g^b\|_\infty^2 \left(2\alpha' + r(\|\psi_g - \psi_g^b\|_\infty^2) \right), \end{aligned}$$

où $\alpha^{'} > 0$ dépend de $I_{\psi_g^b}$ et la fonction $r : \mathbb{R}^+ \mapsto \mathbb{R}$ satisfait $r(y) \underset{y \to 0}{\longrightarrow} 0$.
Ainsi, pour $\|\psi_g - \psi_g^b\|_\infty$ assez petit, nous obtenons

$$\forall \psi_g \in B(\psi_g^b, \epsilon), \quad \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g)\right] - \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g^b)\right] \geq \alpha^{'} \|\psi_g - \psi_g^b\|_\infty^2. \tag{4.19}$$

Puisque $S_n(\ln L_{cc}(\hat{\psi}_n) - \ln L_{cc}(\psi_g^b)) = n(L_n(\hat{\psi}_n) - L_n(\psi_g^b)) + n\,\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g^b) - \ln L_{cc}(\hat{\psi}_n)\right]$,

$$\text{on obtient} \qquad S_n(\ln L_{cc}(\hat{\psi}_n) - \ln L_{cc}(\psi_g^b)) \geq n\,\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g^b) - \ln L_{cc}(\hat{\psi}_n)\right] - O_{\mathbb{P}}(1). \tag{4.20}$$

D'après les équations (4.18), (4.19) et (4.20), on a après quelques développements

$$\mathbb{P}\left(n\|\hat{\psi}_n - \psi_g^b\|_\infty^2 \leq \frac{\|L^{'}\|_2(\sqrt{nK_g} + \sqrt{\eta n} + K_g)\beta + \|L\|_\infty(K_g + \eta) + O_{\mathbb{P}}(1)}{\frac{\alpha^{'}}{\alpha} - \frac{1}{n\beta^2}\left(\|L^{'}\|_2(\sqrt{nK_g} + \sqrt{\eta n} + K_g)\beta + \|L\|_\infty(K_g + \eta)\right)}\right) > 1 - e^{-\eta},$$

tant que le dénominateur du membre gauche est positif.
Il suffit ensuite de choisir β pour que cette condition soit satisfaite, et pour que le membre gauche de la dernière équation soit majoré par une quantité qui ne dépende pas de n pour obtenir le résultat. Essayons avec $\beta = \frac{\beta_0}{\sqrt{n}}$, où β_0 est indépendant de n :

$$\mathbb{P}\left(n\|\hat{\psi}_n - \psi_g^b\|_\infty^2 \leq \frac{\|L^{'}\|_2(\sqrt{K_g} + \sqrt{\eta} + \frac{K_g}{\sqrt{n}})\beta_0 + \|L\|_\infty(K_g + \eta) + O_{\mathbb{P}}(1)}{\frac{\alpha^{'}}{\alpha} - \frac{1}{\beta_0^2}\left(\|L^{'}\|_2(\sqrt{K_g} + \sqrt{\eta} + \frac{K_g}{\sqrt{n}})\beta_0 + \|L\|_\infty(K_g + \eta)\right)}\right) > 1 - e^{-\eta},$$

soit

$$\mathbb{P}\left(n\|\hat{\psi}_n - \psi_g^b\|_\infty^2 \leq \frac{\|L^{'}\|_2(\sqrt{K_g} + \sqrt{\eta} + K_g)\beta_0 + \|L\|_\infty(K_g + \eta) + O_{\mathbb{P}}(1)}{\frac{\alpha^{'}}{\alpha} - \frac{1}{\beta_0^2}\left(\|L^{'}\|_2(\sqrt{K_g} + \sqrt{\eta} + K_g)\beta_0 + \|L\|_\infty(K_g + \eta)\right)}\right) > 1 - e^{-\eta}.$$

L'idée maintenant est de choisir un β_0 assez grand pour que cette inégalité soit toujours vérifiée. Ensuite le résultat est immédiat : $\forall \eta > 0$, $\exists n_0 \in \mathbb{N}^*$, $\forall n \geq n_0$,

$$\mathbb{P}\left(n\|\hat{\psi}_n - \psi_g^b\|_\infty^2 = CO_{\mathbb{P}}(1)\right) > 1 - e^{-\eta},$$

avec C qui dépend de K_g, $\|L\|_\infty$, $\|L^{'}\|_2$, $I_{\psi_g^b}$ et η. □

La dépendance de $O_{\mathbb{P}}(1)$ en K_g, $\|L\|_\infty$, $\|L^{'}\|_2$ et $I_{\psi_g^b}$ n'est pas problématique dans le sens où nous souhaitons un résultat asymptotique sur l'ordre de $\|\hat{\psi}_n - \psi_g^b\|_\infty^2$ par rapport à n. L'hypothèse sur $I_{\psi_g^b}$ joue un rôle similaire à l'hypothèse (H2-A) du théorème 4 : en effet, elle assure que $\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g; Y)\right]$ ne puisse être proche de $\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g^b; Y)\right]$ si ψ_g n'est pas proche de ψ_g^b. Cependant, cette hypothèse est plus forte : elle permet aussi de contrôler la relation entre $\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g; Y)\right] - \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_g^b; Y)\right]$ et $\|\psi_g - \psi_g^b\|_\infty$ pour nous permettre de conclure. Afin de replacer ce résultat dans le contexte de sélection de modèle, nous devons généraliser ce résultat à un univers de modèles ayant des dimensions différentes.

Corollaire 2. *Soit $\{M_g\}_{1\leq g\leq m}$ une collection de modèles de paramètres $\{\psi_g\}_{1\leq g\leq m}\in\{\Psi_g\}_{1\leq g\leq m}$ et de dimension $\{K_g\}_{1\leq g\leq m}$, avec $\Psi_g\subset\mathbb{R}^{K_g}$. Ces modèles sont classés dans un ordre croissant de complexité, avec $K_1\leq K_2\leq ...\leq K_m$. Supposons que :*

(1) Pour g quelconque, il existe $\Psi_g^{\mathcal{O}}$ un ouvert de $\mathbb{R}^{K_g}$ tel que $\Psi_g\subset\Psi_g^{\mathcal{O}}$.
$\psi\in\Psi^{\mathcal{O}}\mapsto\ln L_{cc}(\psi;y)$ est $f^0d\lambda$- presque partout C^1 sur $\Psi^{\mathcal{O}}$, avec $\Psi^{\mathcal{O}}=\Psi_1^{\mathcal{O}}\cup...\cup\Psi_m^{\mathcal{O}}$;

(2) Supposons que $\begin{cases} L(y)=\sup\limits_{\psi\in\Psi^{\mathcal{O}}}|\ln L_{cc}(\psi;y)|<\infty \quad f^0d\lambda\text{-p.s.},\\ \|L\|_\infty=ess\sup\limits_{Y\sim f^0}L(Y)<\infty. \end{cases}$

(3) Supposons que $\begin{cases} L^{'}(y)=\sup\limits_{\psi\in\Psi^{\mathcal{O}}}\left\|\left(\dfrac{\partial\ln L_{cc}}{\partial\psi}\right)_{(\psi;y)}\right\|_\infty<\infty \quad f^0d\lambda\text{-p.s.},\\ \|L^{'}\|_2=\mathbb{E}_{f^0}[L^{'}(Y)^2]^{\frac{1}{2}}<\infty. \end{cases}$

$\forall g\in[\![1;m]\!]$, *soit* $\Psi_g^b=\underset{\psi_g\in\Psi_g}{\arg\max}\,\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g)]$.

Soit $\psi_g^b\in\Psi_g^b$, *et* $g^b=\min\left(\arg\max\limits_{1\leq g\leq m}\mathbb{E}_{f^0}[\ln L_{cc}(\Psi_g^b)]\right)$.

(4) Supposons que $\forall g\in[\![1;m]\!]$, $\quad\forall\psi_g\in\Psi_g$,
$$\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g;Y)]=\mathbb{E}_{f^0}[\ln L_{cc}(\psi_{g^b}^b)]\Longleftrightarrow\ln L_{cc}(\psi_g;y)=\ln L_{cc}(\psi_{g^b}^b;y)\quad f^0d\lambda\text{- p.s.}$$
Soit $K=\left\{g\in[\![1,m]\!]:\ \mathbb{E}_{f^0}[\ln L_{cc}(\Psi_g^b)]=\mathbb{E}_{f^0}[\ln L_{cc}(\Psi_{g^b}^b)]\right\}$.

(5) $\forall g\in K$, *soit* $\hat{\psi}_g=\hat{\psi}_g(Y_1,...,Y_n)\in\Psi_g$ *tel que* $\begin{cases} L_n(\hat{\psi}_g)\geq L_n(\psi_g^b)-O_{\mathbb{P}}\left(\dfrac{1}{n}\right),\\ \hat{\psi}_g\xrightarrow[n\to\infty]{\mathbb{P}}\psi_g^b. \end{cases}$

(6) Supposons de plus $I_{\psi_g^b}=\dfrac{\partial^2}{\partial\psi_g^2}\Big(\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g;Y)]\Big)_{|\psi_g^b}$ *inversible pour $g\in K$.*

Alors
$$\forall g\in K,\ n\left(L_n(\hat{\psi}_g)-L_n(\hat{\psi}_{g^b})\right)=O_{\mathbb{P}}(1).$$

Démonstration. On applique le corollaire 1. Soit $g\in K:\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g^b;Y)]=\mathbb{E}_{f^0}[\ln L_{cc}(\psi_{g^b}^b;Y)]$. Ψ_g est supposé être convexe : s'il ne l'est pas, nous savons néanmoins qu'asymptotiquement $\hat{\psi}_g$ appartient à la boule (convexe) $B(\psi_{g^b}^b,\epsilon)$ avec une grande probabilité. En prenant ϵ assez petit pour garantir que $B(\psi_{g^b}^b,\epsilon)\subset\Psi_g^{\mathcal{O}}$, Ψ_g peut ainsi être remplacé par $B(\psi_{g^b}^b,\epsilon)$. D'après le lemme 8 et les hypothèses qui sont posées : $\exists n_0\in\mathbb{N}^*$, $\forall n\geq n_0$, $\forall\beta>0$, on a
$$\mathbb{P}(A\leq B)>1-e^{-\eta},\tag{4.21}$$
avec $\begin{cases} A=S_n(\ln L_{cc}(\hat{\psi}_g)-\ln L_{cc}(\psi_g^b)),\\ B=\alpha\dfrac{\|\hat{\psi}_g-\psi_g^b\|_\infty^2+\beta^2}{\beta^2}\left(\|L^{'}\|_2(\sqrt{nK_g}+\sqrt{\eta n}+K_g)\beta+\|L\|_\infty(K_g+\eta)\right). \end{cases}$

En posant $\beta=\dfrac{\beta_0}{\sqrt{n}}$ pour tout $\beta_0>0$, nous obtenons finalement
$$B=\alpha\frac{n\|\hat{\psi}_g-\psi_g^b\|_\infty^2+\beta_0^2}{\beta_0^2}\left(\|L^{'}\|_2(\sqrt{K_g}+\sqrt{\eta}+\frac{K_g}{\sqrt{n}})\beta_0+\|L\|_\infty(K_g+\eta)\right).\tag{4.22}$$

De plus, on a

$$\begin{aligned} S_n(\ln L_{cc}(\hat{\psi}_g) - \ln L_{cc}(\psi_g^b)) &= n(L_n(\hat{\psi}_g) - L_n(\psi_g^b)) + n(\mathbb{E}_{f^0}\left[L_n(\psi_g^b)\right] - \mathbb{E}_{f^0}\left[L_n(\hat{\psi}_g)\right]); \\ &\geq n(L_n(\hat{\psi}_g) - L_n(\psi_g^b)). \end{aligned}$$

Ainsi, d'après les équations 4.21 et 4.22, et d'après le précédent corollaire pour lequel nous avons posé les hypothèses adéquates (selon lequel on a avec grande probabilité lorsque n devient grand : $n\,\|\hat{\psi}_g - \psi_g^b\|_\infty^2 = O_{\mathbb{P}}(1)$),

$$n(L_n(\hat{\psi}_g) - L_n(\psi_g^b)) = O_{\mathbb{P}}(1).$$

Ceci est valable pour tout $g \in K$, et donc en particulier pour g^b. Enfin, $L_n(\psi_g^b) = L_n(\psi_{g^b}^b)$ puisque par hypothèse $\ln L_{cc}(\psi_g^b; y) = \ln L_{cc}(\psi_{g^b}^b; y) \quad f^0 d\lambda$-p.p. D'où la conclusion :

$$n(L_n(\hat{\psi}_g) - L_n(\hat{\psi}_{g^b})) = O_{\mathbb{P}}(1).$$

□

Ce corollaire généralise le corollaire précédent à un contexte de sélection de modèle (nous retrouvons (H4-C)). Tous ces résultats permettent encore une fois de formuler différemment le théorème 7, afin d'introduire de nouvelles hypothèses directement exploitables pour aisément vérifier la consistance du critère de sélection ICL_c pour des mélanges de GLMs. Le nouveau théorème s'écrit dans le cas où l'espace des paramètres est compact :

Théorème 9. *(Consistance faible du critère de sélection pénalisé, cas compact).*
Soit $\{M_g\}_{1\leq g\leq m}$ une collection de modèles de paramètres $\{\psi_g\}_{1\leq g\leq m} \in \{\Psi_g\}_{1\leq g\leq m}$ et de dimension $\{K_g\}_{1\leq g\leq m}$, avec $\Psi_g \subset \mathbb{R}^{K_g}$. Ces modèles sont classés dans un ordre croissant de complexité, avec $K_1 \leq K_2 \leq ... \leq K_m$. Supposons que :

(H1-D) $\forall g \in [\![1, m]\!]$, Ψ_G *est un ensemble compact.*
Alors pour g quelconque, posons $\Psi_g^b = \arg\max_{\psi_g \in \Psi_g} \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; Y)]$.
Définissons $g^b = \min\left(\arg\max_{1\leq g\leq m} \mathbb{E}_{f^0}\left[\ln L_{cc}(\Psi_g^b; Y)\right]\right)$;

(H2-D) $\forall g \in [\![1; m]\!]$, $\forall \psi_g \in \Psi_g$, $\forall \psi_{g^b}^b \in \Psi_{g^b}^b$,
$\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g)] = \mathbb{E}_{f^0}[\ln L_{cc}(\psi_{g^b}^b)] \Longleftrightarrow \ln L_{cc}(\psi_g; y) = \ln L_{cc}(\psi_{g^b}^b; y) \quad f^0 d\lambda$- *p.s.*

$\forall g \in [\![1; m]\!]$, *soit $\Psi_g^{\mathcal{O}}$ ouvert de $\mathbb{R}^{K_g}$ sur lequel $\ln L_{cc}$ est définie, avec $\Psi_g \subset \Psi_g^{\mathcal{O}}$.*

(H3-D) $\forall g \in [\![1, m]\!]$, $\begin{cases} L_g(y) = \sup\limits_{\psi_g \in \Psi_g^{\mathcal{O}}} |\ln L_{cc}(\psi_g; y)| < \infty \quad f^0 d\lambda\text{-p.s.}, \\ \|L_g\|_\infty < \infty. \end{cases}$

(H4-D) $\forall g \in [\![1, m]\!]$, $\begin{cases} L_g'(y) = \sup\limits_{\psi_g \in \Psi_g^{\mathcal{O}}} \left\| \left(\dfrac{\partial \ln L_{cc}}{\partial \psi_g} \right)_{(\psi_g; y)} \right\|_\infty < \infty \quad f^0 d\lambda\text{-p.s.}, \\ \|L_g'\|_2 < \infty. \end{cases}$

(H5-D) $\forall g \in [\![1, m]\!], \forall \psi_g^b \in \Psi_g^b$, $I_{\psi_g^b} = \dfrac{\partial^2}{\partial \psi_g^2}\Big(\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; y)]\Big)_{|\psi_g^b}$ *est inversible.*

(H6-D) $\forall g \in [\![1, m]\!]$, $\begin{cases} pen(K_g) > 0 \text{ et } pen(K_g) = o_{\mathbb{P}}(n) \text{ quand } n \to +\infty; \\ \Big(pen(K_g) - pen(K_{g'})\Big) \xrightarrow{\mathbb{P}} \infty \text{ quand } n \to +\infty \text{ et } g > g'. \end{cases}$

Quel que soit g et n, soit $\hat{\psi}_g^{ML_{cc}E} = \hat{\psi}_g^{ML_{cc}E}(Y_1, ..., Y_n) \in \Psi_g$ un estimateur tel que

$$\ln L_{cc}(\hat{\psi}_g^{ML_{cc}E}; Y) \geq \ln L_{cc}(\psi_g^b; Y) - o_{\mathbb{P}}(n);$$

Sélectionnons $\hat{g}$ tel que

$$\hat{g} = \underset{1 \leq g \leq m}{\arg\min}\{-\ln L_{cc}(\hat{\psi}_g^{ML_{cc}E}; y) + pen(K_g)\},$$

Alors

$$\mathbb{P}(\hat{g} \neq g^b) \underset{n \to \infty}{\longrightarrow} 0.$$

Démonstration. Elle reprend les mêmes ingrédients que la preuve du théorème 7, tout en insérant les preuves des lemmes et corollaires que nous venons de voir pour satisfaire les hypothèses (H1-C) à (H4-C). Il n'y a donc pas de difficulté particulière. □

4.3 Extension aux mélanges de GLMs

Quelques auteurs s'intéressent actuellement aux mélanges de modèles linéaires généralisés ; notamment Grun and Leisch (2004), Grun and Leisch (2007), Grun and Leisch (2008) et Leisch (2008). Ils développent en parallèle une librairie R dont nous nous servirons dans les applications, nommée "flexmix", et étudient aussi bien les problèmes d'identifiabilité que ceux d'estimation des paramètres. Par exemple, Leisch (2008) montre dans une partie de ses travaux à quel point certains points aberrants dans les observations peuvent affecter la distribution du mélange final. L'élément qui ressort de notre étude bibliographique est qu'il n'existe aucun développement théorique sur les propriétés d'un critère de sélection pour modèle mélange de GLMs satisfaisant des objectifs de classification. Nous proposons ici de démontrer la convergence du critère ICL_c dans le cadre de la sélection de mélange de GLMs. Cette convergence sera établie au prix de certaines hypothèses inhérentes à cette classe de modèles.

4.3.1 Les GLMs : présentation et concepts

Nous avons déjà utilisé la régression logistique dans les chapitres précédents, sans pour autant l'introduire dans un contexte plus général que sont les modèles linéaires généralisés. Il nous apparaît indispensable de présenter de manière formalisée cette extension naturelle, puisque nous allons par la suite démontrer des résultats sur cette famille de modèles. Les GLMs incluent non seulement la régression linéaire, mais aussi les modèles d'analyse de variance (ou modèles factoriels), les modèles logit et probit pour des variables réponses sous forme de taux, les modèles log-linéaires pour les données de comptage ou encore les modèles à réponse multinomiale.
En pratique les mesures que nous utilisons comme variable réponse contiennent des erreurs dont la distribution n'est pas forcément gaussienne, ce qui explique en grande partie l'utilité des GLMs. En Actuariat, ces modèles sont très populaires car ils ont permis le développement de techniques sophistiquées de tarification, permettant aux assureurs d'effectuer une segmentation des risques de leur portefeuille (Ohlson and Johansson (2010)). La suite de l'introduction à ces modèles est fortement inspirée du livre de référence en la matière, McCullagh and Nelder (1989), ainsi que de la présentation synthétique proposée dans Dutang (2011). Cette partie n'a pas vocation à être exhaustive, mais doit donner au lecteur les éléments suffisants pour sa compréhension des paragraphes traitant de la sélection de mélange de GLMs.

Le modèle linéaire

Pour commencer formalisons les hypothèses et résultats connus dans le cadre simple du modèle linéaire, afin d'être capable par la suite de comprendre en quoi les GLMs en sont une extension immédiate.
Soit la matrice $\mathcal{X} \in \mathcal{M}_{np}(\mathbb{R})$ dont les p colonnes contiennent les variables explicatives (exemple : âge, sexe, ...), et les n lignes sont les valeurs observées de ces covariables par individu. Nous appelons X la matrice de schéma (ou design) suivante :

$$X = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \dots & \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix}$$

Nous désignons par Y la variable réponse (à expliquer), avec $Y = (Y_1, ..., Y_n)^T$ où $Y_j \in \mathbb{R}^d$ ($d = 1$ en unidimensionnel pour l'individu j). Le vecteur $X_j = (X_{j1}, ..., X_{jp})$ représente les facteurs explicatifs de l'individu j.
Sous forme matricielle, le modèle linéaire établit la relation suivante entre X et Y :

$$Y = X\beta + \epsilon$$

où $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ est le vecteur des paramètres à estimer, et $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ est l'erreur commise lors de la mesure de Y (les mesures y sont entachées d'un bruit).
L'étude du modèle linéaire nécessite de poser les hypothèses suivantes :

1. identification (non-colinéarité des covariables) : $rang(X) = p < n$, où n est le nombre d'individus et p le nombre de covariables considérées ;
2. bruit blanc ou résidus centrés : $\mathbb{E}[\epsilon_j | X_j] = 0$;
3. non-corrélation des résidus : $\forall k \neq j, \mathbb{E}[\epsilon_k \epsilon_j \mid X_k, X_j] = 0$;
4. homoscédasticité (variance constante) : $\forall j \in [\![1, n]\!], Var(\epsilon_j | X_j) = \sigma^2$;
5. normalité des résidus : $\epsilon_j | X_j \sim \mathcal{N}(0, \sigma^2)$.

Le théorème de Gauss-Markov garantit que l'estimateur des moindres carrés est le meilleur estimateur linéaire (il est sans biais et de variance minimale). Cet estimateur $\hat{\beta}$ minimise l'erreur L^2 entre l'observation Y_{obs} et la réponse modélisée Y_{mod} :

$$\hat{\beta} = \arg\min_{\beta} \sum_{j=1}^{n} (Y_{j,obs} - Y_{j,mod})^2 = \arg\min_{\beta} \sum_{j=1}^{n} (Y_{j,obs} - X_j\beta)^2.$$

Cet estimateur est donné par $\hat{\beta} = \dfrac{(X^T Y)}{(X^T X)^{-1}}$. Nous trouvons par la même méthode un estimateur de la variance des résidus, donné par $\hat{\sigma^2} = \dfrac{1}{n-p} \sum_{j=1}^{n} (Y_j - X_j\hat{\beta})^2$.
Quelques propriétés de cet estimateur sont bien connues, notamment :

- $\hat{\beta}$ est un vecteur gaussien ;
- $\hat{\beta}$ est indépendant de $\hat{\sigma}^2 \sim \chi^2_{n-p}$;
- $Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$.

Ces propriétés sont fondamentales pour la construction d'intervalles de confiance de nos estimateurs.

Remarque 1 : l'hypothèse 5 induit que l'estimateur des moindres carrés est identique à l'estimateur par maximum de vraisemblance.

Remarque 2 : c'est un abus de langage que de parler de relation linéaire entre Y et X. En fait c'est une relation affine du vecteur β qui nous intéresse. Voici quelques exemples de modèles de régression linéaire :

- $Y_j = \beta_0 + \beta_1 \ln x_j + \epsilon_j$;
- $Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \epsilon_j$;
- le modèle logistique présenté au chapitre 1 est un modèle linéaire.

A titre de contre exemple, le modèle $Y_j = \beta_0 + \exp(\beta_1 x_j) + \epsilon_j$ n'est pas linéaire...

En regardant les hypothèses, nous pouvons d'ores et déjà remarquer que certaines d'entre elles sont fortes. Lorsque nous essayons d'estimer numériquement les paramètres, une colinéarité dans la matrice de schéma empêche l'inversibilité de celle-ci et ne permet donc pas d'obtenir les résultats. Une solution consiste donc à enlever une à une les variables explicatives. D'autre part la variance entre individus est supposée constante : nous pouvons nous affranchir de cette hypothèse par l'utilisation de certaines transformations (exemple : Box-Cox), mais ces transformations ne sont pas toujours satisfaisantes au vu des données. Enfin, nous constatons que la principale limitation de la régression linéaire concerne le domaine de définition de la variable réponse. En effet, le prédicteur est un réel, alors que parfois nous aimerions que notre réponse ait un domaine de définition différent, à ajuster suivant nos applications.

Evolution vers les GLMs

Pour présenter de manière simple et claire les GLMs, il suffit de se rendre compte que ce type de modèles est caractérisé par trois entités :

1. une **entité aléatoire** qui dicte la loi de l'erreur : Y_j suit une distribution de la famille exponentielle $\mathcal{F}_{exp}(\theta_j, \phi_j, a, b, c)$, avec paramètres θ_j et ϕ_j ; et fonctions a, b et c ;
2. une **entité déterministe** qu'est le prédicteur, souvent noté η : $\eta_j = X_j\beta$;
3. une fonction de **lien** l, monotone, dérivable, et qui admet une fonction réciproque notée l^{-1}. Elle lie la moyenne μ_j de la réponse au prédicteur : $l(\mathbb{E}[Y_j]) = \eta_j$.

En prenant $l = Id$ et $Y \sim \mathcal{N}(X\beta, \sigma^2)$, nous retombons ainsi sur le modèle linéaire classique. Cependant, il existe de multiples choix possibles pour la fonction de lien l ainsi que pour la loi de Y. Le tableau 4.1 résume ces choix et leur application principale dans le domaine de l'assurance. Nous pouvons calculer la vraisemblance d'un modèle linéaire généralisé grâce à l'expression générale de la densité de la famille exponentielle. En effet, si Y suit une loi de la famille exponentielle alors sa densité est donnée par :

$$f_Y(y_j; \theta, \phi) = \exp\left\{\frac{y_j\theta - b(\theta)}{a(\phi)} + c(y_j, \phi)\right\}, \tag{4.23}$$

où $a(.)$, $b(.)$ et $c(.)$ sont des fonctions spécifiques suivant le modèle considéré. La fonction $a(\phi)$ est de la forme $\frac{\phi}{\omega}$, où ω correspond à un poids (une "exposition" dans le jargon assurantiel), très souvent constant égal à 1 (cas individuel). D'un point de vue vocabulaire, nous appelons :

- lien canonique : tout lien qui permet de vérifier $\theta_j = \mu_j$, où $\mu_j = \mathbb{E}[Y_j]$;

TABLE 4.1 – GLMs et différentes lois de la famille exponentielle associées.

Loi	Notation	Lien	Moyenne	Utilisation
Normale	$\mathcal{N}(\mu, \sigma^2)$	Id : $\eta = \mu$	$\mu = X\beta$	Régression linéaire
Bernoulli	$\mathcal{B}(\mu)$	logit : $\eta = \ln(\frac{\mu}{1-\mu})$	$\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)}$	Taux
Poisson	$\mathcal{P}(\mu)$	log : $\eta = \ln(\mu)$	$\mu = \exp(X\beta)$	Fréquence
Gamma	$\mathcal{G}(\mu, \nu)$	inverse : $\eta = \frac{1}{\mu}$	$\mu = (X\beta)^{-1}$	Sévérité
Inverse Gaussienne	$\mathcal{IN}(\mu, \lambda)$	inverse^2 : $\eta = -\frac{1}{\mu^2}$	$\mu = (X\beta)^{-2}$	Sévérité

– paramètre de tendance : le paramètre θ_j ;
– paramètre de dispersion : le paramètre ϕ_j.

Nous pouvons expliciter ces paramètres et fonctions dans chacune des classes de la famille GLM, en fonction des paramètres initiaux des distributions. Un autre atout de cette représentation réside dans la facilité avec laquelle de nombreux résultats peuvent être dérivés. A partir de l'équation (4.23), nous obtenons par exemple directement la log-vraisemblance pour une observation y_j :

$$\ln L(\theta, \phi; y_j) = \ln f_Y(y_j; \theta, \phi) = \frac{y_j\theta - b(\theta)}{a(\phi)} + c(y_j, \phi).$$

Il est aisé de remonter au calcul de l'espérance et de la variance de la loi des Y : pour cela, les formules (4.24) et (4.25) peuvent être utilisées (McCullagh and Nelder (1989), p.29). De même, cet ouvrage détaille les méthodes numériques de calibration des modèles, ainsi que les tests concernant la bonne adéquation du modèle aux données observées (voir p. 33, 37 et 42). Afin de définir les ensembles de définition des différents paramètres en adoptant la forme d'écriture générale de la densité de la famille exponentielle, nous utiliserons essentiellement les propriétés suivantes :

$$\mathbb{E}[Y] = b^{'}(\theta), \tag{4.24}$$
$$\mathbb{V}ar[Y] = b^{''}(\theta)a(\phi), \tag{4.25}$$

où $b^{'}(\theta)$ et $b^{''}(\theta)$ désignent les dérivées première et seconde de $b(.)$ par rapport à θ.

4.3.2 Caractéristiques des mélanges de GLMs

Définissons tout d'abord les mélanges de GLMs auxquels nous nous intéressons. Il s'agit de mélanges **discrets** (à support discret) dont les composantes appartiennent toutes à la même famille de GLM, ce qui est concrètement le type de modèles utilisé dans la pratique. Ainsi, l'ensemble de ces modèles est défini par

$$M_G = \left\{ f(.; \psi_G) = \sum_{i=1}^{G} \pi_i \, f_{glm}(.; \theta_i, \phi_i) \mid \psi_G = (\pi_1, ..., \pi_G, \theta_1, ..., \theta_G, \phi_1, ..., \phi_G) \in \Psi_G \right\},$$

avec $\Psi_G \subset \Pi_G \times (\mathbb{R}^d)^{2G}$, et $f_{glm}(y; \theta_i, \phi_i) = \exp\left\{ \frac{y\theta_i - b(\theta_i)}{a(\phi_i)} + c(y; \phi_i) \right\}$.

Notre objectif est d'étudier le comportement de la vraisemblance classifiante conditionnelle pour chaque classe de fonction appartenant à la famille des GLMs. Les fonctions $a()$, $b()$, et $c()$ constituent également un intérêt certain dans le but de potentiellement identifier des similitudes (entre les différents GLMs) sur les restrictions à imposer, afin de conserver de bonnes propriétés pour la fonction de vraisemblance. A la fin de ce panorama, nous devons être capables de formuler les contraintes à imposer sur l'espace des paramètres de chaque classe de fonctions de la famille exponentielle, ainsi que sur les fonctions auxiliaires utilisées dans la densité générale (4.23). Nos résutats sont exprimés pour une réponse Y unidimensionnelle dans un but de simplification d'écriture, mais sont généralisables au cas où $Y \in \mathbb{R}^d$.

Mélange de régressions linéaires

La loi Normale fait partie de la famille exponentielle, et constitue à ce titre un choix possible de modélisation de l'erreur dans un modèle linéaire généralisé. Le support de la loi Normale est la droite des réels, et ses paramètres μ et σ^2 appartiennent respectivement à l'ensemble des réels et aux réels strictement positifs. Ainsi, une variable aléatoire Y de loi normale $\mathcal{N}(\mu, \sigma^2)$, dont la densité vaut

$$f_{\mathcal{N}}(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right),$$

admet une représentation sous forme de famille exponentielle. En effet, en considérant que la densité de Y est également donnée par

$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\Phi)} + c(y, \Phi)\right\},$$

avec
$$\begin{cases} \theta = \mu, \text{ d'où l'ensemble de définition pour } \theta : \theta \in \mathbb{R}, \\ b(\theta) = \dfrac{\theta^2}{2} = \dfrac{\mu^2}{2}, \text{ d'où } b(\theta) \in \mathbb{R}^+, \\ \phi = a(\phi) = \sigma^2, \text{ d'où l'ensemble de définition pour } \phi : \phi \in \mathbb{R}^{+*}, \\ c(y; \phi) = -\dfrac{1}{2}\left(\dfrac{y^2}{\sigma^2} + \ln 2\pi\sigma^2\right), \text{ d'où } c(y; \phi) \in \mathbb{R}; \end{cases}$$

nous retrouvons après quelques calculs la densité originelle d'une gaussienne.

Lorsque nous travaillons avec des mélanges de régressions linéaires, la densité pour une observation peut s'écrire sous la forme suivante (en reprenant nos notations) après quelques calculs : $\forall \psi_G \in \Psi_G$,

$$\begin{aligned} f(y_j; \psi_G) = L(\psi_G; y_j) &= \sum_{i=1}^{G} \pi_i \, f_{\mathcal{N}}(y_j; \mu_i, \sigma_i^2) \\ &= \sum_{i=1}^{G} \pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{(y_j - \mu_i)^2}{\sigma_i^2}\right) \\ &= \sum_{i=1}^{G} \pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{(y_j - X_j\beta_i)^2}{\sigma_i^2}\right), \end{aligned}$$

où $\beta_i = (\beta_{i0}, \beta_{i1}, ..., \beta_{ip})^T$ et $X_j = (1, X_{j1}, X_{j2}, ..., X_{jp})$.

En considérant un lien *identité* et une erreur gaussienne dans un modèle mélange de GLMs, nous retombons sur le cas des mélanges gaussiens auxquels nous ajoutons une dépendance en fonction de variables explicatives (par l'équation de régression). C'est tout l'intérêt de la présentation des résultats sur mélanges gaussiens qui a été faite au préalable, et qui va nous servir d'inspiration pour les résultats à venir. Cette densité de mélange nous permet d'exprimer la vraisemblance classifiante conditionnelle pour les mélanges de régressions linéaires pour une observation y_j. Rappelons que

$$\ln L_{cc}(\psi_G; y_j) \quad = \quad \ln L(\psi_G; y_j) + \sum_{i=1}^{G} \frac{\pi_i \, f_{\mathcal{N}}(y_j; \mu_i, \sigma_i^2)}{\sum_{k=1}^{G} \pi_k \, f_{\mathcal{N}}(y_j; \mu_k, \sigma_k^2)} \ln \left(\frac{\pi_i \, f_{\mathcal{N}}(y_j; \mu_i, \sigma_i^2)}{\sum_{k=1}^{G} \pi_k \, f_{\mathcal{N}}(y_j; \mu_k, \sigma_k^2)} \right).$$

D'où en développant

$$\ln L_{cc}(\psi_G; y_j) = \ln \left(\sum_{i=1}^{G} \pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2} \frac{(y_j - X_j\beta_i)^2}{\sigma_i^2} \right) \right) +$$

$$\sum_{i=1}^{G} \frac{\pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2} \frac{(y_j - X_j\beta_i)^2}{\sigma_i^2} \right)}{\sum_{k=1}^{G} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2} \frac{(y_j - X_j\beta_k)^2}{\sigma_k^2} \right)} \ln \left(\frac{\pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2} \frac{(y_j - X_j\beta_i)^2}{\sigma_i^2} \right)}{\sum_{k=1}^{G} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2} \frac{(y_j - X_j\beta_k)^2}{\sigma_k^2} \right)} \right).$$

Clairement, les mêmes contraintes que celles sur les mélanges gaussiens doivent être imposées : ces contraintes sur μ_i et σ_i^2 sont répercutables sur β_i et σ_i^2, et donc aussi sur θ_i et ϕ_i. L'annexe D.1.2 détaille les calculs des limites de L_{cc}. Nous obtenons finalement que :

- la variance σ_i^2 doit rester bornée, donc ϕ_i doit également être bornée ;
- σ_i^2 ne doit pas tendre vers 0, ce qui induit la même contrainte pour ϕ_i ;
- les coefficients de régression β_i des composantes doivent rester bornés ($|\beta_i| \neq \infty$). Sachant que $\theta_i = \mu_i = X\beta_i$, nous en déduisons que θ_i doit aussi rester borné.

Pour résumer, il faut se placer dans un espace compact bien choisi pour assurer la bornitude de la log-vraisemblance classifiante conditionnelle ainsi que de sa dérivée. Si les contraintes sur les paramètres θ_i et ϕ_i se révèlent être relativement similaires après l'étude de toutes les classes de GLMs, il sera ainsi possible de formuler des résultats généraux de convergence de l'estimateur $ML_{cc}E$ pour cette grande famille.

Mélange de régressions de Poisson

Un autre choix de modélisation de l'erreur pourrait être une loi de Poisson lorsque nous nous intéressons à des données de comptage. La loi de Poisson est à valeurs dans l'ensemble des entiers naturels, et son paramètre μ appartient à l'ensemble des réels strictement positifs. Le tableau 4.2 donne la correspondance entre le paramètre μ et les paramètres de tendance et de dispersion de la famille exponentielle.

La densité d'une observation avec des mélanges de régressions de Poisson s'exprime sous la forme suivante après quelques calculs : $\forall \psi_G \in \Psi_G$,

$$\begin{aligned} f(y_j;\psi_G) = L(\psi_G;y_j) &= \sum_{i=1}^{G} \pi_i \, f_{\mathcal{P}}(y_j;\mu_i) \\ &= \sum_{i=1}^{G} \pi_i \, \exp(-\mu_i) \, \frac{\mu_i^{y_j}}{y_j!} \\ &= \sum_{i=1}^{G} \pi_i \, \exp\left(-\exp(X_j\beta_i)\right) \, \frac{[\exp(X_j\beta_i)]^{y_j}}{y_j!}, \end{aligned}$$

où $\beta_i = (\beta_{i0}, \beta_{i1}, ..., \beta_{ip})^T$ et $X_j = (1, X_{j1}, X_{j2}, ..., X_{jp})$.
Cette densité de mélange nous permet d'exprimer la vraisemblance classifiante conditionnelle qui en découle (toujours pour une observation y_j) :

$$\begin{aligned} \ln L_{cc}(\psi_G;y_j) &= \ln L(\psi_G;y_j) + \sum_{i=1}^{G} \frac{\pi_i \, f_{\mathcal{P}}(y_j;\mu_i)}{\sum_{k=1}^{G} \pi_k \, f_{\mathcal{P}}(y_j;\mu_k)} \ln\left(\frac{\pi_i \, f_{\mathcal{P}}(y_j;\mu_i)}{\sum_{k=1}^{G} \pi_k \, f_{\mathcal{P}}(y_j;\mu_k)}\right) \\ &= \ln\left(\sum_{i=1}^{G} \pi_i \, f_{\mathcal{P}}(y_j;\mu_i)\right) + \sum_{i=1}^{G} \frac{\pi_i \, f_{\mathcal{P}}(y_j;\mu_i)}{\sum_{k=1}^{G} \pi_k \, f_{\mathcal{P}}(y_j;\mu_k)} \ln\left(\frac{\pi_i \, f_{\mathcal{P}}(y_j;\mu_i)}{\sum_{k=1}^{G} \pi_k \, f_{\mathcal{P}}(y_j;\mu_k)}\right) \\ &= \ln\left(\sum_{i=1}^{G} \pi_i e^{-e^{X_j\beta_i}} \frac{\left[e^{X_j\beta_i}\right]^{y_j}}{y_j!}\right) + \sum_{i=1}^{G} \frac{\pi_i e^{-e^{X_j\beta_i}} \frac{\left[e^{X_j\beta_i}\right]^{y_j}}{y_j!}}{\sum_{k=1}^{G} \pi_k e^{-e^{X_j\beta_k}} \frac{\left[e^{X_j\beta_k}\right]^{y_j}}{y_j!}} \ln\left(\frac{\pi_i e^{-e^{X_j\beta_i}} \frac{\left[e^{X_j\beta_i}\right]^{y_j}}{y_j!}}{\sum_{k=1}^{G} \pi_k e^{-e^{X_j\beta_k}} \frac{\left[e^{X_j\beta_k}\right]^{y_j}}{y_j!}}\right) \end{aligned}$$

Faisons tendre le paramètre μ_i vers les frontières de son domaine de définition et étudions les limites de la vraisemblance L_{cc}. Nous obtenons comme contrainte après calculs (cf annexe D.2.2) que les coefficients de régression β_i des composantes doivent rester bornés ($|\beta_i| \neq \infty$), ce qui signifie la même contrainte pour les paramètres θ_i.

Mélange de régressions logistiques

Dans nos applications, c'est le mélange que nous utilisons : en effet, le nombre de rachats à l'échelle d'un portefeuille suit une loi binomiale $\mathcal{B}(n,p)$. Cette loi fait partie de la famille exponentielle, a un support défini par l'ensemble des entiers naturels, et ses paramètres n et p représentent respectivement le nombre d'assurés en portefeuille et la probabilité individuelle de rachat. n est un entier naturel, et $p \in [0,1]$. La variable aléatoire Y représentant la proportion de rachat dans la population prend donc des valeurs dans $[0, 1/n, ..., 1]$ et admet pour densité

$$P(Y = y) = f(y;n,p) = C_n^{ny} p^{ny} (1-p)^{n-ny}.$$

Nous pouvons encore une fois transformer l'écriture de cette densité sous la forme de celle de la famille exponentielle, en fixant le paramétrage donné par le tableau 4.2.

En travaillant sur une observation, la décision de rachat est une loi de Bernoulli de paramètre p. La densité des mélanges de régressions logistiques a été donnée à plusieurs reprises

déjà dans ce mémoire, et vaut : $\forall \psi_G \in \Psi_G$,

$$\begin{aligned} f(y_j;\psi_G) = L(\psi_G;y_j) &= \sum_{i=1}^{G} \pi_i \, f_{\mathcal{B}}(y_j;p_i) \\ &= \sum_{i=1}^{G} \pi_i \, p_i \\ &= \sum_{i=1}^{G} \pi_i \, \frac{\exp(X_j\beta_i)}{1+\exp(X_j\beta_i)}, \end{aligned}$$

où $\beta_i = (\beta_{i0}, \beta_{i1}, ..., \beta_{ip})^T$ et $X_j = (1, X_{j1}, X_{j2}, ..., X_{jp})$.
Nous en déduisons la log-vraisemblance classifiante conditionnelle d'une observation y_j pour les mélanges de régressions logistiques :

$$\begin{aligned} \ln L_{cc}(\psi_G;y_j) &= \ln L(\psi_G;y_j) + \sum_{i=1}^{G} \frac{\pi_i \, f_{\mathcal{B}}(y_j;p_i)}{\sum_{k=1}^{G} \pi_k \, f_{\mathcal{B}}(y_j;p_k)} \ln\left(\frac{\pi_i \, f_{\mathcal{B}}(y_j;p_i)}{\sum_{k=1}^{G} \pi_k \, f_{\mathcal{B}}(y_j;p_k)}\right) \\ &= \ln\left(\sum_{i=1}^{G} \pi_i \, f_{\mathcal{B}}(y_j;p_i)\right) + \sum_{i=1}^{G} \frac{\pi_i \, f_{\mathcal{B}}(y_j;p_i)}{\sum_{k=1}^{G} \pi_k \, f_{\mathcal{B}}(y_j;p_k)} \ln\left(\frac{\pi_i \, f_{\mathcal{B}}(y_j;p_i)}{\sum_{k=1}^{G} \pi_k \, f_{\mathcal{B}}(y_j;p_k)}\right). \end{aligned}$$

D'où en développant,

$$\ln L_{cc}(\psi_G;y_j) = \ln\left(\sum_{i=1}^{G} \pi_i \frac{e^{X_j\beta_i}}{1+e^{X_j\beta_i}}\right) + \sum_{i=1}^{G} \frac{\pi_i \, \dfrac{e^{X_j\beta_i}}{1+e^{X_j\beta_i}}}{\sum_{k=1}^{G} \pi_k \, \dfrac{e^{X_j\beta_k}}{1+e^{X_j\beta_k}}} \ln\left(\frac{\pi_i \, \dfrac{e^{X_j\beta_i}}{1+e^{X_j\beta_i}}}{\sum_{k=1}^{G} \pi_k \, \dfrac{e^{X_j\beta_k}}{1+e^{X_j\beta_k}}}\right).$$

Toujours par l'étude des limites de cette vraisemblance aux frontières de l'espace des paramètres, les contraintes à imposer deviennent flagrantes (annexe D.3.2). En fait, l'unique cas critique pour la bornitude de la log-vraisemblance classifiante conditionnelle correspond à $\beta_i \to -\infty$. De par la relation bijective qu'il existe entre les paramètres θ_i et ϕ_i et les paramètres originels de cette distribution, nous étendons ces contraintes aux contraintes à imposer sur l'espace des paramètres de la famille exponentielle. Ainsi, il suffit d'imposer que θ_i reste borné.

Mélange de régressions Gamma

Parfois, l'erreur peut être de loi Gamma lorsque nous désirons modéliser la charge des sinistres. Cette loi continue est à valeur dans l'ensemble des réels positifs, et ses paramètres μ et ν appartiennent tous deux à l'ensemble des réels strictement positifs. Pour Y une variable aléatoire de loi Gamma $\Gamma(\mu,\nu)$, la densité est donnée par

$$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu}{\mu}\, y\right).$$

Après quelques calculs, la densité d'un mélange de régressions Gamma pour une observa-

tion y_j s'exprime comme suit : $\forall \psi_G \in \Psi_G$,

$$\begin{aligned} f(y_j;\psi_G) = L(\psi_G;y_j) &= \sum_{i=1}^{G} \pi_i \, f_\Gamma(y_j;\mu_i,\nu_i) \\ &= \sum_{i=1}^{G} \pi_i \, \frac{1}{\Gamma(\nu_i)} \left(\frac{\nu_i}{\mu_i}\right)^{\nu_i} y_j^{\nu_i-1} \exp\left(-\frac{\nu_i}{\mu_i}\, y_j\right) \\ &= \sum_{i=1}^{G} \pi_i \, \frac{1}{\Gamma(\nu_i)} \left(\nu_i X_j \beta_i\right)^{\nu_i} y_j^{\nu_i-1} \exp\left(-\nu_i X_j \beta_i \, y_j\right), \end{aligned}$$

où $\beta_i = (\beta_{i0}, \beta_{i1}, ..., \beta_{ip})^T$ et $X_j = (1, X_{j1}, X_{j2}, ..., X_{jp})$.
Cette densité de mélange nous permet d'exprimer la vraisemblance classifiante conditionnelle qui en découle (toujours pour une observation y_j) :

$$\begin{aligned} \ln L_{cc}(\psi_G;y_j) &= \ln L(\psi_G;y_j) + \sum_{i=1}^{G} \frac{\pi_i \, f_\Gamma(y_j;\mu_i)}{\sum_{k=1}^{G} \pi_k \, f_\Gamma(y_j;\mu_k)} \ln\left(\frac{\pi_i \, f_\Gamma(y_j;\mu_i)}{\sum_{k=1}^{G} \pi_k \, f_\Gamma(y_j;\mu_k)}\right) \\ &= \ln\left(\sum_{i=1}^{G} \pi_i \, f_\Gamma(y_j;\mu_i)\right) + \sum_{i=1}^{G} \frac{\pi_i \, f_\Gamma(y_j;\mu_i)}{\sum_{k=1}^{G} \pi_k \, f_\Gamma(y_j;\mu_k)} \ln\left(\frac{\pi_i \, f_\Gamma(y_j;\mu_i)}{\sum_{k=1}^{G} \pi_k \, f_\Gamma(y_j;\mu_k)}\right). \end{aligned}$$

D'où en développant,

$$\ln L_{cc}(\psi_G;y_j) = \ln\left(\sum_{i=1}^{G} \frac{\pi_i}{\Gamma(\nu_i)} \left(\nu_i X_j \beta_i\right)^{\nu_i} y_j^{\nu_i-1} \, e^{-\nu_i X_j \beta_i \, y_j}\right) +$$

$$\sum_{i=1}^{G} \frac{\frac{\pi_i}{\Gamma(\nu_i)} \left(\nu_i X_j \beta_i\right)^{\nu_i} y_j^{\nu_i-1} \, e^{-\nu_i X_j \beta_i \, y_j}}{\sum_{k=1}^{G} \frac{\pi_k}{\Gamma(\nu_k)} \left(\nu_k X_j \beta_k\right)^{\nu_k} y_j^{\nu_k-1} \, e^{-\nu_k X_j \beta_k \, y_j}} \ln\left(\frac{\frac{\pi_i}{\Gamma(\nu_i)} \left(\nu_i X_j \beta_i\right)^{\nu_i} y_j^{\nu_i-1} \, e^{-\nu_i X_j \beta_i \, y_j}}{\sum_{k=1}^{G} \frac{\pi_k}{\Gamma(\nu_k)} \left(\nu_k X_j \beta_k\right)^{\nu_k} y_j^{\nu_k-1} \, e^{-\nu_k X_j \beta_k \, y_j}}\right).$$

Nous devons imposer certaines contraintes pour conserver les propriétés de bornitude de la vraisemblance classifiante conditionnelle : notamment, il ne faut pas que (cf annexe D.4.2)

- $\nu_i \to 0$ ou $\nu_i \to +\infty$;
- $\beta_i \to -\infty$ ou $\beta_i \to +\infty$.

En effet, quand les paramètres β_i et ν_i vers les frontières de leur domaine de définition, l'étude des limites de la vraisemblance L_{cc} montre que cette dernière explose. Traduisons maintenant les contraintes équivalentes sur l'espace des paramètres de la famille exponentielle (cf tableau 4.2) : il faut éviter que $\phi_i \to +\infty$ ou que $\phi_i \to 0$, et s'assurer que θ_i soit borné.

Mélange d'Inverses Gaussiennes

L'Inverse Gaussienne fait partie de la famille exponentielle, et constitue à ce titre un choix possible de modélisation de l'erreur dans un modèle linéaire généralisé. Elle est utilisée dans la modélisation de la sévérité des sinistres en Assurance, et a comme support l'ensemble des réels positifs. Ses deux paramètres μ et σ^2 appartiennent à l'ensemble des réels strictement positifs. Ainsi, une variable aléatoire Y de loi Inverse Gaussienne $\mathcal{IN}(\mu, \sigma^2)$, dont la densité vaut

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left(-\frac{1}{2}\, \frac{(y-\mu)^2}{\mu^2 \sigma^2 y}\right),$$

admet une représentation sous forme de famille exponentielle.

Dans le contexte des mélanges d'inverses gaussiennes, la densité pour une observation y_j peut s'écrire : $\forall \psi_G \in \Psi_G$,

$$\begin{aligned} f(y_j;\psi_G) = L(\psi_G;y_j) &= \sum_{i=1}^{G} \pi_i \, f_{\mathcal{IN}}(y_j;\mu_i,\sigma_i^2) \\ &= \sum_{i=1}^{G} \pi_i \, \frac{1}{\sqrt{2\pi\sigma_i^2 y_j^3}} \exp\left(-\frac{1}{2}\frac{(y_j-\mu_i)^2}{\mu_i^2\sigma_i^2 y_j}\right) \\ &= \sum_{i=1}^{G} \pi_i \, \frac{1}{\sqrt{2\pi\sigma_i^2 y_j^3}} \exp\left(-\frac{1}{2}\frac{\left(y_j-\sqrt{\frac{1}{X_j\beta_i}}\right)^2}{\frac{1}{X_j\beta_i}\sigma_i^2 y_j}\right), \end{aligned}$$

où $\beta_i = (\beta_{i0}, \beta_{i1}, ..., \beta_{ip})^T$ et $X_j = (1, X_{j1}, X_{j2}, ..., X_{jp})$.
Ainsi, pour une observation :

$$\begin{aligned} \ln L_{cc}(\psi_G;y_j) &= \ln L(\psi_G;y_j) + \sum_{i=1}^{G} \frac{\pi_i \, f_{\mathcal{IN}}(y_j;\mu_i,\sigma_i^2)}{\sum_{k=1}^{G}\pi_k \, f_{\mathcal{IN}}(y_j;\mu_k,\sigma_k^2)} \ln\left(\frac{\pi_i \, f_{\mathcal{IN}}(y_j;\mu_i,\sigma_i^2)}{\sum_{k=1}^{G}\pi_k \, f_{\mathcal{IN}}(y_j;\mu_k,\sigma_k^2)}\right) \\ &= \ln\left(\sum_{i=1}^{G}\pi_i \, f_{\mathcal{IN}}(y_j;\mu_i,\sigma_i^2)\right) + \sum_{i=1}^{G} \frac{\pi_i \, f_{\mathcal{IN}}(y_j;\mu_i,\sigma_i^2)}{\sum_{k=1}^{G}\pi_k \, f_{\mathcal{IN}}(y_j;\mu_k,\sigma_k^2)} \ln\left(\frac{\pi_i \, f_{\mathcal{IN}}(y_j;\mu_i,\sigma_i^2)}{\sum_{k=1}^{G}\pi_k \, f_{\mathcal{IN}}(y_j;\mu_k,\sigma_k^2)}\right). \end{aligned}$$

D'où en développant,

$$\begin{aligned} \ln L_{cc}(\psi_G;y_j) = \ln\left(\sum_{i=1}^{G}\frac{\pi_i}{\sqrt{2\pi\sigma_i^2 y_j^3}}\exp\left(-\frac{1}{2}\frac{\left(y_j-\sqrt{\frac{1}{X_j\beta_i}}\right)^2}{\frac{1}{X_j\beta_i}\sigma_i^2 y_j}\right)\right) + \\ \sum_{i=1}^{G}\frac{\frac{\pi_i}{\sqrt{2\pi\sigma_i^2 y_j^3}}\exp\left(-\frac{1}{2}\frac{\left(y_j-\sqrt{\frac{1}{X_j\beta_i}}\right)^2}{\frac{1}{X_j\beta_i}\sigma_i^2 y_j}\right)}{\sum_{k=1}^{G}\frac{\pi_k}{\sqrt{2\pi\sigma_k^2 y_j^3}}\exp\left(-\frac{1}{2}\frac{\left(y_j-\sqrt{\frac{1}{X_j\beta_k}}\right)^2}{\frac{1}{X_j\beta_k}\sigma_k^2 y_j}\right)} \ln\left(\frac{\frac{\pi_i}{\sqrt{2\pi\sigma_i^2 y_j^3}}\exp\left(-\frac{1}{2}\frac{\left(y_j-\sqrt{\frac{1}{X_j\beta_i}}\right)^2}{\frac{1}{X_j\beta_i}\sigma_i^2 y_j}\right)}{\sum_{k=1}^{G}\frac{\pi_k}{\sqrt{2\pi\sigma_k^2 y_j^3}}\exp\left(-\frac{1}{2}\frac{\left(y_j-\sqrt{\frac{1}{X_j\beta_k}}\right)^2}{\frac{1}{X_j\beta_k}\sigma_k^2 y_j}\right)}\right). \end{aligned}$$

En annexe D.5.2, nous détaillons les limites obtenues pour la log-vraisemblance classifiante conditionnelle lorsque les paramètres de la distribution tendent vers les frontières de leur espace de définition. Nous devons finalement éviter :

- $\sigma_i^2 \to 0$ ou $\sigma_i^2 \to +\infty$;
- $\beta_i \to -\infty$ ou $\beta_i \to +\infty$.

Ces limites définissent les contraintes à satisfaire pour que l'estimateur $ML_{cc}E$ existe d'une part, et soit convergent d'autre part. Leur équivalent sur l'espace des paramètres de θ et ϕ est immédiat grâce à l'utilisation des relations bijectives les liant (cf tableau 4.2). Ainsi nous devons éviter $\phi_i \to 0$ ou $\phi_i \to +\infty$, de même θ_i doit rester borné.

Résumé des contraintes à imposer sur l'espace des paramètres

Le tableau 4.2 reprend l'ensemble des résultats de l'étude de chaque classe de la famille des GLMs. Il permet de récapituler le support contraint de chaque paramètre θ et ϕ, ce qui va nous être très utile pour la formulation des hypothèses de convergence de l'estimateur $ML_{cc}E$ et du critère de sélection ICL. En effet, nous voyons bien que l'ensemble des membres de la famille GLM se comporte de manière identique d'un point de vue des contraintes à imposer sur les paramètres de la famille exponentielle. Dès que le paramètre de tendance et/ou la dispersion sont bornés, la log-vraisemblance classifiante conditionnelle reste finie (à condition que la dispersion ne tende pas vers 0), de même que sa dérivée (dont particulièrement la dérivée de l'entropie !). Ces résultats montrent tout l'intérêt de se placer dans des ensembles compacts pour l'espace des paramètres.

4.3.3 Propriétés de convergence avec les mélanges de GLMs

Nous avons longuement discuté des hypothèses qui interviennent dans les théorèmes et lemmes précédents. Finalement, il apparait que le point essentiel à vérifier est la bornitude de la vraisemblance classifiante conditionnelle, ainsi que de sa dérivée. L'étude plus précise de chaque famille des GLM a montré qu'il n'était pas nécessaire d'imposer des contraintes sur les fonctions $a()$, $b()$ et $c()$ de la famille exponentielle. Nous savons donc qu'en imposant des bornes sur les paramètres de localisation et de dispersion des membres des GLM, nous vérifions l'ensemble des hypothèses requises (notamment le fait que cette famille soit $\mathbb{P}$-Glivenko-Cantelli). Evidemment, certaines autres hypothèses sont indispensables dans la pratique (support borné pour les densités, matrice d'information inversible, compacité/convexité de l'espace des paramètres, ...), mais elles ne semblent pas poser de problème concrètement.

Loi de Y :	Normale $\mathcal{N}(\mu, \sigma^2)$	Binomiale $B(n, \mu)$	Poisson $\mathcal{P}(\mu)$	Gamma $\mathcal{G}(\mu, \nu)$	Inverse Gaussienne $\mathcal{IN}(\mu, \sigma^2)$
Supports	$y \in \mathbb{R}$ $\mu \in \mathbb{R}$ $\sigma^2 \in \mathbb{R}^{+*}$	$y \in [\![0, n]\!]$ $n \in \mathbb{N}^*$ $\mu \in [0, 1]$	$y \in \mathbb{N}$ $\mu \in \mathbb{R}^+$	$y \in \mathbb{R}^+$ $\mu \in \mathbb{R}^{+*}$ $\nu \in \mathbb{R}^{+*}$	$y \in \mathbb{R}^+$ $\mu \in \mathbb{R}^{+*}$ $\sigma^2 \in \mathbb{R}^{+*}$
Tendance $\theta(\mu)$ Support de θ	μ $\theta \in \mathbb{R}$	$\ln[\mu/(1-\mu)]$ $\theta \in \mathbb{R}$	$\ln \mu$ $\theta \in \mathbb{R}$	$-\mu^{-1}$ $\theta \in \mathbb{R}^{-*}$	$-(2\mu^2)^{-1}$ $\theta \in \mathbb{R}^{-*}$
Dispersion ϕ Support de ϕ	σ^2 $\phi \in \mathbb{R}^{+*}$	1	1	ν^{-1} $\phi \in \mathbb{R}^{+*}$	σ^2 $\phi \in \mathbb{R}^{+*}$
Fonction $b(\theta)$	$\theta^2/2$	$\ln(1+e^\theta)$	e^θ	$-\ln(-\theta)$	$-(-2\theta)^{1/2}$
Fonction $c(y, \Phi)$	$-\frac{1}{2}\left(\frac{y^2}{\Phi} + \ln(2\pi\phi)\right)$	$\ln(C_n^{ny})$	$-\ln(y!)$		$-\frac{1}{2}\left\{\ln(2\pi\Phi y^3) + \frac{1}{\Phi y}\right\}$
$\mu(\theta) = \mathbb{E}[Y;\theta]$	θ	$e^\theta/(1+e^\theta)$	e^θ	$-1/\theta$	$(-2\theta)^{-1/2}$
Contraintes	$\lvert\theta\rvert < +\infty$ $\phi < +\infty$ $\phi \nrightarrow 0$	$\lvert\theta\rvert < +\infty$	$\lvert\theta\rvert < +\infty$	$\lvert\theta\rvert < +\infty$ $\phi < +\infty$ $\phi \nrightarrow 0$	$\lvert\theta\rvert < +\infty$ $\phi < +\infty$ $\phi \nrightarrow 0$

TABLE 4.2 – Résumé des contraintes sur l'espace des paramètres des GLMs.

Epaisseur limite des queues de distribution

Nous avions fait remarquer lors de la discussion sur l'hypothèse (H1-A) que la queue de distribution de f^0 ne doit pas s'avérer trop épaisse. En effet, l'intégrale paramétrique en ψ_G donnée par l'espérance (sous f^0) de la log-vraisemblance classifiante conditionnelle doit être continue. Le théorème de convergence dominée permet de garantir une telle continuité sous certaines conditions, en l'occurrence que la fonction sous l'intégrale soit bornée par une fonction indépendante de ψ_G qui soit elle-même intégrable. En des termes plus mathématiques, nous voulons étudier $\mathbb{E}_{f^0}\left[\ln L_{cc}(\psi_G, y)\right]$ donnée par $\int_{\mathcal{Y}} \ln L_{cc}(\psi_G, y) f^0(y) dy$.
L'objectif est donc de borner la fonction $g(\psi_G, y) = \ln L_{cc}(\psi_G, y) f^0(y)$ par une fonction h telle que $h = h(y)$ et h intégrable dans $\mathcal{Y}$. Pour visualiser intuitivement ce résultat, nous considérons le supremum en ψ_G de la fonction $\ln L_{cc}$ (qui ne dépend donc plus de ψ_G). Ce supremum se comporte pour les grandes valeurs de y comme la vraisemblance $\ln L_{cc}$ tant que ψ_G est compact. De plus, le terme d'entropie dans la vraisemblance L_{cc} ne pose aucun problème d'intégration dans le passage à l'espérance car nous avons toujours (grâce à la limite $\lim_{x\to 0} x \ln x = 0$) :

$$\forall y \in \mathbb{R}^d, \quad 0 \leq Ent(\psi_G; y) \leq \ln G.$$

Nous allons donc considérer la vraisemblance des données observées dans ce raisonnement, sans se soucier du comportement du terme entropique. A l'aide de ces deux remarques il devient bien plus simple de trouver la fonction h, bien que les calculs soient longs et fastidieux. Nous en donnons pour ainsi dire directement certains résultats, qui n'ont pas vocation à fournir la fonction h la moins contraignante possible mais plutôt à donner une idée de la forme de la queue de distribution limite.

En pratique avec des mélanges de régressions linéaires, il suffit par exemple que $f^0(y) = o\left(\frac{1}{y^3}\right)$, ce qui est largement raisonnable en réalité si l'on pense à la densité gaussienne comme loi qui sous-tend les données observées. Clairement cette loi a un comportement asymptotique qui tend plus vite vers 0 en l'infini puisqu'il est en exponentielle. Pour le cas des régressions de Poisson, le terme en factorielle pose problème : en effet la factorielle l'emporte sur l'exponentielle en l'infini, ce qui suggère de prendre f^0 de la forme $f^0(y) = o\left(\frac{1}{y!}\right)$. Cette densité est en revanche beaucoup moins anodine et ne reflète en général pas la réalité, car la décroissance de la queue de la densité se fait ici à une vitesse supérieure à une décroissance exponentielle ! Il en est de même lorsqu'on nous étudions les mélanges de régressions logistiques, à cause du coefficient binomial qui comporte également un terme en factorielle ($y!$). A contrario, les mélanges de régression Gamma, ainsi que les mélanges d'inverses gaussiennes, ne semblent pas poser de souci quant à la queue de la densité inconnue f^0. Respectivement, nous devrions admettre que $f^0(y) = o(\frac{1}{y})$, et que $f^0(y) = o(\frac{1}{y^2})$. Ces considérations sont tout à fait plausibles au regard des densités usuelles d'observations que nous manipulons.

Convergence du $ML_{cc}E$ et consistance de ICL_c, mélange de GLMs

Nous sommes maintenant en mesure d'expliciter le théorème de convergence de l'estimateur par maximum de vraisemblance classifiante conditionnelle (qui est un M-estimateur) dans le cadre des mélanges de GLMs. Ces hypothèses reposent effectivement sur des propriétés de la fonction de vraisemblance classifiante conditionnelle et de sa dérivée, qui sont directement vérifiables par calcul. Le nouveau théorème en question est le suivant :

Proposition 7. *(Convergence forte de l'estimateur $ML_{cc}E$ pour mélange de GLMs).*
Soit M_G un mélange de GLMs à G composantes, tel que défini en section 4.3.2. L'espace des paramètres Ψ_G, de dimension K_G, satisfait $\Psi_G \subset \mathbb{R}^{K_G}$.
Si l'ensemble des paramètres Ψ_G est compact et que nous imposons les restrictions sur les paramètres donnés par le tableau 4.2, alors l'estimateur $\hat{\psi}_G^{ML_{cc}E}$ est fortement convergent vers le meilleur ensemble de paramètre Ψ_G^b sous-jacent à la distribution des données.

Démonstration. La preuve est immédiate quand on sait que les contraintes imposées sur la famille GLM permettent de satisfaire les hypothèses (H2-B) et (H3-B). L'ensemble des paramètres étant compact, le tour est joué. □

Proposition 8. *(Consistance faible de ICL_c pour mélanges de GLMs, cas compact).*
Soit $\{M_g\}_{1\leq g\leq m}$ une collection de mélanges de GLMs de paramètres $\{\psi_g\}_{1\leq g\leq m} \in \{\Psi_g\}_{1\leq g\leq m}$ et de dimension $\{K_g\}_{1\leq g\leq m}$, avec $\Psi_g \subset \mathbb{R}^{K_g}$. Ces modèles sont classés dans un ordre croissant de complexité, avec $K_1 \leq K_2 \leq ... \leq K_m$.
Le critère de sélection ICL_c est faiblement consistant : le nombre de composantes du mélange de GLMs sélectionné via ce critère converge en probabilité vers le nombre théorique de composantes de la densité mélange sous-jacente.

Démonstration. La preuve est immédiate quand on sait que les contraintes imposées sur la famille GLM (tableau 4.2) permettent de satisfaire les différentes hypothèses du théorème 9. L'expression de la pénalité du critère ICL_c ainsi que le fait que l'estimateur $ML_{cc}E$ soit fortement convergent vers le paramètre théorique complètent l'argumentation de cette preuve. □

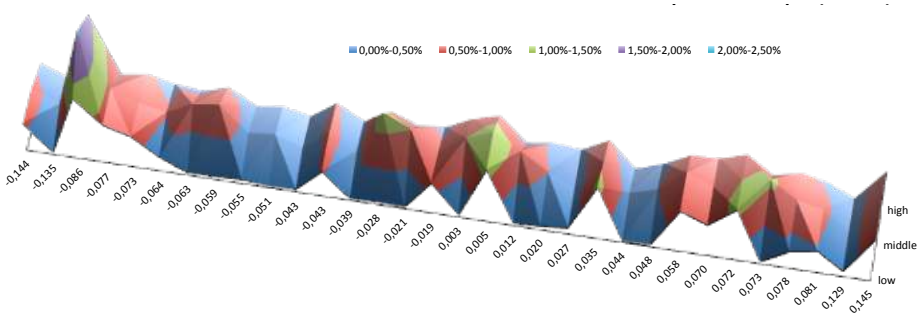
4.4 Applications

Nous proposons dans cette partie de revenir sur les applications développées tout au long de la thèse, et de voir en quoi ce dernier chapitre peut être utile à la fois d'un point de vue théorique mais aussi d'un point de vue opérationnel.

4.4.1 Quelques remarques importantes

Il est intéressant de comparer les résultats obtenus *infra* avec ceux fournis dans le chapitre 3 et les annexes. En effet, l'objectif était de diminuer la dimension des modèles en supprimant certaines des composantes qui se ressemblaient fortement. Notre objectif initial étant d'être capable de distinguer clairement différents types de comportement de rachat, le critère ICL de sélection de modèle est ici employé. Nous allons voir dans ces nouvelles applications que celui-ci répond bel et bien à nos attentes : dans la grande majorité des cas, il permet de s'affranchir de certaines composantes présentes dans un modèle sélectionné par le BIC, tout en améliorant la qualité d'estimation des coefficients de régression. Les composantes restantes ont des caractéristiques nettement mieux différentiables, ce qui vient renforcer l'idée de "clusters comportementaux" facilement identifiables. Il aurait été intéressant de tracer la densité du mélange de GLMs afin de clairement la visualiser (et donc de détecter d'éventuels "overlap" entre densités des composantes), mais ceci est rendu relativement complexe du fait de la multidimensionnalité des données. Les densités de chaque composante du mélange sont

FIGURE 4.2 – Taux de rachat trimestriel des produits Mixtes, en fonction de la catégorie de prime de risque et du taux des obligations 10 ans en Espagne.



conditionnelles à l'observation des variables explicatives, et nous sommes limités aux trois dimensions pour visualiser une surface malgré certains logiciels utiles tels que *GGobi* (il y a toujours plus de deux variables explicatives...). Nous pourrions dès lors fixer certaines des variables explicatives et ne faire varier que les deux qui différencient les groupes selon nous, mais là encore les interprétations ne sont pas forcément évidentes (voir exemple de la figure 4.2). Cette limitation due au fait de travailler avec ce type de modèles nous empêche par exemple de discuter de certaines restrictions à imposer sur l'espace des paramètres qui pourraient jouer sur la forme des clusters désirés.

Un autre point à développer ultérieurement serait de regarder l'évolution (dans le temps calendaire) de la densité du mélange conditionnellement aux valeurs évolutives de certains indices économiques et financiers, afin de visualiser sa déformation engendrée par le changement de contexte économique. Une application intéressante serait alors de comparer les densités de chaque composante conditionnellement à la valeur d'un indice économique, et d'en déduire les différents comportements. D'un point de vue calculatoire, l'estimateur $ML_{cc}E$ est plus difficile à calculer : des solutions algorithmiques sont proposées dans Baudry (2009), faisant intervenir une adaptation de l'algorithme EM pour prendre en compte dans la maximisation le terme d'entropie. En pratique, nous utilisons l'estimateur MLE avec le critère ICL car il ressort de nos discussions avec les chercheurs spécialistes de la question que dans la plupart des cas, les résultats numériques entre MLE et $ML_{cc}E$ sont assez proches (lorsque le modèle n'est pas trop surdimensionné). Nous comparons donc les résultats donnés par le critère de sélection ICL avec ceux fournis via le BIC pour l'estimateur MLE, et donnons un premier aperçu de la différence due à un simple changement de critère prenant en compte l'entropie.

4.4.2 Mise en oeuvre sur nos familles de produits

Le fil général de la présentation des résultats sera le suivant : pour chaque famille de produits nous exposons les calibrages donnés par la sélection via BIC et les prévisions associées, suivi des calibrages fournis par le critère de sélection ICL et ses prévisions. Il est à noter que l'estimation des poids de chaque composante est rendu plus robuste par la diminution du nombre de composantes d'un mélange (ces graphiques ne sont pas exposés ici pour alléger la présentation). Nous limitons le nombre de composantes possibles à 9. Il est aussi important de garder en tête que les modèles sélectionnés résultent bien souvent de maximum locaux (on ne sait jamais vraiment si nous sommes tombés sur le maximum global) ; cependant la comparaison BIC/ICL est évidemment réalisée sur le même maximum !

Mixtos

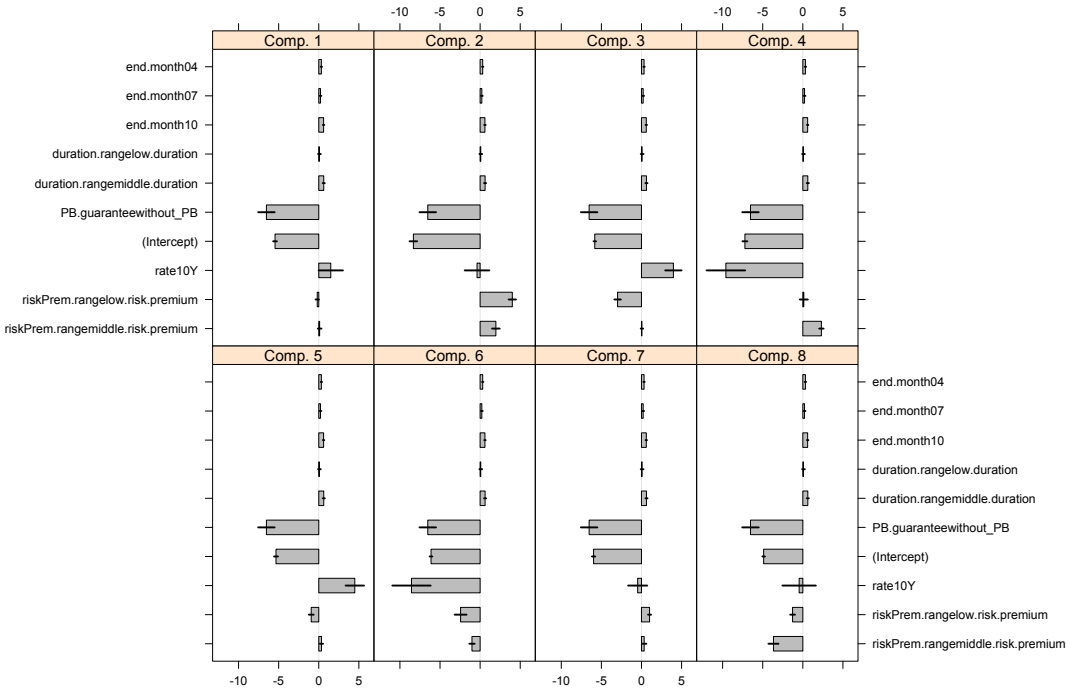


FIGURE 4.3 – Calibrage des coefficients de régression par BIC.

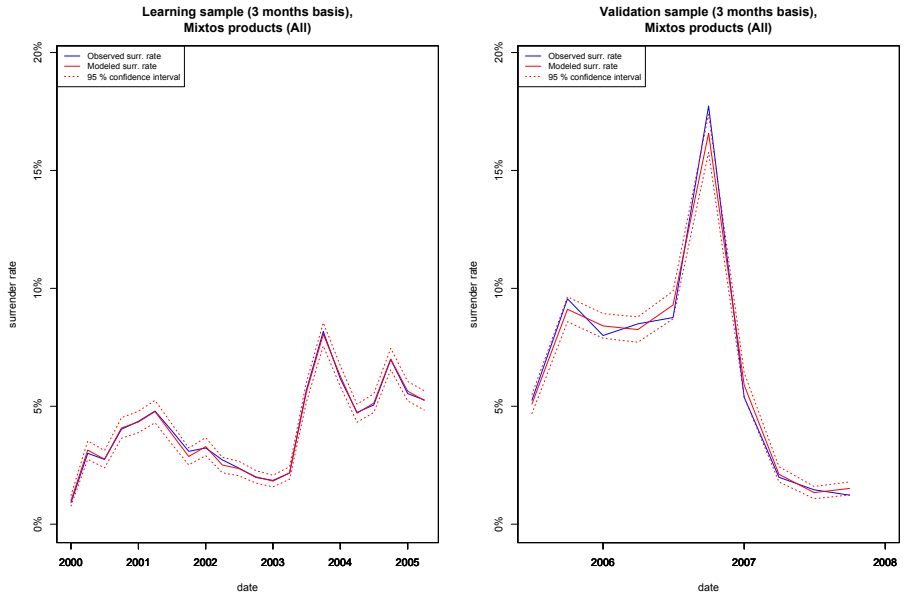


FIGURE 4.4 – Previsions de taux de rachat trimestriel par mélange de GLMs sélectionné via BIC.

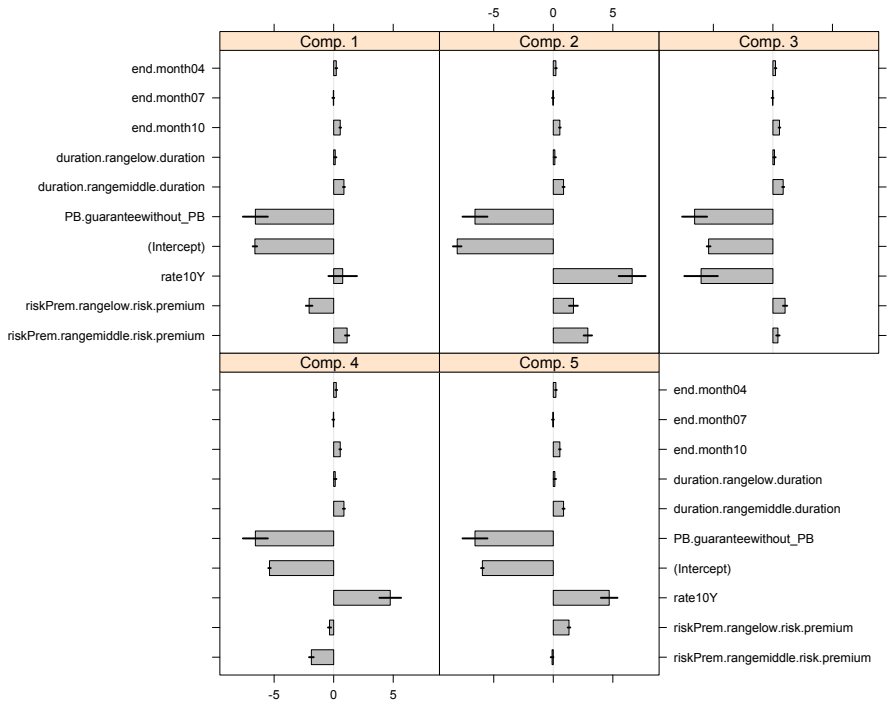


FIGURE 4.5 – Calibrage des coefficients de régression par ICL.

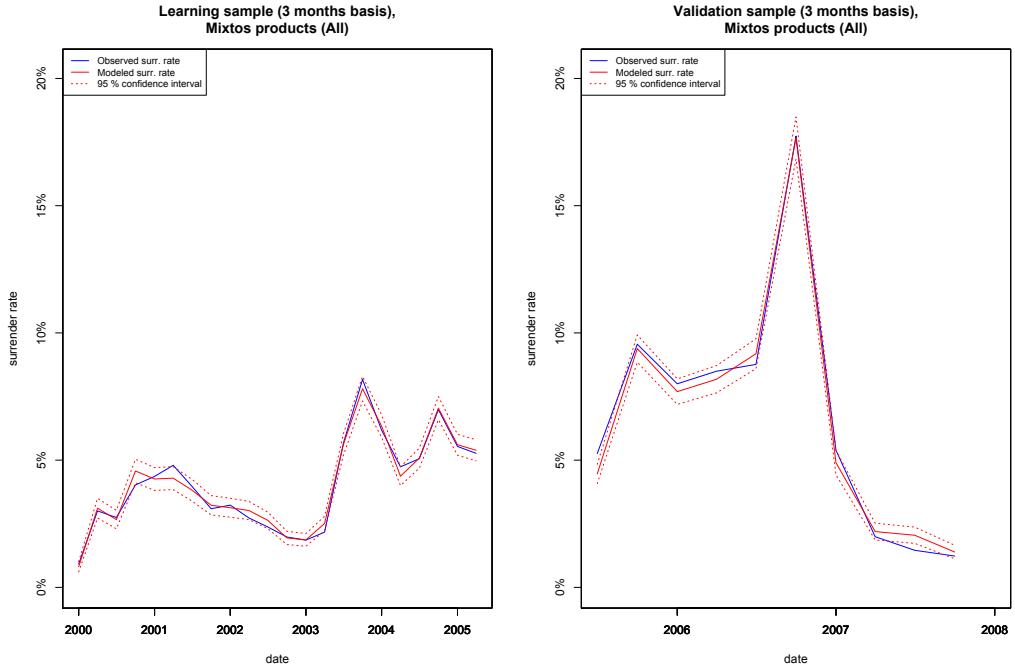


FIGURE 4.6 – Previsions de taux de rachat trimestriel par mélange de GLMs sélectionné via ICL.

Ahorro

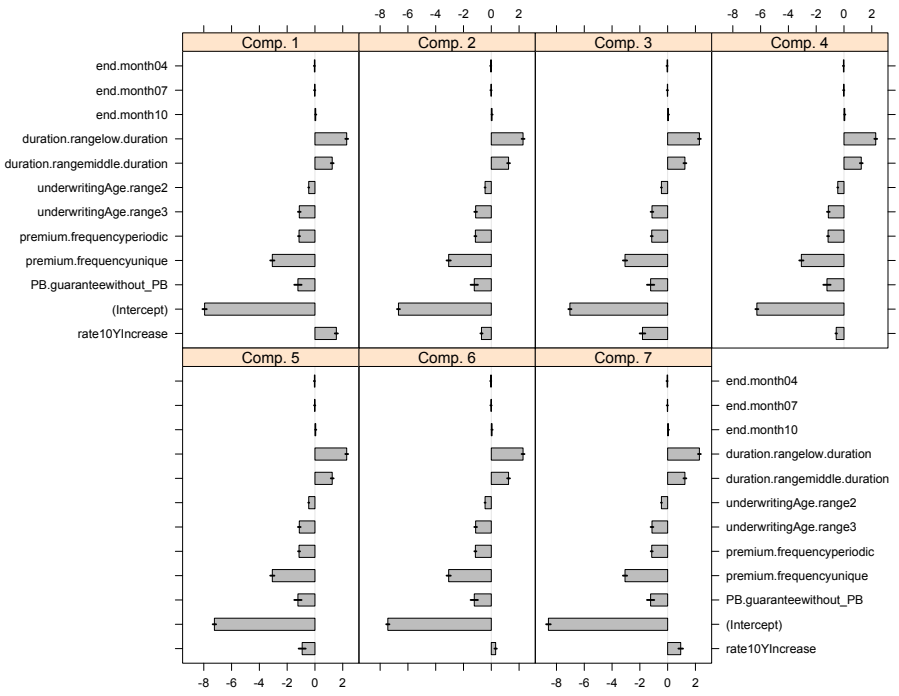


FIGURE 4.7 – Calibrage des coefficients de régression par BIC.

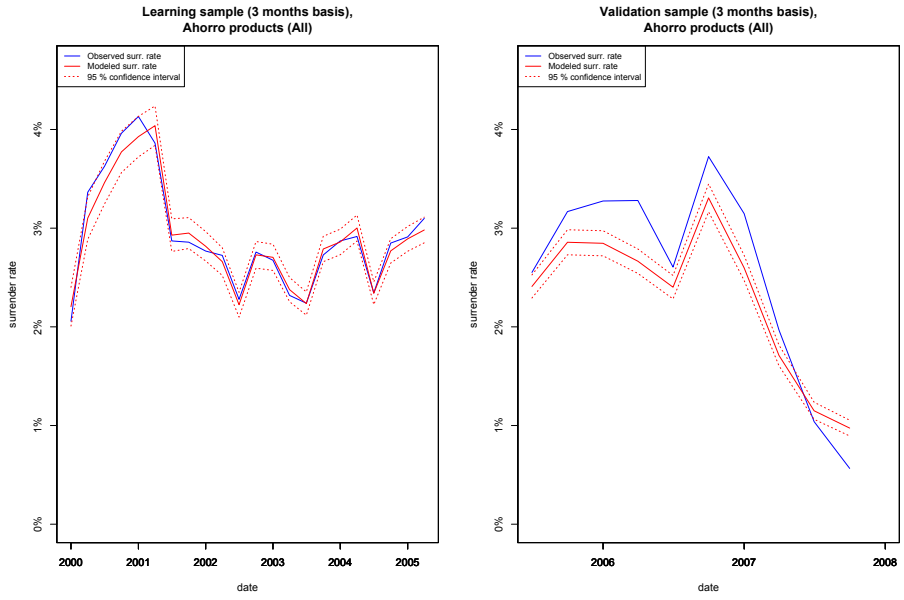


FIGURE 4.8 – Previsions de taux de rachat trimestriel par mélange de GLMs sélectionné via BIC.

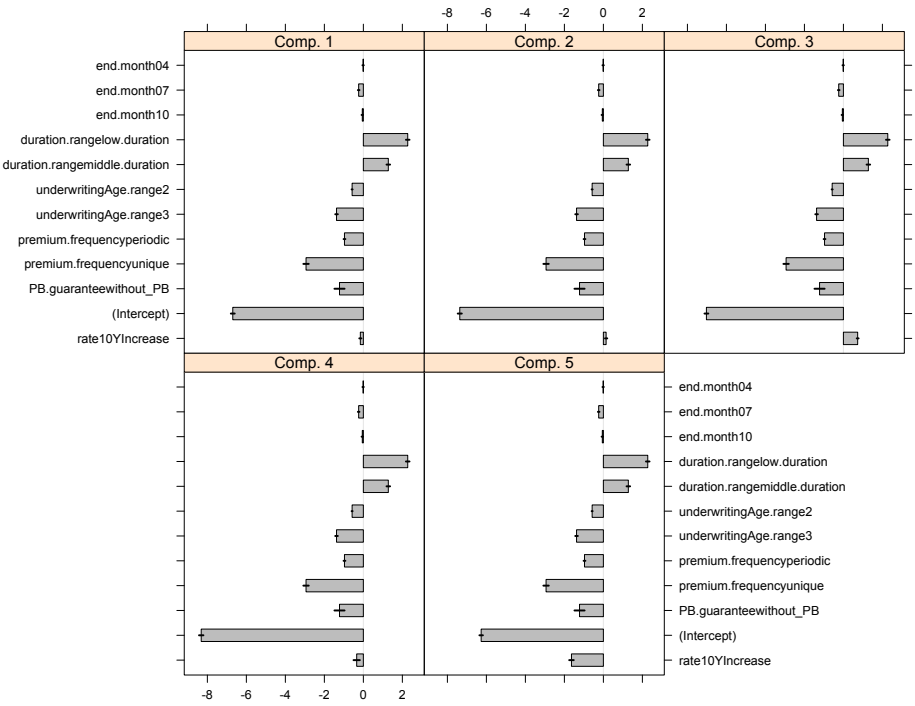


FIGURE 4.9 – Calibrage des coefficients de régression par ICL.

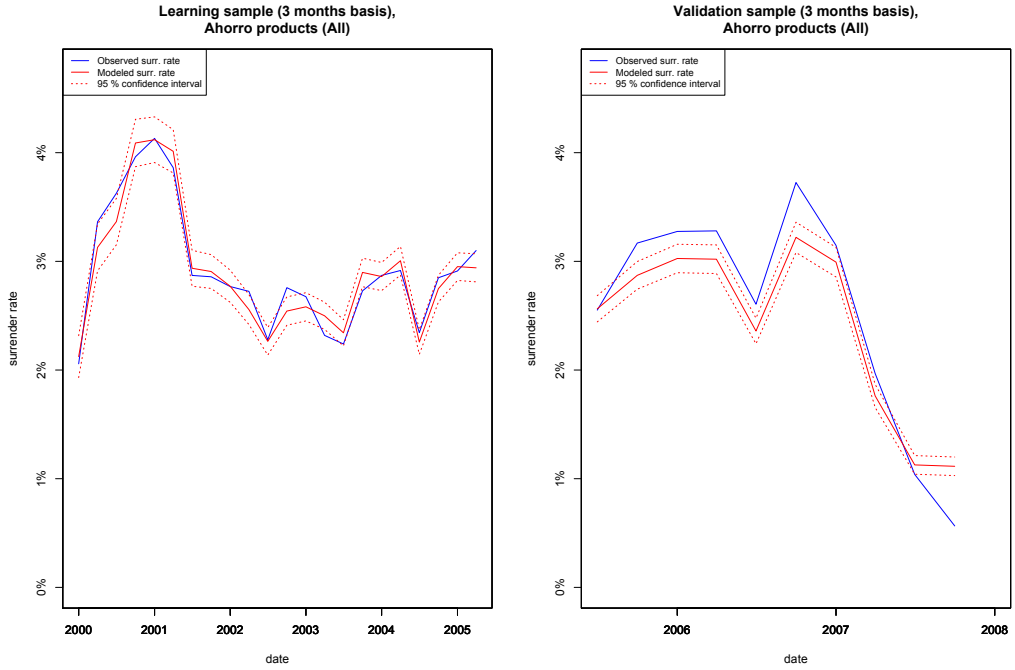


FIGURE 4.10 – Previsions de taux de rachat trimestriel par mélange de GLMs sélectionné via ICL.

Unit-Link

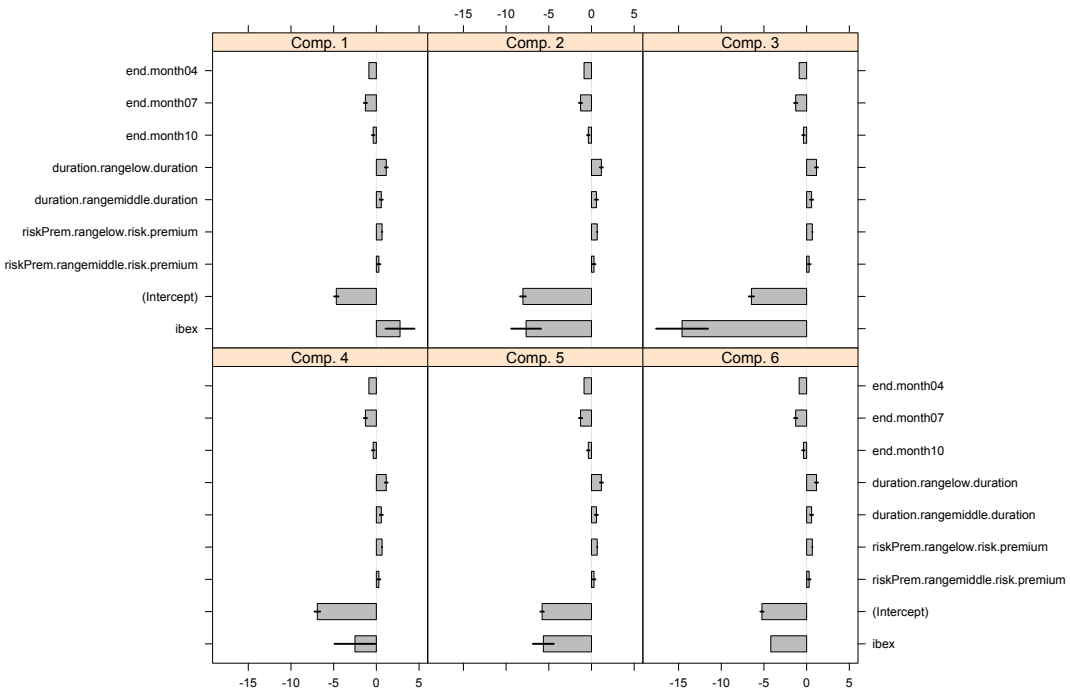


FIGURE 4.11 – Calibrage des coefficients de régression par BIC.

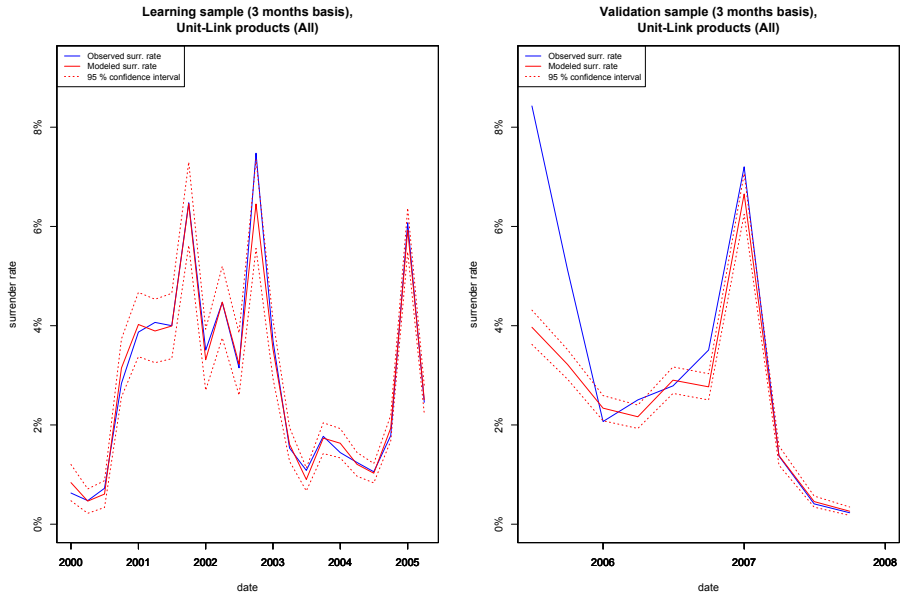


FIGURE 4.12 – Previsions de taux de rachat trimestriel par mélange de GLMs sélectionné via BIC.

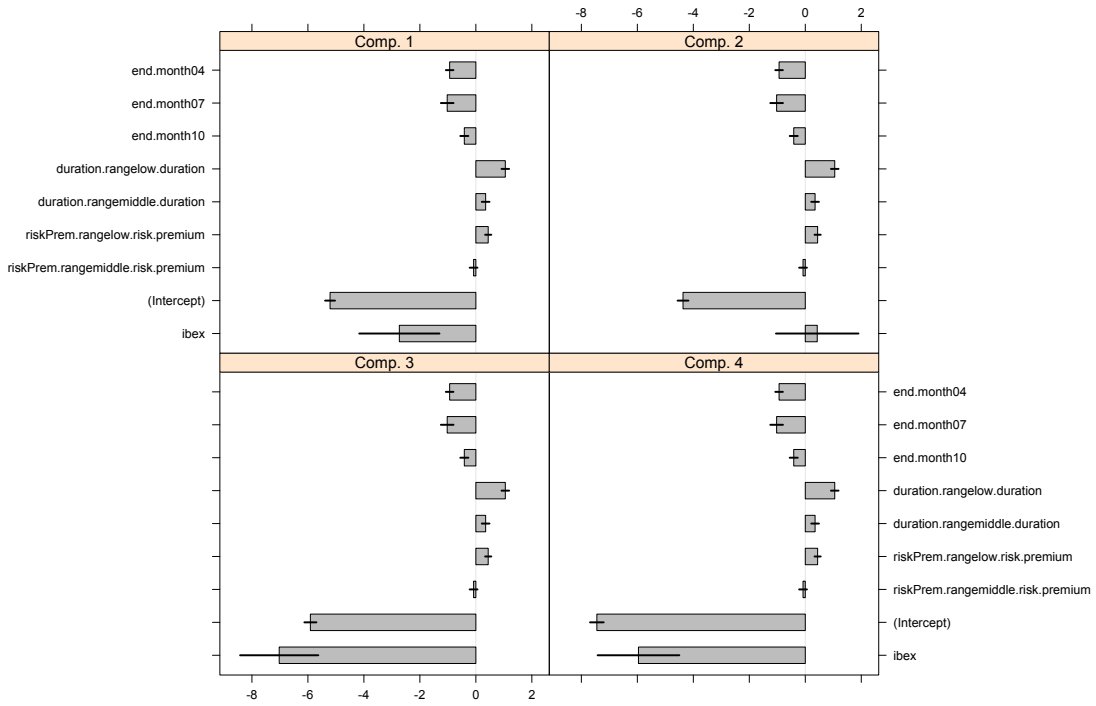


FIGURE 4.13 – Calibrage des coefficients de régression par ICL.

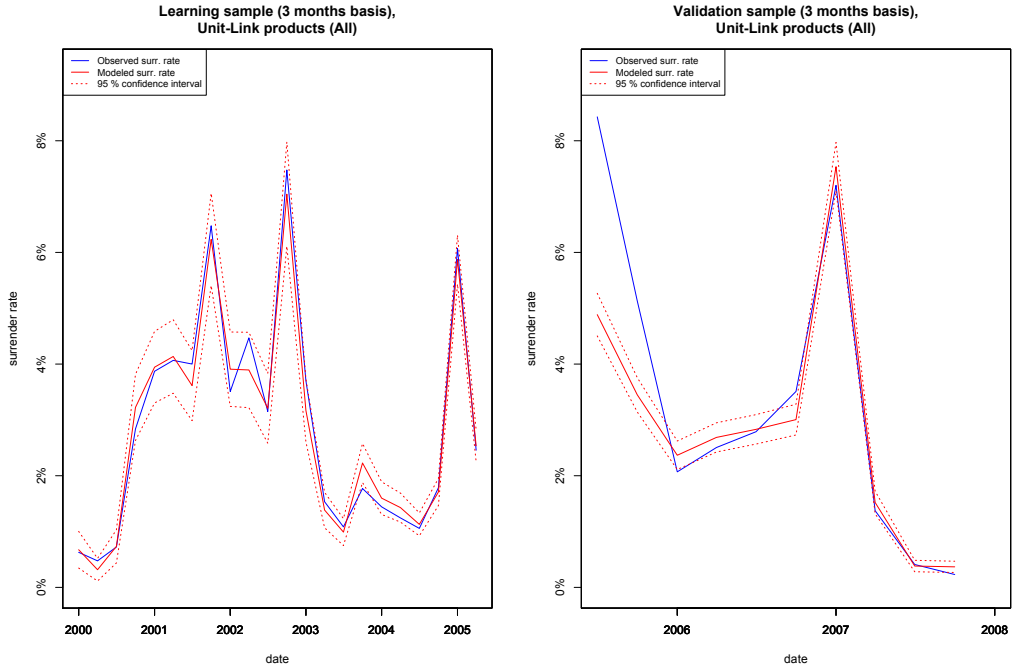


FIGURE 4.14 – Previsions de taux de rachat trimestriel par mélange de GLMs sélectionné via ICL.

Pure Savings

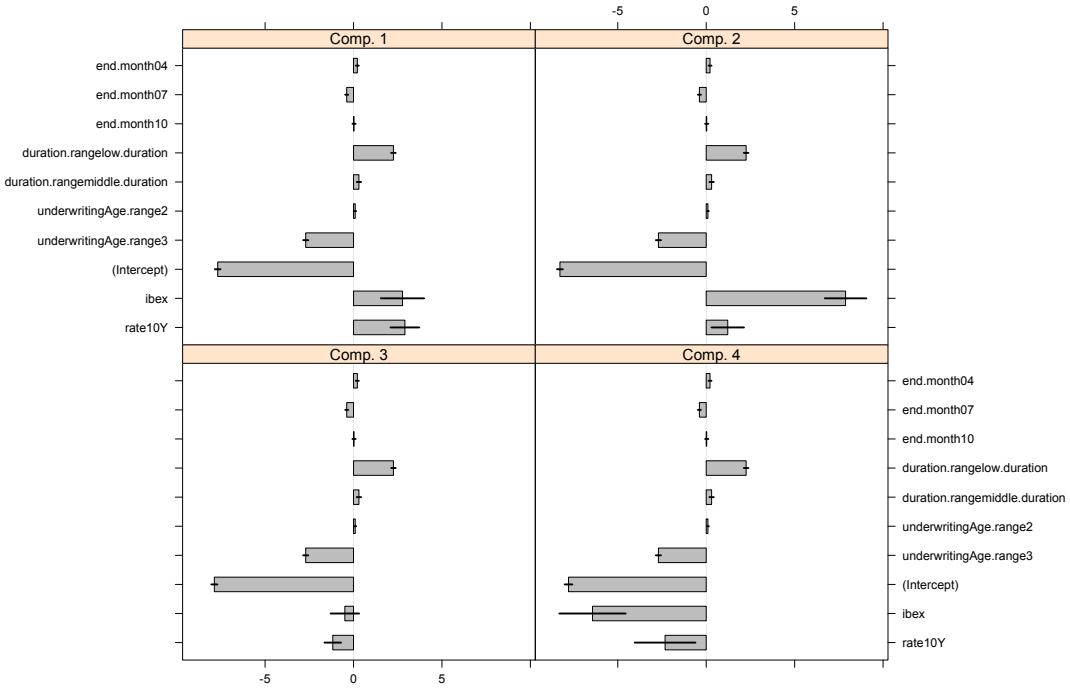


FIGURE 4.15 – Calibrage des coefficients de régression par BIC.

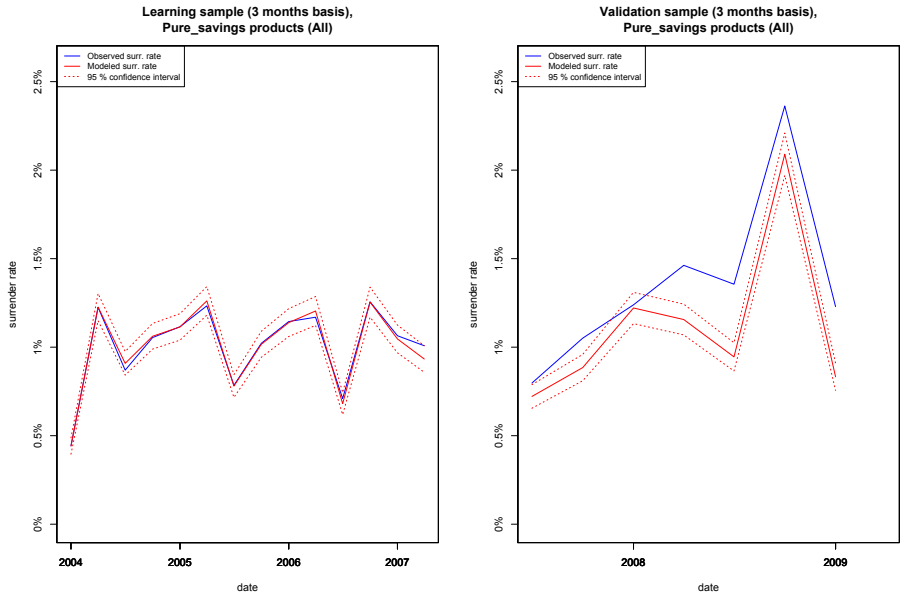


FIGURE 4.16 – Previsions de taux de rachat trimestriel par mélange de GLMs sélectionné via BIC.

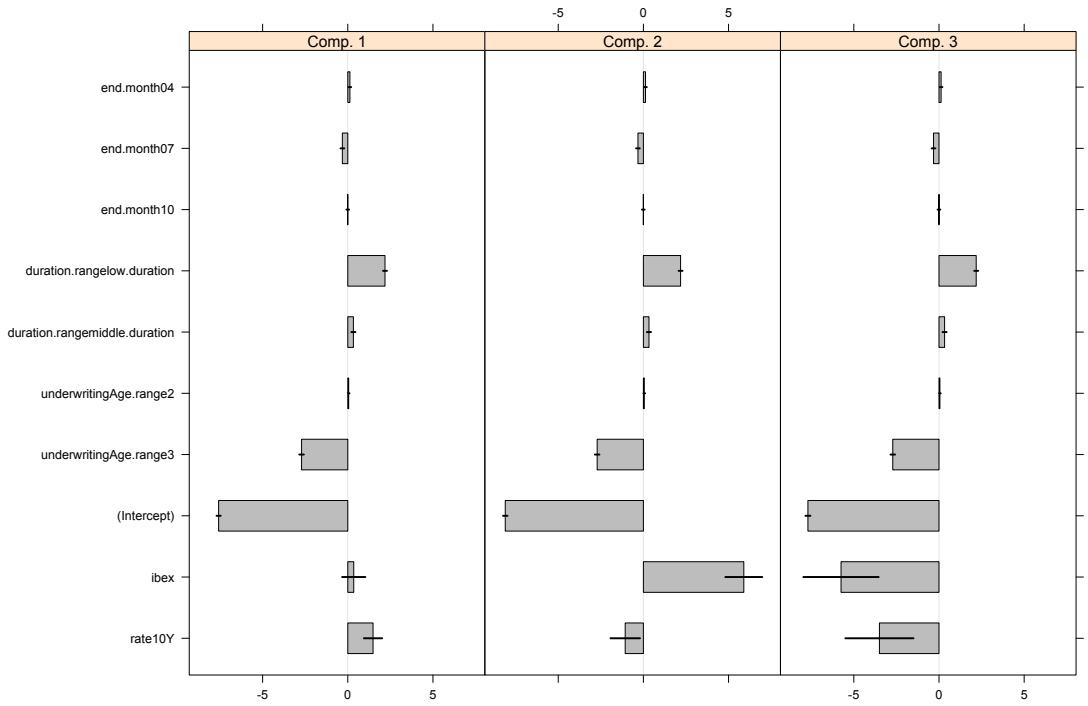


FIGURE 4.17 – Calibrage des coefficients de régression par ICL.

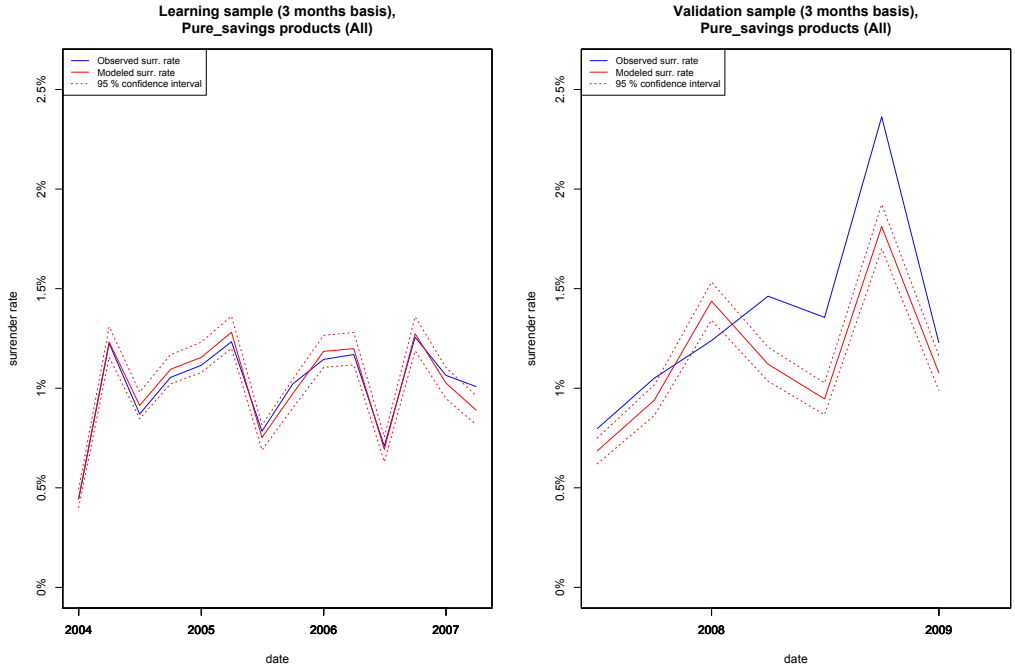


FIGURE 4.18 – Previsions de taux de rachat trimestriel par mélange de GLMs sélectionné via ICL.

Structured Products

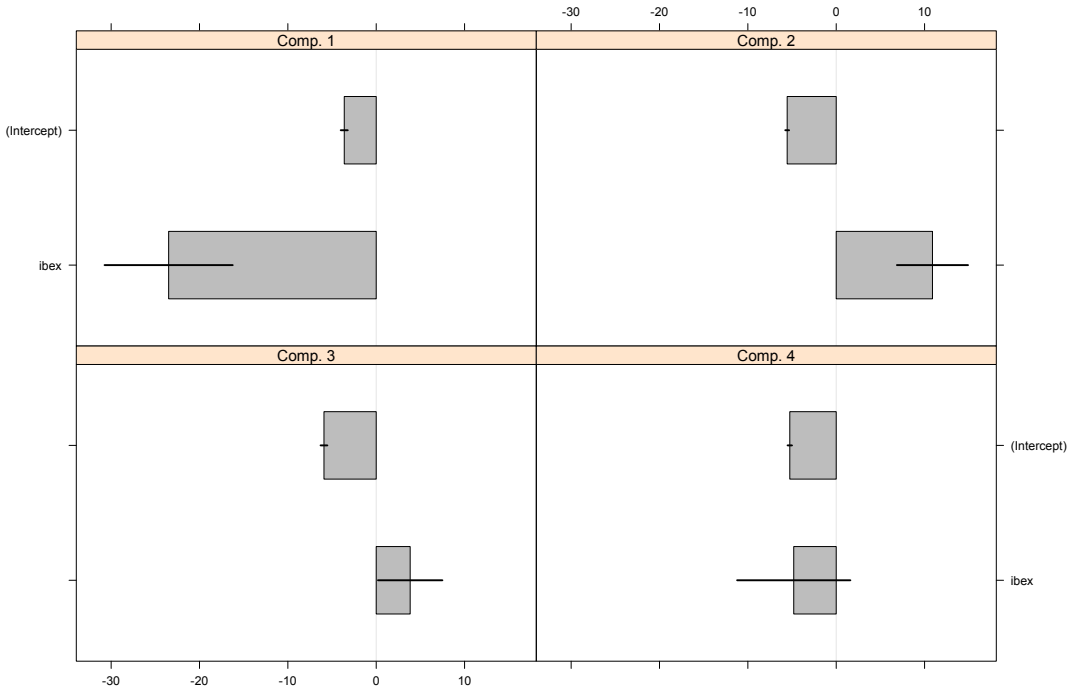


FIGURE 4.19 – Calibrage des coefficients de régression par BIC.

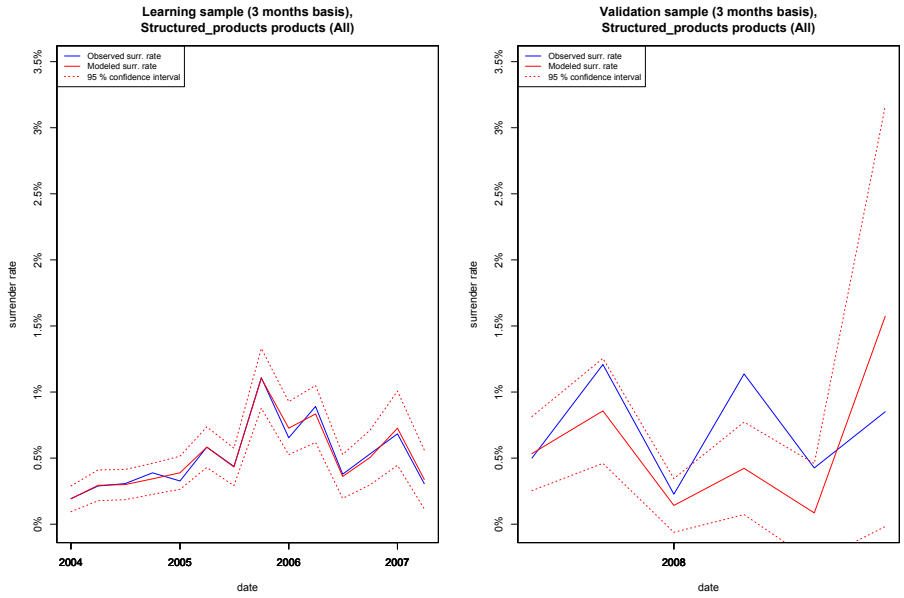


FIGURE 4.20 – Previsions de taux de rachat trimestriel par mélange de GLMs sélectionné via BIC.

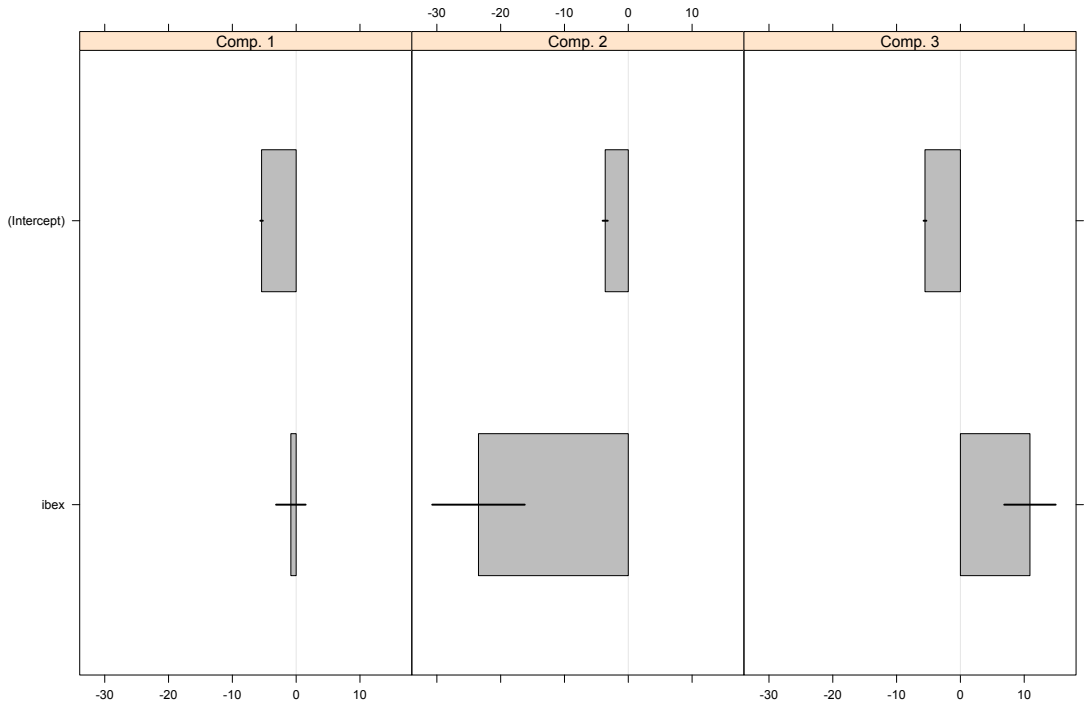


FIGURE 4.21 – Calibrage des coefficients de régression par ICL.

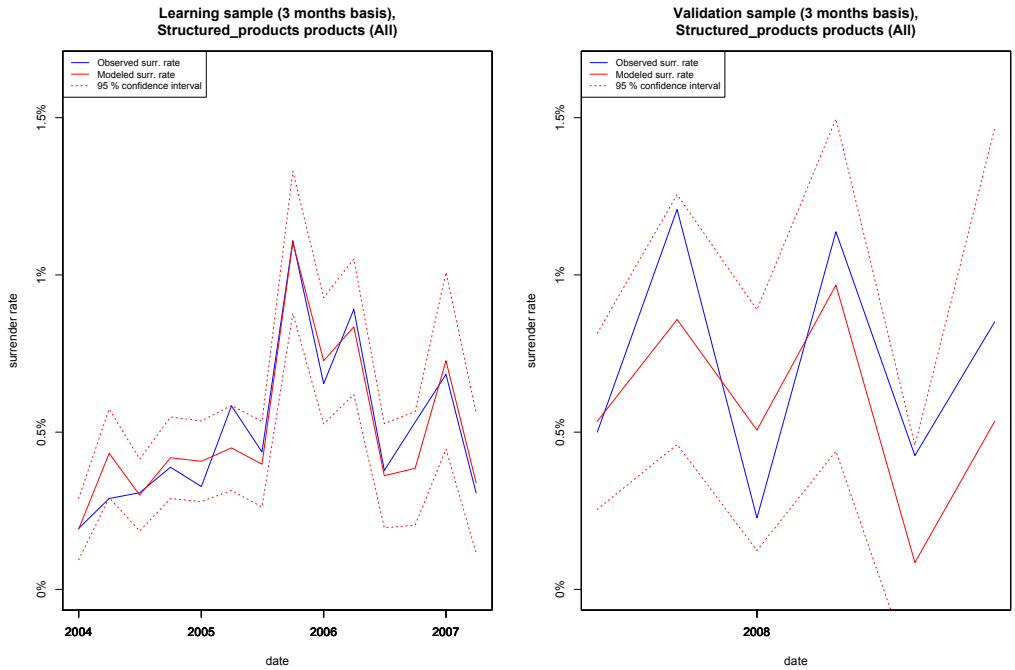


FIGURE 4.22 – Previsions de taux de rachat trimestriel par mélange de GLMs sélectionné via ICL.

Index-Link et Universal Savings

Nous n'avons pas trouvé de différence dans ce cas entre la sélection par BIC ou ICL, dans le sens où les deux critères sélectionnaient systématiquement le même nombre de composantes dans le mélange. Nous imaginons que le peu de nombre de composantes initialement sélectionné par l'un ou l'autre des critères est un signe d'une confiance relativement grande dans l'affectation des observations aux composantes. Les composantes sont originellement bien différenciée, or nous avons vu que l'estimation du nombre de composantes par BIC ou ICL est relativement similaire dans ce cas-là : en effet, le terme d'entropie tend vers une très petite valeur, ce qui provoque sa quasi-disparition.

4.5 Analyse et conclusion

Finalement, nous constatons que le critère ICL permet effectivement de diminuer et de créer des groupes de comportements d'assurés plus cohérents. De plus les prévisions dans le futur n'en demeurent pas moins correctes, alors même que l'estimation des coefficients de régression est globalement améliorée. Ainsi, la variance des estimateurs est plus petite, ce qui renforce le test d'hypothèse d'un coefficient de régression non-nul. Pour synthèse, le critère ICL a permis dans notre étude de diminuer le nombre de composantes dans cinq des sept cas abordés, tout en conservant une bonne approximation de la loi sous-jacente aux données réelles au vu des graphiques précédents. Ceci vient conforter les résultats théoriques de la première partie du chapitre, selon lesquels ce critère permet de sélectionner un mélange de GLMs dont le nombre de composantes est convergent vers le nombre théorique de composantes. Evidemment, ces premiers résultats nécessiteraient d'être davantage développés : ici par exemple, il est fréquent de tomber sur un maximum local lors de l'application numérique. Ceci a pour effet que nous ne sommes pas placés dans les conditions exactes d'application des théorèmes présentés.
La baisse en dimension la plus spectaculaire provient de la famille des produits Mixtes et de celle des produits en UC. Nous pouvions nous y attendre car nous avons constaté qu'il s'agissait effectivement des familles pour lesquelles nous avions des composantes particulièrement ressemblantes. Ensuite, certaines applications ont montré le peu d'intérêt du critère ICL, notamment lorsque le calibrage du mélange a permis d'effectuer une première segmentation robuste : ce résultat était aussi anticipable car les densités de chaque composante semblaient déjà bien séparées dans ces cas-là.

Comme future extension, nous pensions à travailler avec la théorie du minimum de contraste pour trouver des formes de pénalités non-asymptotiques. Dans le même esprit que certaines approches qui servent à calibrer spécialement la pénalité (heuristique de pente, ...), nous pourrions aussi définir des critères de sélection dépendants des données (Yang (2005)).

Bibliographie

Akaike, H. (1973), *Information theory as an extension of the maximum likelihood principle*, second international symposium on information theory edn, B.N Petrov and F. Csaki. 104, 107, 108

Baudry, J. (2009), Sélection de modèle pour la classification non supervisée. Choix du nombre de classes., PhD thesis, Université Paris Sud XI. 115, 116, 117, 118, 124, 127, 131, 132, 134, 135, 160

Baudry, J.-P., Celeux, G. and Marin, J. (2008), Selecting models focussing on the modeler's purpose, Proceedings in Computational Statistics, Physica-Verlag, Heidelberg, pp. 337–348. 104

Bickel, P. and Doksum, K. (2001), *Mathematical Statistics, Vol. I*, Second Edition, Prentice Hall. 131

Biernacki, C. (2000), 'Assessing a mixture model for clustering with the integrated completed likelihood', *IEEE Transactions on PAMI* **22**, 719–725. 132, 133, 134

Biernacki, C. (2009), 'Pourquoi les modèles de mélange pour la classification ?', *MODULAD* **40**. 104, 122

Biernacki, C., Celeux, G., Govaert, G. and Langrognet, F. (2006), 'Model-based cluster and discriminant analysis with the mixmod software', *Computational Statistics and Data Analysis* **51**(2), 587–600. 132

Biernacki, C. and Govaert, G. (1997), 'Using the classification likelihood to choose the number of clusters', *Computer Science and Statistics* **29**, 451–457. 104

Celeux, G. and Govaert, G. (1992), 'A classification em algorithm for clustering and two stochastic versions', *Computational Statistics and Data Analysis* **14**(3), 315–332. 119

Celeux, G. and Soromenho, G. (1996), 'An entropy criterion for assessing the number of clusters in a mixture model', *Classification Journal* (13), 195–212. 116

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton. 106

Dempster, A., N.M., L. and D.B., R. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society* **39**, 1–38. 115

Doob, J. (1934), 'Probability and statistics', *Transactions of the American Mathematical Society* **36**. 106

Dudley, R. (1999), *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics. 126

Dutang, C. (2011), Regression models of price elasticity in non-life insurance, Master's thesis, ISFA. Mémoire confidentiel - AXA Group Risk Management. 147

Fraley, C. and Raftery, A. (1998), 'How many clusters ? which clustering method ? answer via model-based cluster analysis', *The Computer Journal* (41), 578–588. 116

Garel, B. (2007), 'Recent asymptotic results in testing for mixtures', *Computational Statistics and Data Analysis* **51**, 5295–5304. 115

Grun, B. and Leisch, F. (2004), *Bootstrapping finite mixture models*, compstat'2004 symposium edn, Physica Verlag, Heidelberg. 147

Grun, B. and Leisch, F. (2007), 'Fitting finite mixtures of generalized linear regressions in r', *Computational Statistics and Data Analysis* **51**, 5247–5252. 147

Grun, B. and Leisch, F. (2008), Identifiability of finite mixtures of multinomial logit models with varying and fixed effects, Technical Report 24, Department of Statistics, University of Munich. 147

Hai Xan, W., Bin, L., Quan bing, Z. and Sui, W. (2004), 'Estimation for the number of components in a mixture model using stepwise split-and-merge em algorithm', *Pattern Recognition Letters* **25**, 1799–1809. 115

Hathaway, R. (1986), 'A constrained em algorithm for univariate normal mixtures', *Journal of Statistical Computation and Simulation* **23**(3), 211–230. 119

Keribin, C. (1999), Tests de modèles par maximum de vraisemblance, PhD thesis, Université d'Evry Val d'Essonne. 116, 117

Kullback, S. and Leibler, R. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* **22**(1), 79–86. 104, 105

Lebarbier, E. and Mary-Huard, T. (2004), Le critère bic : fondements théoriques et interprétation, Technical Report 5315, INRIA. 111

Leisch, F. (2008), Modelling background noise in finite mixtures of generalized linear regression models, Technical Report 37, Department of Statistics, University of Munich. 147

Mallows, C. (1974), 'Some comments on cp', *Technometrics* **15**, 661–675. 104

Massart, P. (2007), *Concentration inequalities and model selection. Ecole d'été de Probabilités de Saint-Flour 2003.*, Lecture Notes in Mathematics, Springer. 135, 138, 139, 140

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models, 2nd ed.*, Chapman and Hall. 147, 150

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, Wiley Series In Porbability and Statistics. 116, 122, 133, 134

Mun, E.-Y., von Eye, A., Bates, M. and Vaschillo, E. (2008), 'Finding groups using model-based cluster analysis : heterogeneous emotional self-regulatory processes and heavy alcohol use risk', *Developmental Psychology* **44**, 481–495. 111

Nishii, R. (1988), 'Maximum likelihood principle and model selection when the true model is unspecified', *Journal of Multivariate Analysis* (27), 392–403. 106, 107, 113, 134, 137

Ohlson, E. and Johansson, B. (2010), *Non-Life Insurance Pricing with Generalized Linear Models*, Springer. 147

Oliviera-Brochado, A. and Vitorino Martins, F. (2005), Assessing the number of components in mixture models : a review. Working Paper. 115

Oliviera-Brochado, A. and Vitorino Martins, F. (2008), Determining the number of market segments using an experimental design. Working Paper. 116

Raftery, A. (1994), Bayesian model selection in social research (with discussion), Technical Report 94-12, Demography Center Working, University of Washington. 111

Redner, R. (1981), 'Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions', *The Annals of Statistics* **9**(1), 225–228. 106, 118

Redner, R. and Walker, H. (1984), 'Mixture densities, maximul likelihood and the em algorithm', *SIAM Review* **26**(2), 195–239. 117

Ripley, B. (1995), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge. 113

Sarstedt, M., Becker, J.-M., Ringle, C. and Schwaiger, M. (2011), 'Uncovering and treating unobserved heterogeneity with fixmix-pls : which model selection criterion provides an appropriate number of segments ?', *Schmalenbach Business Review* **63**, 34–62. 116

Schlattmann, P. (2003), 'Estimating the number of components in a finite mixture model : the special case of homogeneity', *Computational Statistics and Data Analysis* **41**, 441–451. 115

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464. 104, 111

van der Vaart, A. (1998), *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge. 124, 127, 135

Wald, A. (1949), 'Note on the consistency of the maximum likelihood estimate', *The Annals of Mathematical Statistics* **20**(4), 595–601. 106

Wenbin, L. (2006), 'Penalized minimum matching distance-guided em algorithm', *International Journal of Electronics and Communications* **60**, 235–239. 115

Yang, Y. (2005), 'Can the strengths of aic and bic be shared ? a conflict between model identification and regression estimation', *Biometrika* **92**(4), 937–950. 171

Conclusion et annexes

Conclusion et perspectives

Cette étude nous a permis de mieux comprendre comment appréhender les comportements de rachat grâce à une vision agrégée, par une définition plus complète et moins détaillée. L'étude empirique de différentes lignes de produits du portefeuille entier d'une entité d'AXA a été très instructive, et a servi de point de départ aux choix qui ont été adoptés par la suite. La volonté de comprendre, segmenter et modéliser les rachats à l'échelle de grandes familles de produit en utilisant des bases de données agrégées induit des soucis de modélisation (à cause de l'hétérogénéité entre ces produits), mais se révèle clairement mieux adaptée pour une étude d'impact des rachats au niveau de la gestion actif-passif d'une compagnie d'assurance.

Le risque de rachat est un risque comportemental, donc par nature difficilement modélisable car dépendant de nombreux facteurs aussi bien endogènes qu'exogènes. Suivant les positions de l'assureur et le type de produit, l'impact d'un scénario averse des comportements de rachat peut être très conséquent. A ce titre, nous avons vu qu'il était primordial de prendre en compte la possible corrélation entre les comportements d'assurés lors de la modélisation dans une optique de gestion des risques (affinement du modèle interne partiel). En effet cette dépendance entraîne une déformation de la distribution des rachats, qui provoque une hausse conséquente de la marge de risque. De plus, certaines caractéristiques clefs ne peuvent pas être négligées lors de la modélisation en cas d'évolution de la composition du portefeuille : nous pensons à l'ancienneté du contrat, à la richesse de l'assuré, au réseau de distribution. D'autres comme l'état de l'économie ou la réputation de l'entreprise, plus difficile à prendre en compte, jouent également un rôle prépondérant. Cette remarque suppose l'utilisation de modèles de régression, mais d'une manière que nous avons faite évoluer tout au long de ce projet. Comme une alternative aux précédentes approches bibliographiques, le coeur de ce travail s'articule autour d'une approche probabiliste invoquant une prise en compte spécifique des facteurs de risque structurels et conjoncturels. Nous évitons par cette approche l'hypothèse de rationalité et d'optimalité (au sens financier du terme) des assurés, et définissons une méthodologie **uniformisée** et semble-t-il **efficace** pour traiter le problème de la modélisation des rachats.

La principale "*leçon*" que nous retirons de cette étude est qu'il est inutile de prendre trop de facteurs de risque en compte : la saisonnalité, l'ancienneté de contrat, un troisième facteur de risque discriminant (endogène) et le contexte économique et financier suffisent en général à une bonne modélisation des comportements. La deuxième "découverte" concerne la manière de les considérer : les effets des facteurs idiosynchratiques doivent être fixés égaux entre les composantes des mélanges alors que le contexte économique joue différemment sur les assurés. Cela permet à la fois de définir un cadre logique d'étude et de limiter le nombre de paramètres à estimer, donc la dimension de l'espace (permettant de meilleures prévisions). Les résultats de l'étude sont d'autant plus satisfaisants que la méthode de validation choisie (*back-testing*) fait intervenir de nouveaux contextes économiques en permanence (à cause de la crise financière), garantissant une bonne quantification des effets exogènes par la modélisation. Nous attirons

l'attention du lecteur sur le fait que nous ne prétendons pas avoir trouvé la méthode idéale pour la modélisation des comportements de rachat ; simplement dans les différents cas testés, cette modélisation s'est avérée relativement fine et robuste. Un phénomène extrême qui ne serait pas dû à des mouvements sur les marchés financiers (exemple : politique de vente, image) pourrait avoir des conséquences dramatiques sur le taux de rachat mais ne serait évidemment pas prévu par notre modèle.

Perspectives

Sur le plan théorique, la modélisation de ce phénomène nous a poussé à étudier dans le détail un large panel d'outils mathématiques, que nous n'avons bien sûr pas pu totalement exploiter du fait de leur diversité. Ainsi les méthodes de classification par arbre ont été abordées, la compréhension des modèles linéaires généralisés et de leurs caractéristiques a dû être approfondie. Nous avons également largement développé nos connaissances en analyse de survie non-paramétrique et semi-paramétrique (en vain pour les applications !) et en modèles markoviens cachés. Lorsque nous avons réussi à résoudre le problème opérationnel de la prévision précise des rachats, nous avons ensuite pu nous pencher sur une étude plus poussée concernant le choix du nombre de composantes dans les modélisations mélange. En particulier, nous nous sommes intéressés à cette question pour des mélanges de GLMs après avoir constaté une potentielle sur-paramétrisation dans nos applications. Suite à des discussions sur le sujet avec des chercheurs expérimentés (Laurent Bordes, Bernard Garel, Gilles Celeux, Jean-Patrick Baudry) que je remercie par la même occasion, nous avons pu prouver la consistance d'un nouvel estimateur, ainsi que celle d'un critère de sélection adapté aux problématiques de classification non-supervisée dans le contexte de la modélisation mélange de GLMs. Malheureusement nous avons manqué de temps pour développer tout un ensemble de résultats que nous aurions aimé traiter, et qui constitueront les prochaines recherches que nous effectuerons dans les mois à venir. Il serait par exemple intéressant de s'approprier les notions d'inégalités de concentration et d'heuristique de pente afin de déterminer des formes de pénalité optimale dans un cadre non-asymptotique. Une autre piste très générale concerne l'établissement d'un lien entre modélisation mélange et forme paramétrique spécifique, avec les difficultés et avantages qui en découlent. Enfin nous aimerions reconsidérer l'analyse de survie pour laquelle nous avons "dépensé" beaucoup de temps, mais sur laquelle nous n'avons pas réussi à aboutir. Néanmoins, nous n'étions pas loin d'un résultat et les problématiques traitées dans le cadre des GLMs sont clairement transposables avec des modèles de survie. L'approche par modélisation markovienne cachée ("regime switching") est également attractive d'un point de vue intellectuel ; et il va sans dire que nous aimerions également investiguer ses propriétés théoriques et les résultats qu'elle procurerait dans un contexte d'analyse de survie.

Finalement, cette expérience m'a ouvert les yeux sur mon désir de m'investir davantage dans le monde académique. La diversité des sujets et la curiosité intellectuelle qu'offre ce milieu sont tout à fait uniques et constituent son attrait le plus fort. Je me suis découvert une forte envie de "fouiller" dans les sujets, que seule la Recherche permet d'assouvir. Pour finir, l'écriture d'articles, l'intéraction avec les autres chercheurs ainsi que l'enseignement sont autant de facettes qui m'ont aidé à m'épanouir et faire de cette thèse un vrai succès personnel.

Bibliographie

Akaike, H. (1973), *Information theory as an extension of the maximum likelihood principle*, second international symposium on information theory edn, B.N Petrov and F. Csaki.

Albert, F. S., Bragg, D. G. W. and Bragg, J. M. (1999), 'Mortality rates as a function of lapse rates', *Actuarial research clearing house* **1**.

Atkins, D. C. and Gallop, R. J. (2007), 'Re-thinking how family researchers model infrequent outcomes : A tutorial on count regression and zero-inflated models', *Journal of Family Psychology* .

Austin, P. C. (2007), 'A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting ami mortality', *Statistics in Medicine* **26**, 2937–2957.

Bacinello, A. R. (2005), 'Endogenous model of surrender conditions in equity-linked life insurance', *Insurance : Mathematics and Economics* **37**, 270–296.

Bacinello, A. R., Biffis, E. and P., M. (2008), 'Pricing life insurance contracts with early exercise features', *Journal of Computational and Applied Mathematics* .

Balakrishnan, N. (1991), *Handbook of the Logistic Distribution*, Marcel Dekker, Inc.

Baudry, J. (2009), Sélection de modèle pour la classification non supervisée. Choix du nombre de classes., PhD thesis, Université Paris Sud XI.

Baudry, J.-P., Celeux, G. and Marin, J. (2008), Selecting models focussing on the modeler's purpose, Proceedings in Computational Statistics, Physica-Verlag, Heidelberg, pp. 337–348.

Biard, R., Lefèvre, C. and Loisel, S. (2008), 'Impact of correlation crises in risk theory : Asymptotics of finite-time ruin probabilities for heavy-tailed claim amounts when some independence and stationarity assumptions are relaxed', *Insurance : Mathematics and Economics* **43**(3), 412 – 421.

Bickel, P. and Doksum, K. (2001), *Mathematical Statistics, Vol. I*, Second Edition, Prentice Hall.

Biernacki, C. (2000), 'Assessing a mixture model for clustering with the integrated completed likelihood', *IEEE Transactions on PAMI* **22**, 719–725.

Biernacki, C. (2009), 'Pourquoi les modèles de mélange pour la classification ?', *MODULAD* **40**.

Biernacki, C., Celeux, G., Govaert, G. and Langrognet, F. (2006), 'Model-based cluster and discriminant analysis with the mixmod software', *Computational Statistics and Data Analysis* **51**(2), 587–600.

Biernacki, C. and Govaert, G. (1997), 'Using the classification likelihood to choose the number of clusters', *Computer Science and Statistics* **29**, 451–457.

Bluhm, W. F. (1982), 'Cumulative antiselection theory', *Transactions of Society of actuaries* **34**.

Bohning, D., Dietz, E., Schaub, R., Schlattmann, P. and Lindsay, B. (1994), 'The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family', *Annals of the Institute of Statistical Mathematics* **46**, 373–388.

Bohning, D. and Seidel, W. (2003), 'Editorial : recent developments in mixture models', *Computational Statistics and Data Analysis* **41**, 349–357.

Box, G. and Cox, D. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society B* **26**, 211–252.

Breiman, L. (1994), Bagging predictors, Technical Report 421, Department of Statistics, University of California.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* (24), 123–140.

Breiman, L. (1998), 'Arcing classifiers', *The Annals of Statistics* **26**(3), 801–849.

Breiman, L. (2001), 'Random forests', *Machine Learning* (45), 5–32.

Breiman, L., Friedman, J., Olshen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, Chapman and Hall.

Celeux, G. and Govaert, G. (1992), 'A classification em algorithm for clustering and two stochastic versions', *Computational Statistics and Data Analysis* **14**(3), 315–332.

Celeux, G. and Soromenho, G. (1996), 'An entropy criterion for assessing the number of clusters in a mixture model', *Classification Journal* (13), 195–212.

Costabile, M., Massabo, I. and Russo, E. (2008), 'A binomial model for valuing equity-linked policies embedding surrender options', *Insurance : Mathematics and Economics* **40**, 873–886.

Cox, D. (1972), 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society : Series B* (34), 187–220.

Cox, S. H. and Lin, Y. (2006), Annuity lapse rate modeling : tobit or not tobit ?, *in* 'Society of actuaries'.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton.

Cummins, J. (1975), *An econometric model of the life insurance sector in the U.S. economy*, Lexington books, Health, Lexington/Mass u.a.

De Giovanni, D. (2007), Lapse rate modeling : A rational expectation approach, Finance Research Group Working Papers F-2007-03, University of Aarhus, Aarhus School of Business, Department of Business Studies.

Dempster, A., N.M., L. and D.B., R. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society* **39**, 1–38.

Denuit, M., Lefèvre, C. and Shaked, M. (1998), 'The s-convex orders among real random variables, with applications', *Math. Inequal. Appl.* **1**(4), 585–613.

Doob, J. (1934), 'Probability and statistics', *Transactions of the American Mathematical Society* **36**.

Dutang, C. (2011), Regression models of price elasticity in non-life insurance, Master's thesis, ISFA. Mémoire confidentiel - AXA Group Risk Management.

Engle, R. and Granger, C. (1987), 'Cointegration and error-correction : Representation, estimation and testing', *Econometrica* (55), 251–276.

Fauvel, S. and Le Pévédic, M. (2007), Analyse des rachats d'un portefeuille vie individuelle : Approche théorique et application pratique, Master's thesis, ENSAE. Mémoire non confidentiel - AXA France.

Follmann, D. and Lambert, D. (1989), 'Generalizing logistic regression by non-parametric mixing', *Journal of the American Statistical Association* **84**, 295–300.

Fraley, C. and Raftery, A. (1998), 'How many clusters ? which clustering method ? answer via model-based cluster analysis', *The Computer Journal* (41), 578–588.

Fum, D., Del Missier, F. and A., S. (2007), 'The cognitive modeling of human behavior : Why a model is (sometimes) better than 10,000 words', *Cognitive Systems Research* **8**, 135–142.

Garel, B. (2007), 'Recent asymptotic results in testing for mixtures', *Computational Statistics and Data Analysis* **51**, 5295–5304.

Ghattas, B. (1999), 'Previsions par arbres de classification', *Mathematiques et Sciences Humaines* **146**, 31–49. 193

Ghattas, B. (2000*a*), 'Aggregation d'arbres de classification', *Revue de statistique appliquee* **2**(48), 85–98.

Ghattas, B. (2000*b*), Importance des variables dans les methodes cart. GREQAM - Universite de Marseille.

Ghosh, J. and Sen, P. (1985), 'On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results', **2**, 789–806. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer.

Grun, B. and Leisch, F. (2004), *Bootstrapping finite mixture models*, compstat'2004 symposium edn, Physica Verlag, Heidelberg.

Grun, B. and Leisch, F. (2007), 'Fitting finite mixtures of generalized linear regressions in r', *Computational Statistics and Data Analysis* **51**, 5247–5252.

Grun, B. and Leisch, F. (2008), Identifiability of finite mixtures of multinomial logit models with varying and fixed effects, Technical Report 24, Department of Statistics, University of Munich.

Hai Xan, W., Bin, L., Quan bing, Z. and Sui, W. (2004), 'Estimation for the number of components in a mixture model using stepwise split-and-merge em algorithm', *Pattern Recognition Letters* **25**, 1799–1809.

Hathaway, R. (1986), 'A constrained em algorithm for univariate normal mixtures', *Journal of Statistical Computation and Simulation* **23**(3), 211–230.

Hilbe, J. M. (2009), *Logistic regression models*, Chapman and Hall.

Hin, H. K. and Huiyong, S. (2006), 'Structural prepayment risk behavior of the underlying mortgages for residential mortgage life insurance in a developing market', *Journal of Housing Economics* (15), 257–278.

Hosmer, D. W. and Lemeshow, S. (2000), *Applied Logistic Regression, 2nd ed.*, Wiley.

Kagraoka, Y. (2005), Modeling insurance surrenders by the negative binomial model. Working Paper 2005.

Keribin, C. (1999), Tests de modèles par maximum de vraisemblance, PhD thesis, Université d'Evry Val d'Essonne.

Kim, C. (2005), 'Modeling surrender and lapse rates with economic variables', *North American Actuarial Journal* pp. 56–70.

Kim, C. N., Yang, K. H. and Kim, J. (2008), 'Human decision-making behavior and modeling effects', *Decision Support Systems* **45**, 517–527.

Kuen, S. T. (2005), 'Fair valuation of participating policies with surrender options and regime switching', *Insurance : Mathematics and Economics* **37**, 533–552.

Kullback, S. and Leibler, R. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* **22**(1), 79–86.

Lebarbier, E. and Mary-Huard, T. (2004), Le critère bic : fondements théoriques et interprétation, Technical Report 5315, INRIA.

Lee, S., Son, Y.-J. and Jin, J. (2008), 'Decision field theory extensions for behavior modeling in dynamic environment using bayesian belief network', *Information Sciences* **178**, 2297–2314.

Lefèvre, C. and Utev, S. (1996), 'Comparing sums of exchangeable Bernoulli random variables', *J. Appl. Probab.* **33**(2), 285–310.

Leisch, F. (2008), Modelling background noise in finite mixtures of generalized linear regression models, Technical Report 37, Department of Statistics, University of Munich.

Lemmens, A. and Croux, C. (2006), 'Bagging and boosting classification trees to predict churn', *Journal of Marketing Research* **134**(1), 141–156.

Lindsay, B. and Lesperance, M. (1995), 'A review of semiparametric mixture models', *Journal of Statistical Planning and Inference* **47**, 29–99.

Lindstrom, M. and Bates, D. (1988), 'Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data', *Journal of the American Statistical Association* **83**, 1014–1022.

Liu, Y., Chawla, N., Harper, M., Shriberg, E. and Stolcke, A. (2006), 'A study in machine learning for unbalanced data for sentence boundary detection in speech.', *Computer Speech and Language* **20**(4), 468–494.

Loisel, S. (2008), 'From liquidity crisis to correlation crisis, and the need for quanls in enterprise risk management', pp. 75–77. in Risk Management : The Current Financial Crisis, Lessons Learned and Future Implications, Edited by the SOA, CAS and CIA.

Loisel, S. (2010), 'Contribution à la gestion quantitative des risques en assurance', *Habilitation Thesis, Université Lyon 1* .

Loisel, S. and Milhaud, X. (2011), 'From deterministic to stochastic surrender risk models : Impact of correlation crises on economic capital', *European Journal of Operational Research* **214**(2).

Mallows, C. (1974), 'Some comments on cp', *Technometrics* **15**, 661–675.

Margolin, B., Kim, B. and Risko, K. (1989), 'The ames salmonella/microsome mutagenicity assay : issues of inference and validation', *Journal of the American Statistical Association* **84**, 651–661.

Marshall, A. and Olkin, I. (1979), *Inequalities : Theory of Majorization and Its Applications*, Academic Press, New York.

Martinussen, T. and Scheike, T. (2006), *Dynamic Regression Models for Survival Data*, Springer.

Massart, P. (2007), *Concentration inequalities and model selection. Ecole d'été de Probabilités de Saint-Flour 2003.*, Lecture Notes in Mathematics, Springer.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models, 2nd ed.*, Chapman and Hall.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, Wiley Series In Porbability and Statistics.

McNeil, A., Frey, R. and Embrechts, P. (2005), *Quantitative Risk Management*, Princeton Series In Finance.

Milhaud, X., Gonon, M.-P. and Loisel, S. (2010), 'Les comportements de rachat en assurance vie en régime de croisière et en période de crise', *Risques* (83), 76–81.

Milhaud, X., Maume-Deschamps, V. and Loisel, S. (2011), 'Surrender triggers in life insurance : what main features affect the surrender behavior in a classical economic context ?', *Bulletin Francais d'Actuariat* **22**, ?

Mun, E.-Y., von Eye, A., Bates, M. and Vaschillo, E. (2008), 'Finding groups using model-based cluster analysis : heterogeneous emotional self-regulatory processes and heavy alcohol use risk', *Developmental Psychology* **44**, 481–495.

Nishii, R. (1988), 'Maximum likelihood principle and model selection when the true model is unspecified', *Journal of Multivariate Analysis* (27), 392–403.

Nordahl, H. A. (2008), 'Valuation of life insurance surrender and exchange options', *Insurance : Mathematics and Economics* **42**, 909–919.

Ohlson, E. and Johansson, B. (2010), *Non-Life Insurance Pricing with Generalized Linear Models*, Springer.

Oliviera-Brochado, A. and Vitorino Martins, F. (2005), Assessing the number of components in mixture models : a review. Working Paper.

Oliviera-Brochado, A. and Vitorino Martins, F. (2008), Determining the number of market segments using an experimental design. Working Paper.

Outreville, J. F. (1990), 'Whole-life insurance lapse rates and the emergency fund hypothesis', *Insurance : Mathematics and Economics* **9**, 249–255.

Pan, X., Han, C. S., Dauber, K. and Law, K. H. (2006), 'Human and social behavior in computational modeling and analysis of egress', *Automation in Construction* **15**, 448–461.

Pearson, K. (1894), 'Contributions to the theory of mathematical evolution', *Philosophical Transactions of the Royal Society of London A* **185**, 71–110.

Pesando, J. (1974), 'The interest sensibility of the flow of funds through life insurance companies : An econometric analysis', *Journal Of Finance* **Sept**, 1105–1121.

Planchet, F. and Thérond, P. (2006), *Modèles de durée : applications actuarielles*, Economica (Paris).

Raftery, A. (1994), Bayesian model selection in social research (with discussion), Technical Report 94-12, Demography Center Working, University of Washington.

Ramsay, J., Hooker, G. and Graves, S. (2009), *Functional Data Analysis with R and MATLAB*, Springer.

Ramsay, J. and Silverman, B. (2005), *Functional Data Analysis, Second Edition*, Spinger, Springer Series in Statistics.

Redner, R. (1981), 'Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions', *The Annals of Statistics* **9**(1), 225–228.

Redner, R. and Walker, H. (1984), 'Mixture densities, maximul likelihood and the em algorithm', *SIAM Review* **26**(2), 195–239.

Renshaw, A. E. and Haberman, S. (1986), 'Statistical analysis of life assurance lapses', *Journal of the Institute of Actuaries* **113**, 459–497.

Ripley, B. (1995), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

Ruiz-Gazen, A. and Villa, N. (2007), 'Storms prediction : logistic regression vs random forest for unbalanced data', *Case Studies in Business, Industry and Government Statistics* **1**(2), 91–101.

Sarstedt, M., Becker, J.-M., Ringle, C. and Schwaiger, M. (2011), 'Uncovering and treating unobserved heterogeneity with fixmix-pls : which model selection criterion provides an appropriate number of segments ?', *Schmalenbach Business Review* **63**, 34–62.

Schlattmann, P. (2003), 'Estimating the number of components in a finite mixture model : the special case of homogeneity', *Computational Statistics and Data Analysis* **41**, 441–451.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464.

Shen, W. and Xu, H. (2005), 'The valuation of unit-linked policies with or without surrender options', *Insurance : Mathematics and Economics* **36**, 79–92.

Stanton, R. (1995), 'Rational prepayment and the valuation of mortgage-backed securities', *Review of Financial* **8**, 677–708.

Teicher, H. (1963), 'Identifiability of finite mixtures', *Annals of Mathematical Statistics* **34**, 1265–1269.

Torsten, K. (2009), Valuation and hedging of participating life-insurance policies under management discretion, *in* 'Insurance : Mathematics and Economics Proceedings', Vol. 44, pp. 78–87.

Tsai, C., Kuo, W. and Chen, W.-K. (2002), 'Early surrender and the distribution of policy reserves', *Insurance : Mathematics and Economics* **31**, 429–445.

van der Vaart, A. (1998), *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge.

Vandaele, N. and Vanmaele, M. (2008), 'Explicit portfolio for unit-linked life insurance contracts with surrender option', *Journal of Computational and Applied Mathematics* .

Viquerat, S. (2010), On the efficiency of recursive evaluations in relation to risk theory applications, PhD thesis.

Wald, A. (1949), 'Note on the consistency of the maximum likelihood estimate', *The Annals of Mathematical Statistics* **20**(4), 595–601.

Wang, P. (1994), Mixed Regression Models for Discrete Data, PhD thesis, University of British Columbia, Vancouver.

Wenbin, L. (2006), 'Penalized minimum matching distance-guided em algorithm', *International Journal of Electronics and Communications* **60**, 235–239.

Wolfe, J. (1971), A monte carlo study of sampling distribution of the likelihood ratio for mixtures of multinormal distributions, Technical Report STB-72-2, San Diego : U.S. Naval Personnel and Training Research Laboratory.

Yang, Y. (2005), 'Can the strengths of aic and bic be shared ? a conflict between model identification and regression estimation', *Biometrika* **92**(4), 937–950.

Annexe A

Articles de presse

GESTION D'ACTIFS

EN BREF

RISQUE

Quand la réputation d'une société de gestion est mise en cause

« Les investisseurs ne veulent plus de boîte noire. »

FIGURE A.1 – Les Echos du 14/02/2011 - Classement de l'importance des risques pour une société de gestion. La réputation est classé en 2ème position en 2010, 1ère en 2009.

La Maîtrise des Risques le mensuel de RiskAssur

Edito

Une initiative de plus pour la résiliation des polices souscrites à titre individuel

On est loin des polices d'assurances établies dans le cadre de la loi du 13 juillet 1930 pour la durée de la compagnie et résiliable, tous les 10 ans avec un préavis de 6 mois.

La résiliation annuelle avec un préavis de trois ou d'un mois a été en grand progrès, avant de se voir substituer des polices d'une durée annuelle, avec renouvellement par tacite reconduction, sauf dénonciation avec préavis de 2 mois.

Pour permettre aux assureurs d'augmenter, hors délai de résiliation, les primes à l'occasion de l'échéance annuelle, ils accordent à l'assuré la faculté de résilier sa police, dans un délai, généralement de 2 semaines, à dater de la réception de l'avis d'échéance majoré.

Toutefois, la majoration résultant du changement de l'indice pour l'assurance habitation ou le du fait de l'application d'un malus automobile, n'entrent pas en ligne de compte.

Cependant, les dés sont pipés dès l'envoi de l'avis d'échéance parce que l'assureur ne mentionne pas l'existence d'une hausse de la prime et n'en donne encore moins le détail.

A l'assuré de comparer la prime appelée avec celle d'il y a un an et de demander le détail de l'augmentation constatée, le cas échéant.

Un autre problème s'est posé au sujet du renouvellement des polices par tacite reconduction, tranché par la loi Chatel en 2005, qui oblige les assureurs à informer l'assuré de la date d'expiration de la faculté de dénonciation du renouvellement par tacite reconduction de la police.

A défaut l'assuré dispose d'un délai de 20 jours, à dater du jour où il a eu connaissance du renouvellement de son contrat, généralement par un appel de prime.

Si d'une manière générale, cette information n'est pas donné spontanément et que l'assuré n'a pas pu dénoncer sa police dans les délais contractuels, il reçoit l'avis d'échéance qui déclenche le délai de résiliation de 20 jours.

Il se trouve que pour la ministre de l'Economie, Christine Lagarde, ces deux possibilités de résiliation ne sont pas suffisamment connues des assurés et n'accroissent pas la concurrence entre assureurs qui en est attendue.

Pour cette raison, elle souhaite introduire un délai unique de résiliation dans toutes les polices souscrites à titre individuel et de faire en sorte que les assurés en soient informés.

Pour notre part, nous pensons que le changement d'assureur pour une simple question de prime entraîne des frais inutiles et peut-être évité par une meilleure fidélisation de la clientèle par les assureurs.

Erik Kauf
Rédacteur en Chef

FIGURE A.2 – RiskAssur du 29/04/2011, changement législatif potentiel pour la résiliation.

Annexe B

Méthodes de segmentation

B.1 Méthode CART

B.1.1 Etapes de construction de l'arbre

Les différentes étapes de construction de l'arbre sont résumées dans le schéma suivant :

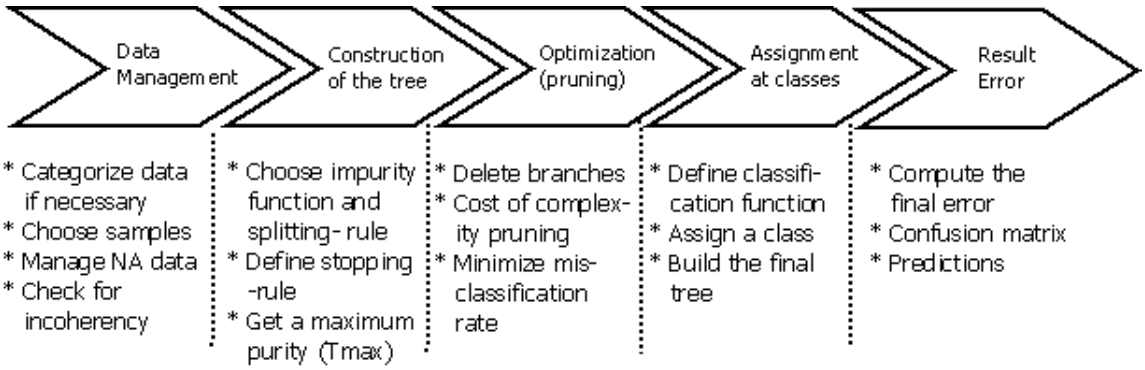


FIGURE B.1 – Etapes chronologiques de la procédure CART

Nous détaillons également par un dessin (en figure B.2) la division d'un noeud, donnant lieu à de nouvelles branches et un gain d'homogénéité.

B.1.2 Choix du paramètre de complexité

`rpart()` élague l'arbre par K validations croisées (K=10 par défaut) sur chaque arbre élagué (nous avons pris K=10). Les assurés sont choisis aléatoirement dans le processus de validations croisées, c'est pourquoi la *cptable* peut différer légèrement entre deux simulations. Sur la table B.1, *relerror* mesure l'erreur d'apprentissage et donne la qualité d'ajustement de l'arbre, *xerror* mesure le taux de mauvaise classification des 10 validations croisées et est considérée comme un meilleur estimateur de l'erreur réelle. *xstd* est l'écart type de *xerror*. L'arbre optimal minimise $err = xerror + xstd$. Si deux arbres ont la même erreur err, nous

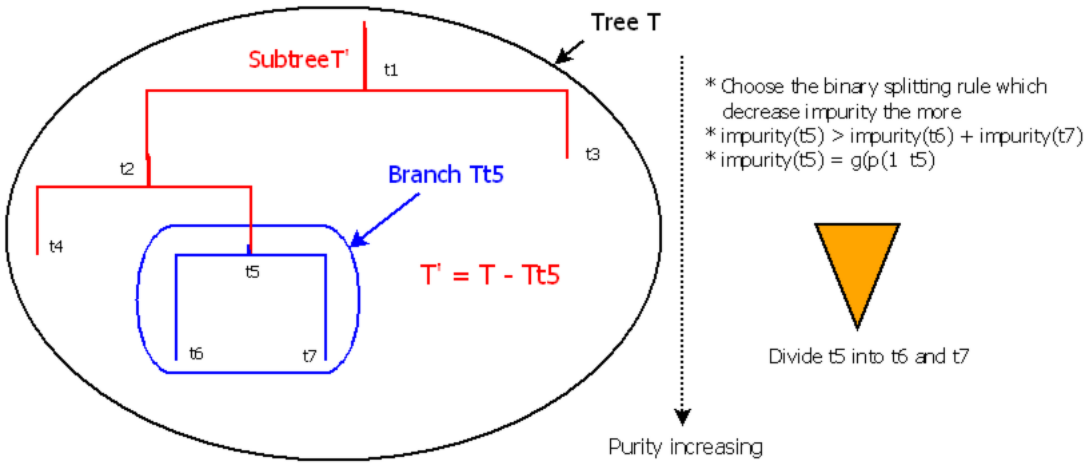


FIGURE B.2 – Construction d'un arbre binaire

choisissons le plus petit. La table B.1 permet de tracer l'erreur d'apprentissage en fonction du paramètre de complexité et de la taille de l'arbre (voir figure B.3).

Remarque B.1.1. *Quelques commentaires sur la lecture de ce tableau :*

- *le troisième arbre avec deux divisions correspond à* $\alpha \in]2.30, 3.10]$,
- `R` *normalise l'erreur, ce qui explique que l'erreur de la racine soit de 100% (1). La vraie erreur de la racine peut être obtenue en affichant l'arbre (ici elle est de 45.465%),*
- *l'arbre maximal* T_{max} *(non élagué) retourné par défaut par la fonction* `rpart()` *correspond à la dernière ligne de la cptable.*

CP	nsplit	rel error	xerror	xstd
3.3981e-01	0	1.000	1.000	0.0084
3.0539e-01	1	0.660	0.660	0.0077
5.9982e-03	2	0.354	0.361	0.0062
7.8237e-04	5	0.336	0.337	0.0061
5.2158e-04	10	0.331	0.333	0.0060
4.5638e-04	15	0.328	0.333	0.0060
3.9119e-04	19	0.326	0.333	0.0060
3.6945e-04	21	0.325	0.333	0.0060
3.2599e-04	32	0.319	0.333	0.0060
3.1295e-04	34	0.318	0.333	0.0060
2.6079e-04	39	0.317	0.332	0.0060
2.1733e-04	53	0.31360	0.334	0.0060
1.9559e-04	59	0.312	0.332	0.0060
1.8255e-04	68	0.310	0.332	0.0060
1.3040e-04	73	0.309	0.332	0.0060
1.0432e-04	82	0.308	0.332	0.0060
9.7796e-05	88	0.307	0.333	0.0060
8.6930e-05	97	0.306	0.334	0.0060
6.5198e-05	100	0.306	0.334	0.0060
4.3465e-05	117	0.305	0.337	0.0061
3.7256e-05	132	0.304	0.339	0.0061
3.2599e-05	139	0.304	0.340	0.0061
2.6079e-05	159	0.303	0.340	0.0061
0.0000e+00	174	0.303	0.341	0.0061

TABLE B.1 – Paramètres de complexité, *cptable*

B.1.3 Plus loin dans la théorie des CART

Spécification des règles binaires

Criterion 1. *Ces règles dépendent seulement d'un seuil μ et d' une variable x_l, $1 \leq l \leq d$:*
- *$x_l \leq \mu, \mu \in \mathbb{R}$ dans le cas d'une variable continue ordonnée (si nous avons m valeurs distinctes pour x_l, l'ensemble des valeurs possibles card(D) vaut M - 1) ;*
- *$x_l \in \mu$ où μ est un sous-ensemble de $\{\mu_1, \mu_2, ..., \mu_M\}$ et les μ_m sont les modalités de la variable catégorielle (dans ce cas le cardinal du sous-ensemble D des règles binaires vaut $2^{M-1} - 1$).*

Qu'est ce qu'une fonction d'impureté ?

Définition. *Une fonction d'impureté est une fonction réelle g définie sur un ensemble de probabilités discrètes d'un ensemble fini :*

$$g : (p_1, p_2, ..., p_J) \rightarrow g(p_1, p_2, ..., p_J),$$

symétrique en p_1, p_2, ..., p_J et qui vérifie :

1. *le maximum de g est à l'équiprobabilité : $argmax\ g(p_1, p_2, ..., p_J) = \left(\frac{1}{J}, \frac{1}{J}, ..., \frac{1}{J}\right)$,*
2. *le minimum de g est obtenu par les "dirac" : $argmin\ g(p_1, p_2, ..., p_J) \in \{e_1, ..., e_J\}$, où e_j est le j^{eme} élément dans la base canonique de$\mathbb{R}^J$.*

Différentes fonctions d'impureté

D'habitude nous considérons les fonctions suivantes (qui satisfont le critère de concavité) :
- impur(t) = - $\sum_{j=1}^{J} p(j|t)\ \ln(p(j|t))$;

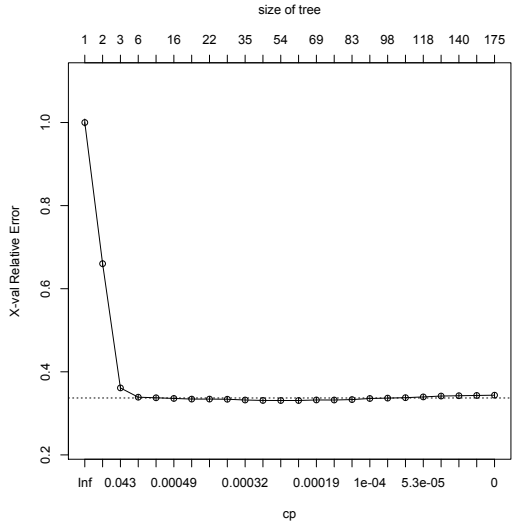


FIGURE B.3 – L'estimateur du taux de mauvaise classification de l'arbre optimal en fonction du paramètre de complexité cp (ou α). T_{max} contient ici 175 feuilles et correspond à $cp = 0$. Remarquez la forme avec un forte pente négative suivie d'un plateau, puis une légère remontée de l'erreur.

- impur(t) = $\sum_{j\neq k} p(j|t)\ p(k|t)$ (index de Gini)

Remarque B.1.2. *Dans une approche variance,*

- *l'index de Gini est aussi égal à* $1 - \sum_j p_j^2$ *;*
- *nous utilisons également la twoing rule : choisir* Δ *qui maximise* $\frac{p_L p_R}{4}\Big[\sum_j |p(j|t_L) - p(j|t_R)|\Big]^2$ *;*
- *dans un problème avec une réponse binaire, l'index de Gini se réduit à* $impur(t) = 2p(1|t)p(2|t)$.

Commentaires sur l'erreur de prévision

Nous pouvons écrire de manière formelle l'expression de la portion d'observations mal classées par la fonction *class* suivant l'estimation choisie de l'erreur de prévision :

– l'estimate "resubstitution" :

$$\hat{\tau}(class) = \frac{1}{N}\sum_{(x_n,j_n)\in\epsilon} \mathbb{1}\{class(x_n,\epsilon) \neq j_n\} \tag{B.1}$$

– l'estimation par échantillon de validation : quasiment comme dans $(B.1)$:

$$\hat{\tau}^{ts}(class) = \frac{1}{N'}\sum_{(x_n,j_n)\in W} \mathbb{1}\{class(x_n,\epsilon) \neq j_n\} \tag{B.2}$$

– l'estimation par validations croisées :

$$\hat{\tau}^{cv}(class) = \frac{1}{N}\sum_{k=1}^{K}\sum_{(x_n,j_n)\in\epsilon_k} \mathbb{1}\{class(x_n,\epsilon^k) \neq j_n\} \tag{B.3}$$

Remarquons aussi que

$$\begin{aligned}\mathbb{E}[\hat{\tau}(class)] &= \mathbb{E}\left[\tfrac{1}{N}\sum_{(x_n,j_n)\in\epsilon} \mathbb{1}\{class(x_n,\epsilon) \neq j_n\}\right] \\ &= \frac{1}{N}\sum_{(x_n,j_n)\in\epsilon} \mathbb{E}[\mathbb{1}\{class(x_n,\epsilon) \neq j_n\}] \\ &= P(class(X,\epsilon) \neq Y) = \tau(class).\end{aligned}$$

et que tous les estimateurs présentés ci-dessus sont non-biaisés :
$\mathbb{E}[\hat{\tau}(class)] = \mathbb{E}[\hat{\tau}^{cv}(class)] = \mathbb{E}[\hat{\tau}^{ts}(class)]$
L'erreur de prévision et le taux de mauvaise classification sont deux concepts différents. L'erreur de mauvaise classification est l'erreur dans les noeuds de l'arbre alors que l'erreur de prévision est liée à la classification finale de la variable d'intérêt et est calculée une fois l'arbre construit.

Par défaut, R calcule un estimateur par validations croisées de l'erreur d'apprentissage. Ce sont les résultats du tableau des paramètres de complexité. Toutefois cette procédure de validations croisées ne correspond pas à la fameuse technique de validations croisées dans la théorie du reéchantillonage. La première calcule l'arbre optimal pour une taille donnée en minimisant l'erreur d'apprentissage alors que la dernière permet d'obtenir une estimation plus réaliste de l'erreur de prévision mais ne traite pas le problème qui est de trouver un arbre optimal.

Pénalisation de la mauvaise classification

Les méthodes en structure d'arbre ont subi beaucoup de critiques à cause de la taille des arbres finaux sélectionnés en pratique et de l'usage de l'estimation par resubstitution (cf 1.1.1). Le coût de mal classer une observation n'est souvent pas le même pour toutes les classes dans les applications, d'où l'idée de pénaliser la mauvaise classification d'une observation (par rapport à sa classe observée, apprentissage supervisé) par un facteur positif.

Définition. *Le coût de mauvais classement d'une observation est défini par*

$$\Gamma : C \times C \to \mathbb{R}_+ \text{ , such that}$$
$$\Gamma(i|j) \geq 0 \text{ and } \Gamma(i|i) = 0$$

Définissons ainsi

- la probabilité de mal classer une observation par $P_{class}(i|j) = P(class(x, \epsilon) = i \mid j)$ (la fonction *class* classe x dans la classe i au lieu de la classe j),
- $\tau_{class}(j) = \sum_i \Gamma(i|j) P_{class}(i|j)$: le coût moyen de mauvaise classification.

Nous obtenons $\tau_{class} = \tau(T)$ et

$$\tau(T) = \sum_j \pi(j)\tau_{class}(j) = \frac{1}{N} \sum_j N_j \tau_{class}(j)$$

Ghattas (1999) définit dans ce contexte la fonction de classification pénalisée d'assignation d'une classe à un noeud terminal t :

$$class(x, \epsilon) = \underset{i \in C}{argmin} \sum_{j \in C} \Gamma(i|j)\ p(j|t) \tag{B.4}$$

D'après $(B.4)$, l'estimation du taux de mauvaise classification est maintenant

$$r(t) = \min_{i \in C} \sum_{j \in C} \Gamma(i|j)\ p(j|t)$$

Sachant que $\tau(t) = r(t)p(t)$, le taux de mauvaise classification par substitution de l'arbre T est donné par

$$\hat{\tau}(T) = \sum_{t \in \tilde{T}} \hat{\tau}(t). \tag{B.5}$$

Corollaire 3. *L'estimateur $\hat{\tau}(T)$ du taux de mauvaise classification de l'arbre s'abaissent à chaque division, et ce quelle que soit la division. Ainsi, si nous notons T_s l'arbre obtenu par division de T à une feuille, nous avons*

$$\hat{\tau}(T_s) \leq \hat{\tau}(T) \tag{B.6}$$

Soient t_L et t_R les descendants du noeud t dans l'arbre T_s.
D'après $(B.5)$ et $(B.6)$,

$$\begin{aligned} \sum_{t \in \tilde{T}_s} \hat{\tau}(t) &\leq \sum_{t \in \tilde{T}} \hat{\tau}(t) \\ \sum_{t \in \tilde{T}} \hat{\tau}(t) - \hat{\tau}(t) + \hat{\tau}(t_L) + \hat{\tau}(t_R) &\leq \sum_{t \in \tilde{T}} \hat{\tau}(t) \\ \hat{\tau}(t_L) + \hat{\tau}(t_R) &\leq \hat{\tau}(t) \end{aligned} \tag{B.7}$$

Elagage de l'arbre

Le problème d'un arbre final trop complexe qui "overfit" les données peut être résolu assez facilement. La solution consiste à appliquer ces deux idées plutôt que d'essayer de trouver la bonne règle d'arrêt des divisions (qui n'est pas la bonne approche) :

1. ne pas arrêter la construction de l'arbre et obtenir le plus grand arbre T_{max} ; puis l'élaguer par étape jusqu'à la racine (le critère d'élagage et de recombinaison de l'arbre est beaucoup plus important que le critère de division) ;
2. utiliser de meilleurs estimateurs du vrai taux de mauvaise classification pour sélectionner l'arbre de bonne taille parmi les sous-arbres élagués (utiliser les validations croisées ou l'échantillon témoin/test pour cela).

L'idée est de chercher des sous-arbres de T_{max} avec un taux de mauvaise classification minimum. Elaguer une branche T^t d'un arbre T signifie supprimer tous les descendants du noeud t dans T. L'arbre élagué résultant est noté $T^{'} = T - T^t$, et $T^{'} < T$.
D'après ($B.7$) nous avons

$$\hat{\tau}(t) \geq \hat{\tau}(T^t). \tag{B.8}$$

T_{max} contient tellement de noeuds qu'un nombre incalculable de manières d'élaguer l'arbre jusqu'à la racine existe, ce qui nous amène à définir un critère pour la sélection de la procédure d'élagage qui donne le "meilleur" sous-arbre. Le critère naturel de comparaison pour les arbres de même taille est l'erreur de mauvaise classification : l'algorithme d'élagage commence par T_{max} et élague progressivement jusqu'à obtenir la racine de telle manière qu'à chaque étape de l'élagage le taux de mauvaise classification soit aussi faible que possible. Ce travail fournit une suite d'arbres de plus en plus petits : $T_{max} > T_1 > T_2 > ... > T_{root}$. ($T_{root}$ est le noeud racine, sans aucune division).
D'après ($B.6$), remarquons que : $T_1 < T_{max} \Rightarrow \hat{\tau}(T_{max}) \leq \hat{\tau}(T_1)$. L'erreur de l'arbre maximal est toujours inférieure ou égale à l'erreur de l'arbre élagué, le but étant de diminuer le nombre de feuilles de T_{max}. Une idée naturelle consiste à pénaliser un grand nombre de feuilles dans l'arbre final, par l'introduction dans l'erreur d'un terme de coût de complexité. Le nouveau taux de mauvaise classification ou *cost-complexity measure* devient :

$$\hat{\tau}_\alpha(T) = \hat{\tau}(T) + \underbrace{\alpha\, Card(\tilde{T})}_{\text{complexity term}}, \text{ where } \alpha > 0, \tag{B.9}$$

où $Card(\tilde{T})$ est le nombre de noeuds terminaux de T.
En fait nous désirons juste trouver le sous-arbre $T(\alpha) \leq T_{max}$ qui minimise $\tau_\alpha(T)$ pour chaque α :

$$\tau_\alpha(T(\alpha)) = \min_{T \leq T_{max}} \tau_\alpha(T) \tag{B.10}$$

Pour les questions d'existence et d'unicité de l'arbre $T(\alpha)$, voir l'ouvrage de Breiman (1984). La valeur α est clairement liée à la taille de l'arbre final élagué ; si α est petit alors la pénalité associée à un grand nombre de feuilles est petite et l'arbre $T(\alpha)$ sera grand. Les cas extrêmes sont :

- $\alpha = 0$: chaque feuille contient une seule observation (T_{max} très grand). Toutes les observations sont bien classées et $\tau(T_{max}) = 0$. T_{max} minimise $\tau_0(T)$;
- $\alpha \to +\infty$: la pénalité pour le nombre de feuilles est grande et le sous-arbre qui minimise l'erreur sera la racine !

Algorithme 1. *Pour connaître les branches à élaguer et le α optimal associé,*

1. *Soient les feuilles t_L et t_R les descendants immédiats du noeud parent t ; en commençant par T_{max}, nous cherchons la division qui n'a pas donné de diminution de l'erreur, i.e. pour laquelle $\hat{\tau}(t) = \hat{\tau}(t_L) + \hat{\tau}(t_R)$ (voir (B.7)). Elaguer t_L et t_R, et recommencer de même jusqu'à ce que ce ne soit plus possible. Nous obtenons $T_1 < T$;*
2. *Pour T_1^t branche de T_1, définissons $\hat{\tau}(T_1^t) = \sum_{t \in \tilde{T}_1^t} \hat{\tau}(t)$. D'après (B.8), les noeuds non-terminaux t de l'arbre T_1 satisfont la propriété : $\hat{\tau}(t) > \hat{\tau}(T_1^t)$ (pas d'égalité grâce à la première étape).*
3. *Notons $\{t\}$ la sous-branche de T_1^t qui consiste en l'unique noeud $\{t\}$, $card(\{t\}) = 1$. Ainsi, $\hat{\tau_\alpha}(\{t\}) = \hat{\tau}(t) + \alpha$ et*

$$\hat{\tau_\alpha}(T_1^t) = \hat{\tau}(T_1^t) + \alpha \; Card(\tilde{T}_1^t) \tag{B.11}$$

Nous avons vu que $\hat{\tau}(T_1^t) < \hat{\tau}(\{t\})$, mais l'introduction d'un terme de complexité fait que cette inégalité avec $\hat{\tau_\alpha}$ n'est pas toujours respectée. Tant que $\hat{\tau_\alpha}(T_1^t) < \hat{\tau_\alpha}(\{t\})$ il est inutile d'élaguer, mais il existe un seuil α_c tel que $\hat{\tau_{\alpha_c}}(T_1^t) = \hat{\tau_{\alpha_c}}(\{t\})$. On a donc

$$\begin{aligned} \hat{\tau}(T_1^t) + \alpha_c \; Card(\tilde{T}_1^t) &= \hat{\tau}(t) + \alpha_c \\ \alpha_c &= \frac{\hat{\tau}(t) - \hat{\tau}(T_1^t)}{Card(\tilde{T}_1^t) - 1} \end{aligned}$$

Tant que $\alpha < \alpha_c$, il n'est pas nécessaire d'élaguer l'arbre au noeud t, mais dès que $\alpha = \alpha_c$ l'élagage de cette sous-branche est intéressante car l'erreur est équivalente et l'arbre est plus simple ;

4. *Faire ceci pour tous les noeuds t de T_1 et choisir le noeud t dans T_1 qui minimise la quantité α_c. Soit $\alpha_1 = \alpha_c$. En élaguant T_1 au noeud t, nous obtenons $T_2 = T_1 - T_1^t$. Répéter 3. et 4. récursivement avec T_2, obtenez α_2 et ainsi de suite jusqu'à la racine.*

Au final, nous obtenons par construction (avec les cas extrêmes) une suite $\alpha_1 < \alpha_2 < ... < \alpha_{root}$ qui correspondant aux arbres élagués $T_1 > T_2 > ... > T_{root}$. T_{root} est juste le noeud racine. Pour définir l'arbre optimal de cette suite, $(B.10)$ nous dit que le meilleur arbre élagué est celui avec le taux de mauvaise classification minimum.

B.2 La régression logistique

B.2.1 Résultats numériques de l'analyse statique

Les coefficients de régression, leur écart-type, la confiance que nous pouvons avoir dans l'estimation de ces coefficients et leur effet sont disponibles dans la table B.2. Les coefficients de régression de l'analyse dynamique du début du chapitre... ne sont pas donnés ici car ils n'ont pas vraiment d'intérêt (l'analyse logistique dynamique avait pour but de montrer que les prévisions n'étaient pas robustes).

B.2.2 Un peu de théorie

La modélisation "logit" est pertinente car nous voulons étudier un événement binaire (le rachat), or la régression logistique analyse des données issues de loi binomiale de la forme

$Y_i \sim B(n_i, p_i)$, avec n_i le nombre d'expériences de bernoulli et p_i la probabilité de succès (rachat ici). Si nous notons Y la variable à expliquer (i.e. la décision de rachat), nous avons

$$Y = \begin{cases} 1, & \text{si l'assuré rachète sa police,} \\ 0, & \text{sinon.} \end{cases}$$

Nous pouvons maintenant adapter l'équation de régression logistique à notre contexte et nous obtenons la probabilité de rachat p :

$$\begin{aligned} logit &= \ln\left(\frac{P[Y=1|X_0=x_0,...,X_k=x_k]}{P[Y=0|X_0=x_0,...,X_k=x_k]}\right) \\ &= \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \end{aligned}$$

Finalement,

$$\left.\begin{array}{rcl} \Phi(logit(p)) &=& \Phi(\Phi^{-1}(p)) = p \\ \Phi(logit(p)) &=& \Phi(\beta_0 + \sum_{j=1}^{k} \beta_j X_j) \end{array}\right\} \Rightarrow p = \Phi(\beta_0 + \sum_{j=1}^{k} \beta_j X_j)$$

Cette écriture permet de comprendre plus facilement l'expression de la fonction de vraisemblance en 1.2.2.

B.2.3 L'algorithme de Newton-Raphson

Maxmiser la fonction de log-vraisemblance (**??**) amène à la résolution du système (k + 1) équations

$$\left\{\begin{array}{rcl} \dfrac{\partial l}{\partial \hat{\beta}_0} &=& \sum_{i=1}^{n} Y_i - \Phi(\beta_0 + \sum_{j=1}^{k} \beta_k X_k) = 0 \\ \dfrac{\partial l}{\partial \hat{\beta}_j} &=& \sum_{i=1}^{n} X_{ij}(Y_i - \Phi(\beta_0 + \sum_{j=1}^{k} \beta_k X_k)) = 0 \end{array}\right.$$

$\forall j = 1, ..., k$.
Le problème est que les solutions n'admettent pas de formules fermées et l'utilisation d'un algorithme d'optimisation est alors indispensable. Souvent l'algorithme de Newton-Raphson (basé en fait sur un développement de Taylor à l'ordre 1) est utilisé à cette fin. En SAS et en R, cet algorithme est inclus et lance le processus itératif suivant :

$$\beta^{(i+1)} = \beta^{(i)} - \left(\frac{\partial^2 \ln(L(\beta))}{\partial\beta\partial\beta'}\right)^{-1} \times \left(\frac{\partial \ln(L(\beta))}{\partial\beta}\right) \tag{B.12}$$

Lorsque la différence entre $\beta^{(i+1)}$ et $\beta^{(i)}$ est plus petite qu'un certain seuil (disons par exemple 10^{-4}), les itérations s'arrêtent et nous obtenons la solution finale.

B.2.4 Estimation de la matrice de covariance

La matrice de variance Z des coefficients $\hat{\beta}$ s'écrit

$$\begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) & \cdots & Cov(\hat{\beta}_0, \hat{\beta}_k) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & Var(\hat{\beta}_1) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_k, \hat{\beta}_0) & Cov(\hat{\beta}_k, \hat{\beta}_1) & \cdots & Var(\hat{\beta}_k) \end{pmatrix} \tag{B.13}$$

et est estimée par l'inverse de la matrice d'information de Fisher, qui vaut

$$I(\beta) = -\mathbb{E}\Big[\frac{\partial^2 \ln(L(\beta))}{\partial\beta\partial\beta'}\Big].$$

Un des caractéristiques intéressantes est que le dernier terme de cette équation est déjà calculé par l'algorithme de Newton-Raphson, ce qui permet d'estimer les coefficients de régression et leur matrice de covariance simultanément.
Comme d'habitude, l'estimateur par maximum de vraisemblance $\hat{\beta}$ converge asymptotiquement vers une loi normale de moyenne la vraie valeur de β et de variance l'inverse de la matrice de Fisher $I(\beta)$. Le terme dans l'espérance est appelé la *Hessienne* et est également utilisé dans les tests de significativité des coefficients de régression β.

B.2.5 Statistique de déviance et tests

Evaluation statistique de la régression

Pour vérifier la pertinence du modèle, nous utilisons la statistique du test du ratio de vraisemblance : l'hypothèse nulle de ce test est $\beta_1 = \beta_2 = ... = \beta_k = 0$ (H_0) ;
Et l'hypothèse alternative est "au moins un des coefficients de régression n'est pas nul" (H_1).
Soit $\mathrm{l}(\beta)$ la log-vraisemblance du modèle de régression logistique avec $k+1$ coefficients de régression, et $\mathrm{l}(\beta_0)$ la log-vraisemblance du modèle de régression logistique le plus simple (avec seulement l'ordonnée à l'origine β_0), la statistique du ratio de vraisemblance vaut

$$\Lambda = 2 \times \Big(l(\beta) - l(\beta_0)\Big). \tag{B.14}$$

Cette statistique suit une loi du χ^2_k à k degrès de liberté (d.f.).
Si la p-valeur est plus petite que le niveau de confiance que nous nous accordons, alors le modèle est globalement significatif et H_0 est rejetée.
Plus intuitivement, les statisticiens utilisent parfois le coefficient R^2 (ou coefficient de MC Fadden) : $R^2 = 1 - \dfrac{l(\beta)}{l(\beta_0)}$.
Un coefficient R^2 proche de 0 signifie que le ratio de vraisemblance est proche de 1, et donc que la log-vraisemblance du modèle complet est proche de celle du modèle le plus simple. Ainsi il n'est pas très utile d'introduire des variables explicatives supplémentaires pour la modélisation. A l'opposé, si R^2 est proche de 1, alors il y a une grande différence en termes de vraisemblance entre les deux modèles et il est intéressant de considérer le modèle complet qui est bien meilleur.

Pertinence et significativité d'une variable explicative

L'idée de ce test est de comparer la valeur du coefficient estimé β_j (associé à la variable explicative X_j) à sa variance, elle-même extraite de la matrice hessienne.
L'hypothèse nulle (H_0) est : $\beta_j = 0$; et l'hypothèse alternative (H_1) est donnée par : $\beta_j \neq 0$.
Nous utilisons la statistique de Wald qui suit une distribution du χ^2_1 pour réaliser ce test : $\Lambda = \dfrac{\hat{\beta_j}^2}{Var(\hat{\beta_j})}$.
Choisissons par exemple un seuil de confiance de 5%, et notons $\chi^2_{95\%}(1)$ le 95^{eme} percentile de la loi du chi-deux à 1 d.f. H_0 est vraie si le ratio est plus petit que ce quantile, sinon nous rejetons H_0.

Coef. (var. type)	modality : correspondance	coefficient estimate	std error	p-value	effect
β_0 (continuous)		10.63398	1.48281	7.42e-13	> 0
$\beta_{duration}$ (categorical)	1 : [0,12] (in month)	0 (reference)			nul
	2 :]12,18]	-1.31804	0.15450	$< 2e-16$	< 0
	3 :]18,24]	-2.66856	0.14016	$< 2e-16$	< 0
	4 :]24,30]	-2.75744	0.14799	$< 2e-16$	< 0
	5 :]30,36]	-3.09368	0.14294	$< 2e-16$	< 0
	6 :]36,42]	-3.54961	0.15080	$< 2e-16$	< 0
	7 :]42,48]	-3.72161	0.14980	$< 2e-16$	< 0
	8 :]48,54]	-4.10431	0.15772	$< 2e-16$	< 0
	9 : > 54	-5.49307	0.14037	$< 2e-16$	< 0
$\beta_{premium\ frequency}$ (categorical) (in month)	Monthly	0 (reference)			nul
	Bi-monthly	0.92656	0.62071	0.135504	> 0
	Quarterly	-0.03284	0.10270	0.749148	< 0
	Half-yearly	-0.22055	0.16681	0.186128	< 0
	Annual	0.43613	0.10690	4.51e-05	> 0
	Single	-0.28494	0.38155	0.455177	< 0
$\beta_{underwriting\ age}$ (categorical)	1 : [0,20[(years old)	0 (reference)			nul
	2 : [20,30[	0.28378	0.13912	0.041376	> 0
	3 : [30,40[	-0.01146	0.13663	0.933163	< 0
	4 : [40,50[	-0.26266	0.14077	0.062054	< 0
	5 : [50,60[	-0.42098	0.15136	0.005416	< 0
	6 : [60,70[	-0.66396	0.19531	0.000675	< 0
	7 : > 70	-0.75323	0.23417	0.001297	< 0
$\beta_{face\ amount}$ (categorical)	1 :	0 (reference)			nul
	2 :	-5.79014	1.46592	7.82e-05	< 0
	3 :	-7.14918	1.46631	1.08e-06	< 0
$\beta_{risk\ premium}$ (categorical)	1 :	0 (reference)			nul
	2 :	0.36060	0.11719	0.002091	> 0
	3 :	0.26300	0.14041	0.061068	> 0
$\beta_{saving\ premium}$ (categorical)	1 :	0 (reference)			nul
	2 :	0.93642	0.13099	8.74e-13	> 0
	3 :	1.32983	0.14955	$< 2e-16$	> 0
$\beta_{contract\ type}$ (categorical)	PP con PB	0 (reference)			nul
	PP sin PB	-16.79213	114.05786	0.882955	< 0
	PU con PB	-7.48389	1.51757	8.16e-07	< 0
	PU sin PB	-12.43284	1.08499	$< 2e-16$	< 0
β_{gender}	Female	0 (reference)			nul
	Male	-0.08543	0.04854	0.078401	< 0

TABLE B.2 – Estimations des coefficients de la régression logistique pour les contrats Mixtes.

Annexe C

Résultats des mélanges de Logit

C.1 Tests de validation des prédictions sur produits "Mixtos"

C.1.1 Test de Pearson

```
> normality.test(obj, "Pearson")

Pearson chi-square normality test

data:  validation.residuals
P = 0.8, p-value = 0.8495
```

C.1.2 Test de Mann-Whitney-Wilcoxon

```
> distribution.test(obj, "Wilcoxon-Mann-Whitney")

Wilcoxon rank sum test

data:  obj[[2]] and obj[[3]]
W = 50, p-value = 1
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -0.04310852  0.03889232
sample estimates:
difference in location
         -0.0001792725
```

C.2 Famille de produits Ahorro

C.2.1 Données formatées pour la modélisation

```
"issue.date";"termination.date";"line.of.business";"contract.type";"PB.guarantee";"product.no";
"1";"1999-01-01";"2008-01-01";"Saving";"Ahorro_PP_con_PB";"with_PB";"462";
"2";"1999-01-01";"2008-01-01";"Saving";"Ahorro_PP_con_PB";"with_PB";"313";
"3";"1999-01-01";"2008-01-01";"Saving";"Ahorro_PP_con_PB";"with_PB";"328";
"4";"1999-01-01";"2008-01-01";"Saving";"Ahorro_PP_con_PB";"with_PB";"462";
"5";"1999-01-01";"2004-11-01";"Saving";"Ahorro_PP_con_PB";"with_PB";"328";
```

```
"premium.frequency";"gender";"lapse.age";"underwriting.age";"underwritingAge.range";"face.amount";
"unique";"Male";NA;49;"2";9562.04;
"highly.periodic";"Female";NA;32;"1";44560.44;
"highly.periodic";"Male";NA;49;"2";23064.41;
"highly.periodic";"Female";NA;41;"2";36986.28;
"highly.periodic";"Male";26;20;"1";11706.51;

"fa.range";"risk.premium";"riskPrem.range";"saving.premium";"savingPrem.range";"duration";
"high.face.amount";0;"low.risk.premium";601.01;"middle.saving.premium";35.7802197802198;
"high.face.amount";0;"low.risk.premium";490.16;"middle.saving.premium";35.7802197802198;
"high.face.amount";0;"low.risk.premium";1667.22;"high.saving.premium";35.7802197802198;
"high.face.amount";0;"low.risk.premium";1050.42;"high.saving.premium";35.7802197802198;
"high.face.amount";0;"low.risk.premium";336.7;"low.saving.premium";23.4175824175824;

"duration.range";"lapse.reason";"lapse.bit";"surrender.bit"
"high.duration";"In force";"0";"0"
"high.duration";"In force";"0";"0"
"high.duration";"In force";"0";"0"
"high.duration";"In force";"0";"0"
"high.duration";"Surrender";"1";"1"
```

C.2.2 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille Nous représentons dans la figure C.1 l'historique de l'exposition (en vert), du taux de rachat (en noir) et du taux de chute (en rouge) du portefeuille. Nous voyons bien que les rachats font partie des chutes, mais que les chutes englobent d'autres événements ; ici par exemple un produit largement distribué est arrivé à maturité début 2005. Nous observons une forte baisse du taux de rachat dans l'année 2007, le niveau moyen de rachat ayant diminué fin 2001 pour rester relativement stable ensuite (peu de volatilité jusqu'en 2007). Le taux de rachat semble présenter des creux et des pics réguliers traduisant une certaine périodicité, sans doute dûe au cycle annuel de vente des produits (période de fête,...).

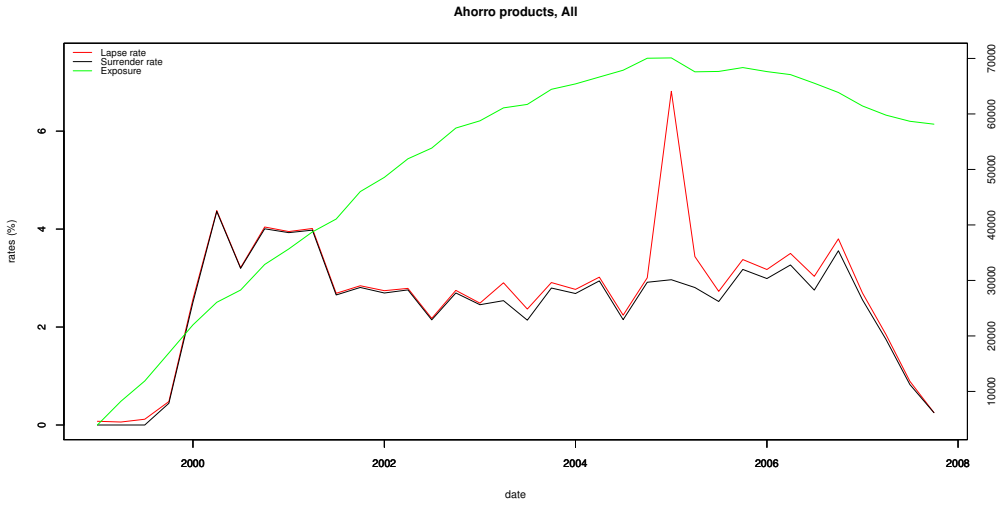
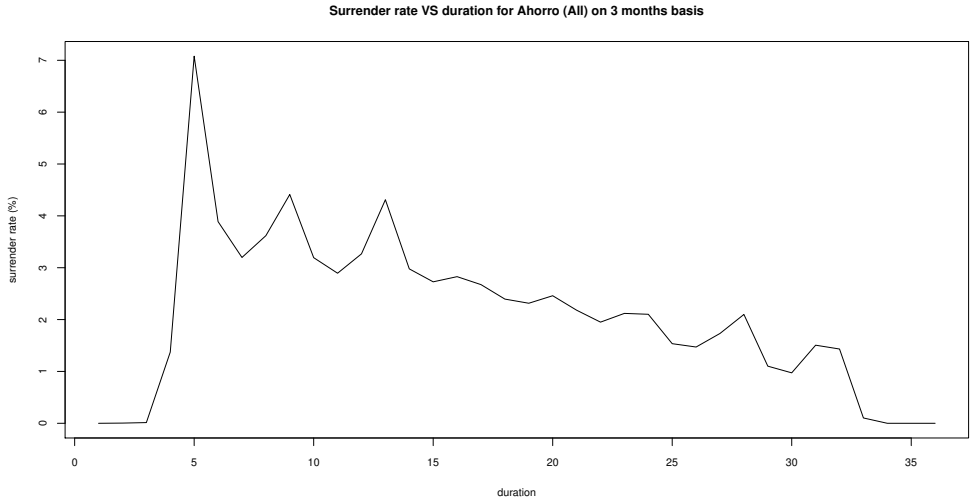


FIGURE C.1 – Exposition et taux de rachat trimestriel du portefeuille de produits Ahorro.

FIGURE C.2 – Rachat par ancienneté de contrat (en trimestre) pour les produits Ahorro.



Profil des rachats par ancienneté de contrat Le profil des rachats en fonction de l'ancienneté des contrats est un élément clef de modélisation. En effet, il s'agit ici de détecter un éventuel aspect monotone de la courbe afin de savoir si la catégorisation de la variable "ancienneté" serait judicieuse. En figure C.2, des pics périodiques apparaissent à chaque date anniversaire du contrat : cela est dû au fait que les contrats "Ahorro" en Espagne sont rachetables sans frais à chaque anniversaire de la police (aucun rachat n'est autorisé la première année, sauf cas exceptionnel). Ce profil suggère la catégorisation de cette variable continue, dans le sens où un unique coefficient de régression serait insuffisant à rendre compte de cette forme spécifique. Chaque fois qu'une catégorisation de variable continue sera effectuée dans la suite, ce sera par la méthode des quantiles : trois modalités avec chacune la même exposition

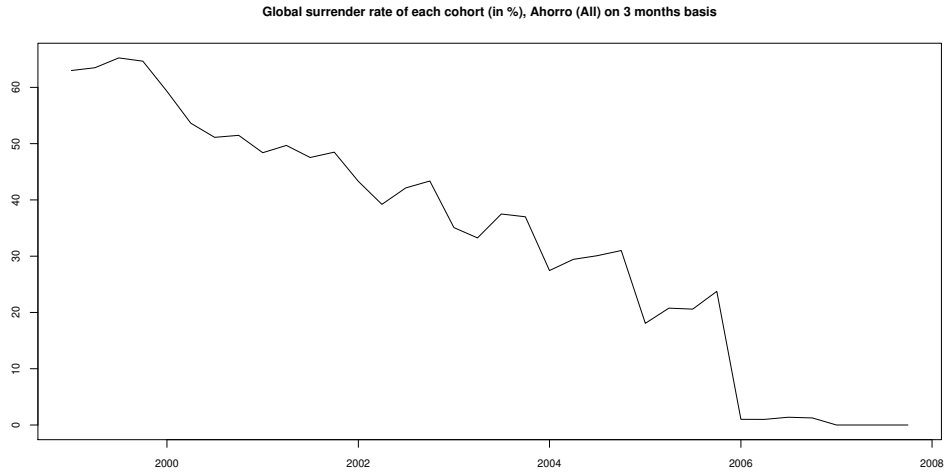
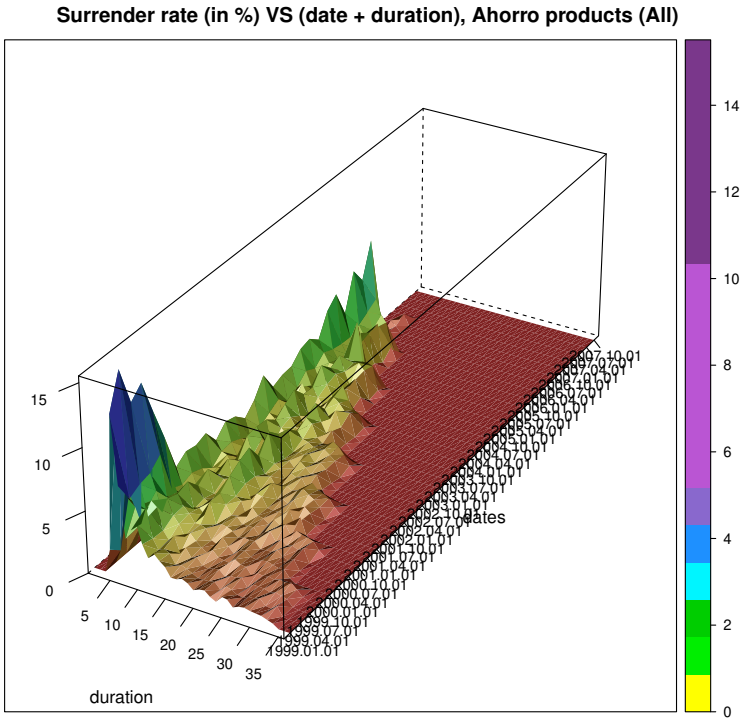


FIGURE C.3 – Pourcentage global de rachat par cohorte pour les produits Ahorro.

FIGURE C.4 – Profil 3D du taux de rachat par date et par ancienneté de contrat (par trimestre), produit Ahorro.



(l'ancienneté était catégorisée différemment au départ, aussi il est possible qu'un résultat soit basé sur cette ancienne catégorisation mais ceci est marginal). Les pics de rachat s'amenuisent avec le temps pour la simple et bonne raison que l'exposition devient moindre.

Taux de rachat par cohorte Lorsque l'on regarde le taux de rachat global par cohorte, l'idée est de voir si certaines cohortes ont globalement beaucoup plus racheté que d'autres. La figure C.3 permet de détecter une partie de l'hétérogénéité des comportements susceptible d'exister : dans ce cas précis, rien ne semble anormal (c'est pourquoi nous basculons ce graphe en annexe), le taux chutant à 0 pour les très jeunes cohortes car personne n'a encore racheté (les assurés sont dans leur première année de contrat). Nous retrouvons d'ailleurs les caractéristiques de la figure C.2 à travers les baisses périodiques observées.

Taux de rachat par date et par ancienneté de contrat La vision 3D offerte par la figure C.4 est utile dans un contexte global. Il est relativement facile d'observer des comportement anormaux en croisant les effets des dates et de l'ancienneté du contrat. Ici par exemple, nous observons que les assurés rachètent majoritairement avant leur quatrième année de contrat (du trimestre 4 au trimestre 12) quelle que soit la date ; bien qu'en 2000 énormément de personnes rachetaient dès le premier anniversaire de la police (pics bleus). C'est typiquement le mélange de ces comportements qui donne une surdispersion des données et qui empêche la modélisation par une approche simplifiée.

C.2.3 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre Au vu de la matrice de confusion sur l'échantillon de validation, le taux d'erreur de classification de l'arbre s'élève à 6,77 %. Ce bon résultat est à prendre avec précaution car la spécificité est assez mauvaise (34 %), bien que la sensibilité soit excellente (99,6 %). Nous nous servons de ce classifieur relativement précis pour en extraire les variables discriminantes dans le paragraphe suivant.

	Rachats non-observés	Rachat observés
Rachats non-prédits	1731	3363
Rachats prédits	196	47281

Importance des variables explicatives Les variables qui apparaissent comme les plus discriminantes dans la figure C.5 sont l'ancienneté de contrat, suivie de la richesse de l'assuré, de l'option de participation aux bénéfices (PB), de la prime d'épargne (corrélée à la richesse, donc nous ne considérerons qu'une des deux variables dans la modélisation), de l'ancienneté de contrat catégorisée (celle que nous considérerons par la suite pour mieux refléter le profil spécifique des rachats vu au graphe C.2) et ainsi de suite. Ce classement nous sert de base dans le choix des inputs aux futures modélisations, sachant qu'il confirme quasiment tout le temps les statistiques descriptives du taux de rachat en fonction de ces variables explicatives (nous nous abstenons donc dans le mémoire d'exposer l'ensemble des statistiques descriptives des rachats en fonction de chaque variable, ce qui serait long et fastidieux). La relation entre le taux de rachat et les variables explicatives continues n'étant que très rarement monotone, nous considérons très souvent dans la suite le classement par importance des variables catégorisées. Les trois principales que nous retenons ici sont donc l'option de PB, l'ancienneté de contrat et la fréquence de la prime. La saisonnalité n'apparait pas car elle ne fait pas partie des variables

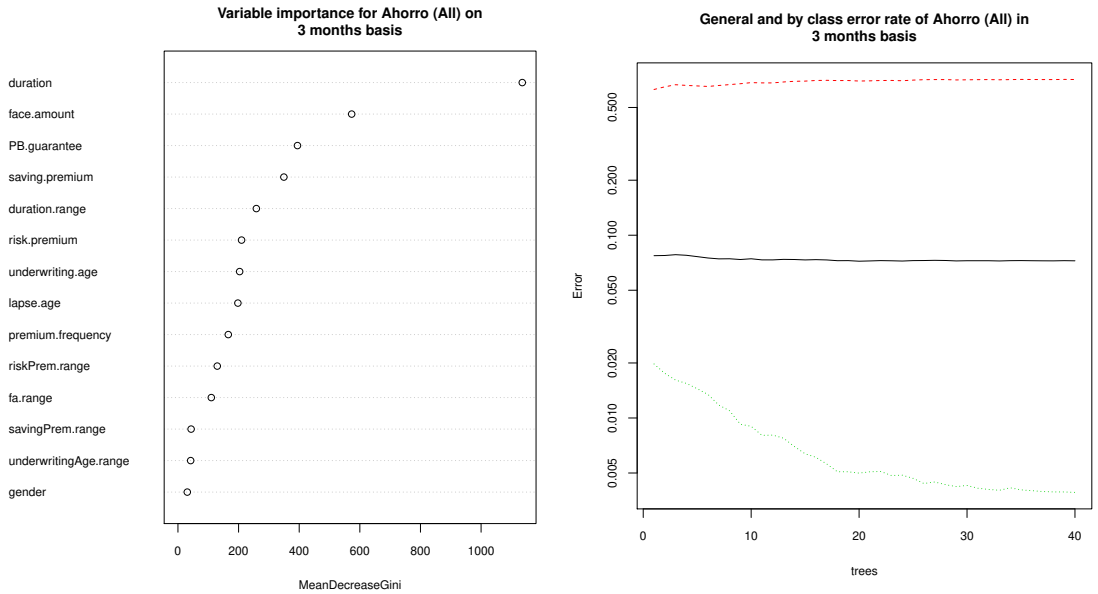


FIGURE C.5 – Importance des variables explicatives, produit Ahorro.

en input de la méthode CART mais nous la prendrons toujours en compte, hormis avec des produits pour lesquels cet effet semble peu logique (produits structurés par exemple).

C.2.4 Boxplot des coefficients du modèle

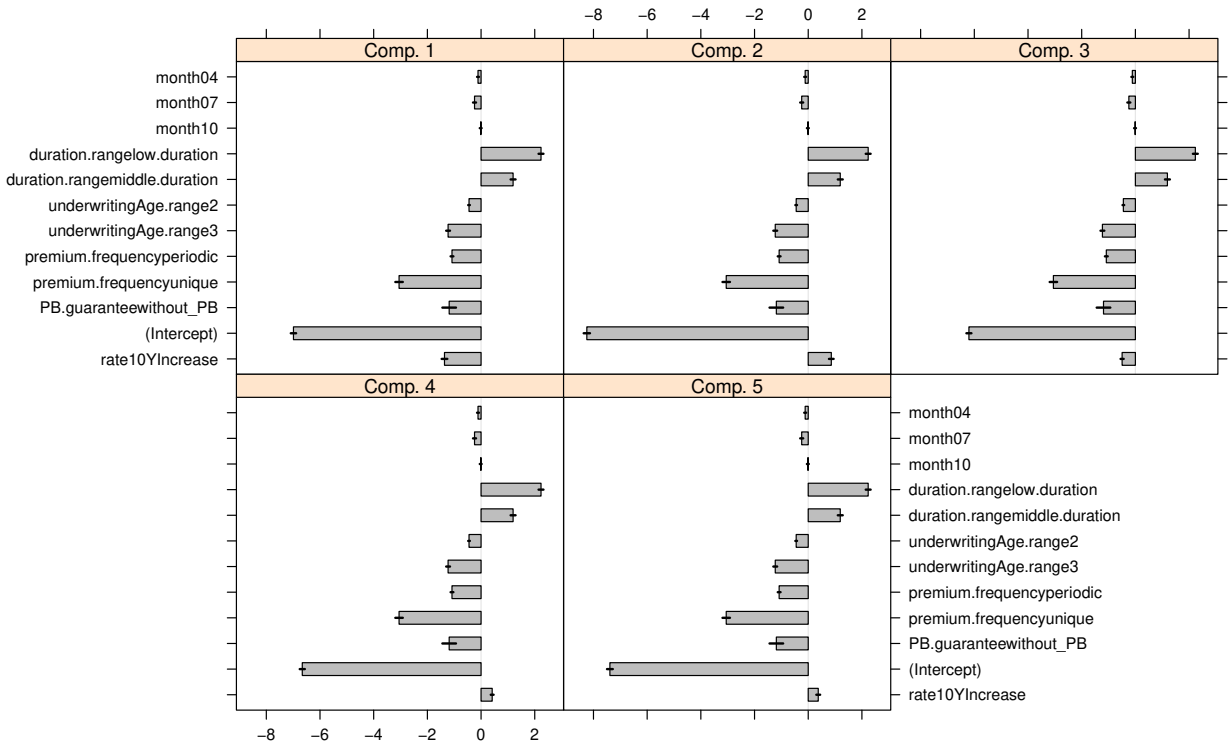


FIGURE C.6 – Coefficients de régression des composantes, produits Ahorro.

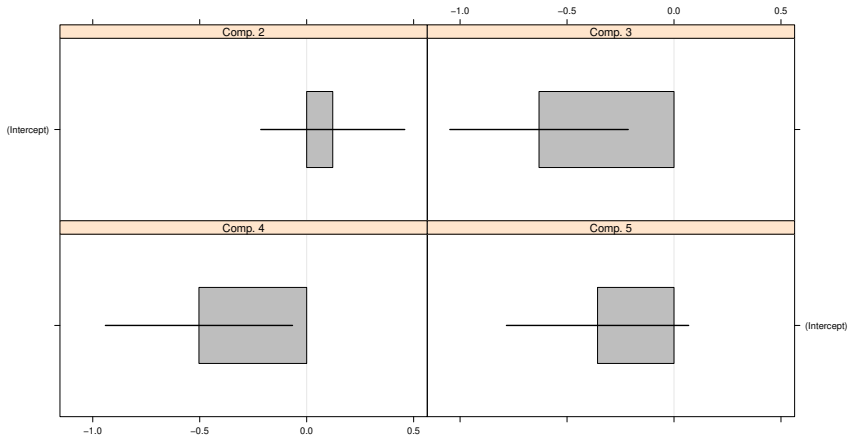
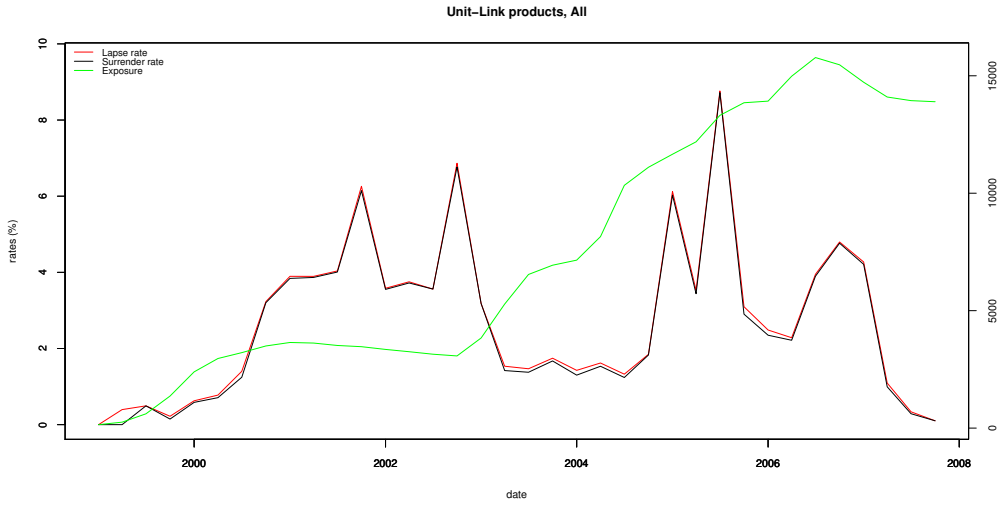


FIGURE C.7 – Coefficients de régression estimés des poids des composantes, produits Ahorro.

C.3 Famille de produits Unit-Link

C.3.1 Analyse descriptive

FIGURE C.8 – Exposition et taux de rachat trimestriel du portefeuille de produits UC.

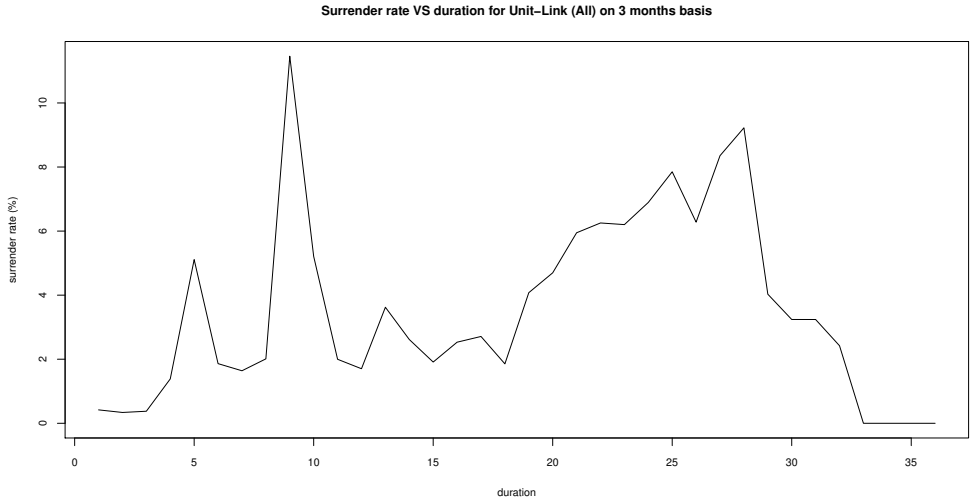


Evolution de l'exposition et du taux de rachat du portefeuille D'après le graphique C.8, le niveau moyen du taux de rachat change régulièrement et fait preuve d'une volatilité assez importante. Les changements de niveau sont brusques et de forte amplitude, l'exposition se stabilise lors des périodes de crise (2000 et 2007), traduisant la réticence des agents à souscrire de nouvelles affaires sur ce type de produit en environnement fortement incertain. Nous retrouvons une forme de périodicité à travers les petites hausses et baisses régulières du taux, mais qui n'est pas non plus forcément évidente. Comme pour les produits de pure épargne, le taux de rachat semble s'effondrer à des niveaux anormalement bas en 2007.

Profil des rachats par ancienneté de contrat Le graphe C.9 ne montre aucune relation monotone entre le taux de rachat et l'ancienneté de contrat, et met en évidence un fort pic de rachat à la fin de la deuxième année de contrat (8 et 9 trimestres d'ancienneté) mais ce comportement semble marginal et n'a pas d'explication rationnelle (de type frais spécifique pour un rachat à tel ou tel moment). L'allure de cette courbe nous fait pencher encore une fois pour une catégorisation de la variable "ancienneté" dans la modélisation, qui a l'avantage de mieux rendre compte de cette forme non-monotone mais qui a l'inconvénient d'augmenter le nombre de paramètres à estimer.

Taux de rachat par cohorte Le taux de rachat global par cohorte du graphique C.10 prouve une forte hétérogénéité des comportements de rachat en fonction de la date d'entrée en portefeuille. Le fait que les vieilles cohortes aient un taux élevé (environ 80 %) est tout à fait normal compte tenu de leur ancienneté, mais le pic de taux à presque 90 % pour les cohortes de fin 2003 paraît étonnant. Le couplage des graphes C.9 et C.10 pourrait d'ailleurs expliquer le pic observé fin 2005 sur le graphique C.8 si le pic de rachat autour de 2 ans d'ancienneté de contrat est dû aux cohortes entrées fin 2003. Ce pic ne semble pas dû aux marchés financiers

FIGURE C.9 – Rachat par ancienneté de contrat (en trimestre) pour les produits UC.



(car ceux-ci sont en nette hausse en 2005), mais plutôt à une nouvelle vague de produits UC que les agents d'AXA ont déployé en faisant racheter par là-même leur ancien contrat UC aux assurés.

Taux de rachat par date et par ancienneté de contrat Nous retrouvons dans le graphe C.11 la confirmation que ce sont bien les cohortes qui ont souscrit fin 2003 qui rachètent fin 2005, mais pas pour des conditions de marché défavorables. Le pic de rachat en 2002 semble par contre correspondre à des comportements de rachat rationnels (dûs à une baisse du marché), ce qui induit une forte hétérogénéité pour la modélisation car les sources de rachat varient en plus de la volatilité des marchés.

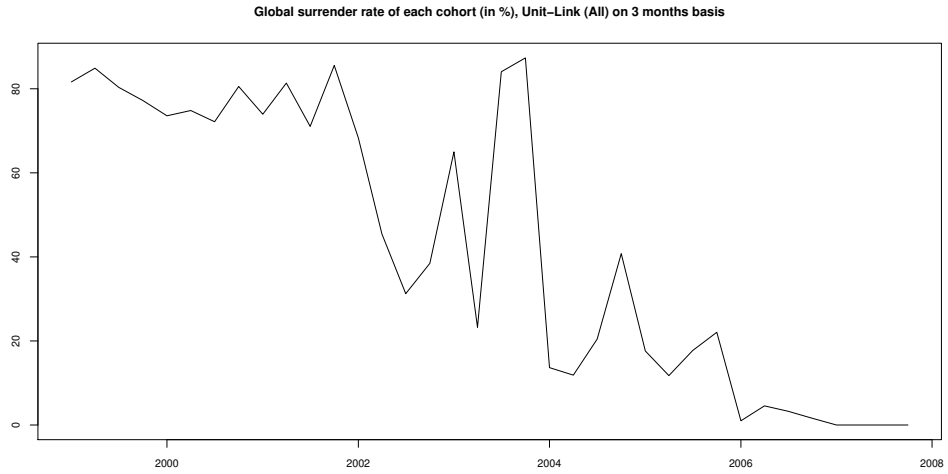
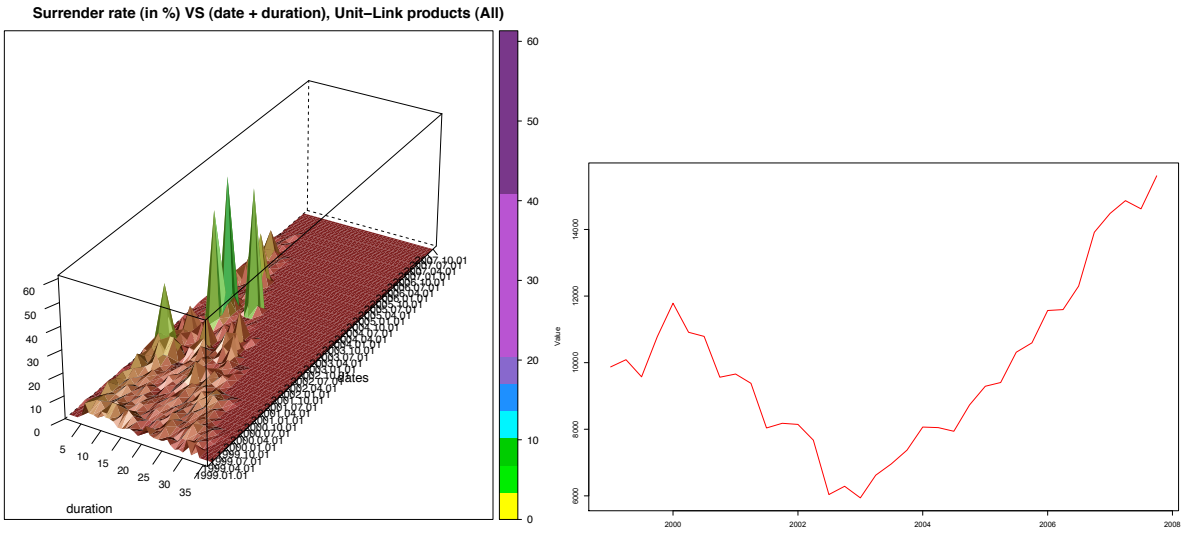


FIGURE C.10 – Pourcentage global de rachat par cohorte pour les produits UC.

FIGURE C.11 – A gauche : profil 3D du taux de rachat des UC par date et par ancienneté de contrat (par trimestre). A droite : évolution trimestrielle en valeur de l'indice boursier espagnol Ibex 35.



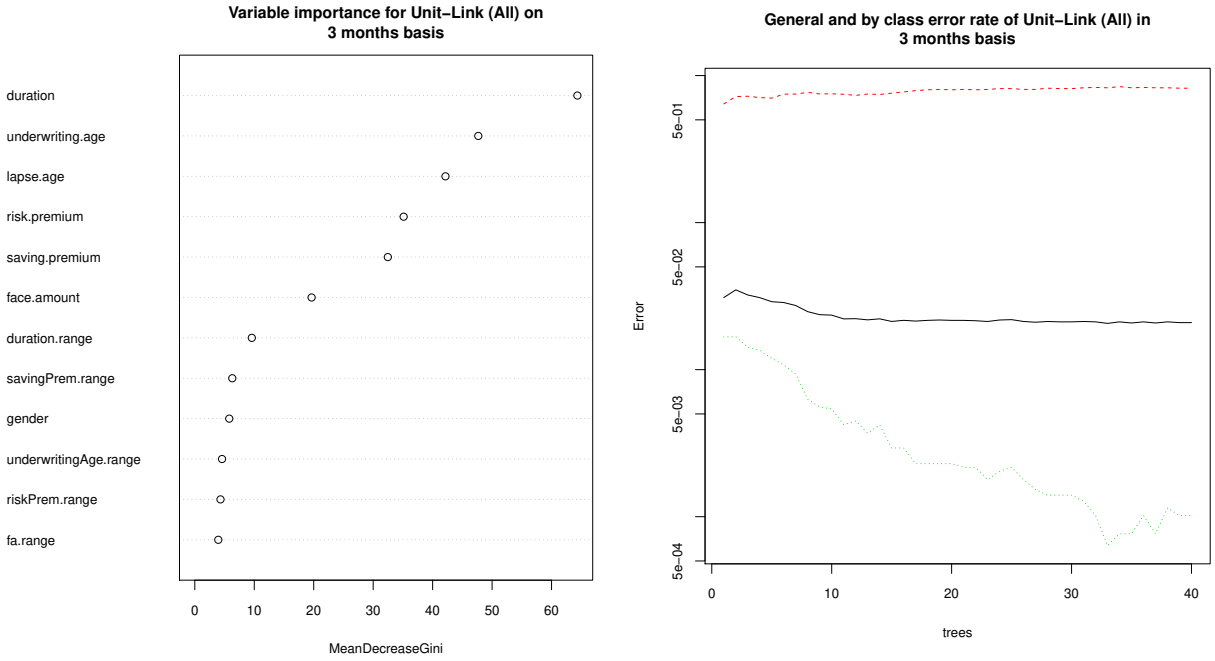
C.3.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre L'erreur de classification (sur l'échantillon de validation) s'élève à 4,9 %, dont des indices de performance aux résultats exceptionnels de 92,6 % (pour la sensibilité) et de 96,5 % (pour la spécificité). Les comportements de rachat semblent donc très bien prédits par le modèle grâce aux variables dont nous disposons. Nous verrons dans la modélisation par régression logistique dynamique que les facteurs conjoncturels font voler en éclat ce constat.

	Rachats non-observés	Rachat observés
Rachats non-prédits	4541	164
Rachats prédits	193	2410

Importance des variables explicatives Le classement de l'importance des variables explicatives disponible en figure C.12 fait la part belle à l'ancienneté de contrat, l'âge de souscription, l'âge du rachat (corrélé à l'âge de souscription et l'ancienneté), la prime de risque et la prime d'épargne. Nous ne souhaitons sélectionner que les deux ou trois variables les plus importantes dans la modélisation pour minimiser la complexité du modèle final, ce qui donne (en considérant les variables catégorisées) l'ancienneté de contrat et la prime de risque (nous aurions pu considérer la tranche d'âge mais les statistiques desciptives nous montrent qu'en réalité cette variable n'est pas si discriminante).

FIGURE C.12 – Importance des variables explicatives, produit UC.



C.3.3 Boxplot des coefficients du modèle

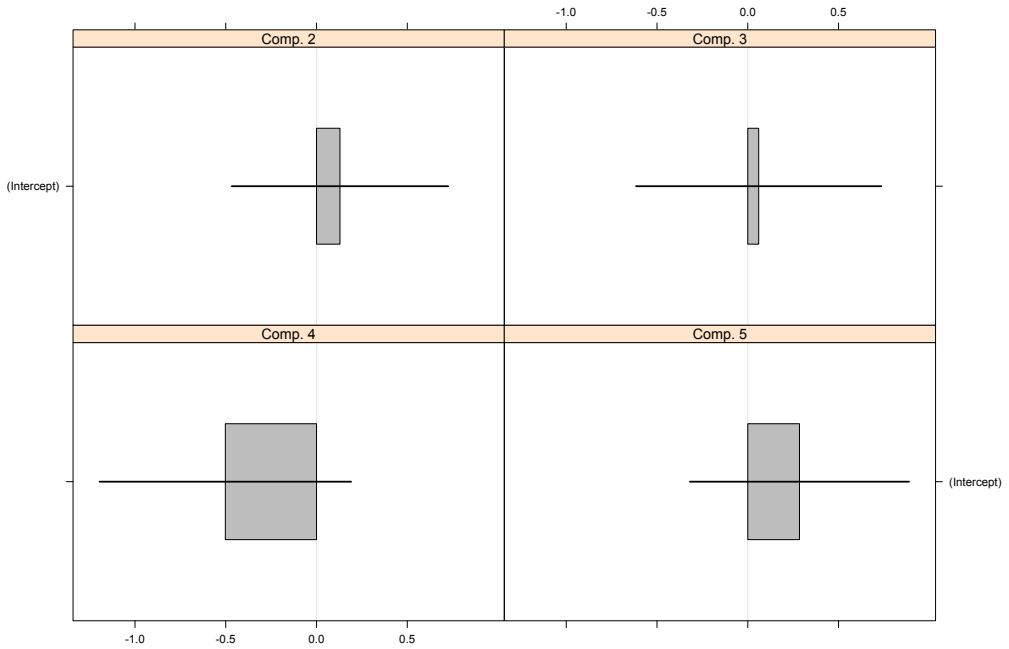
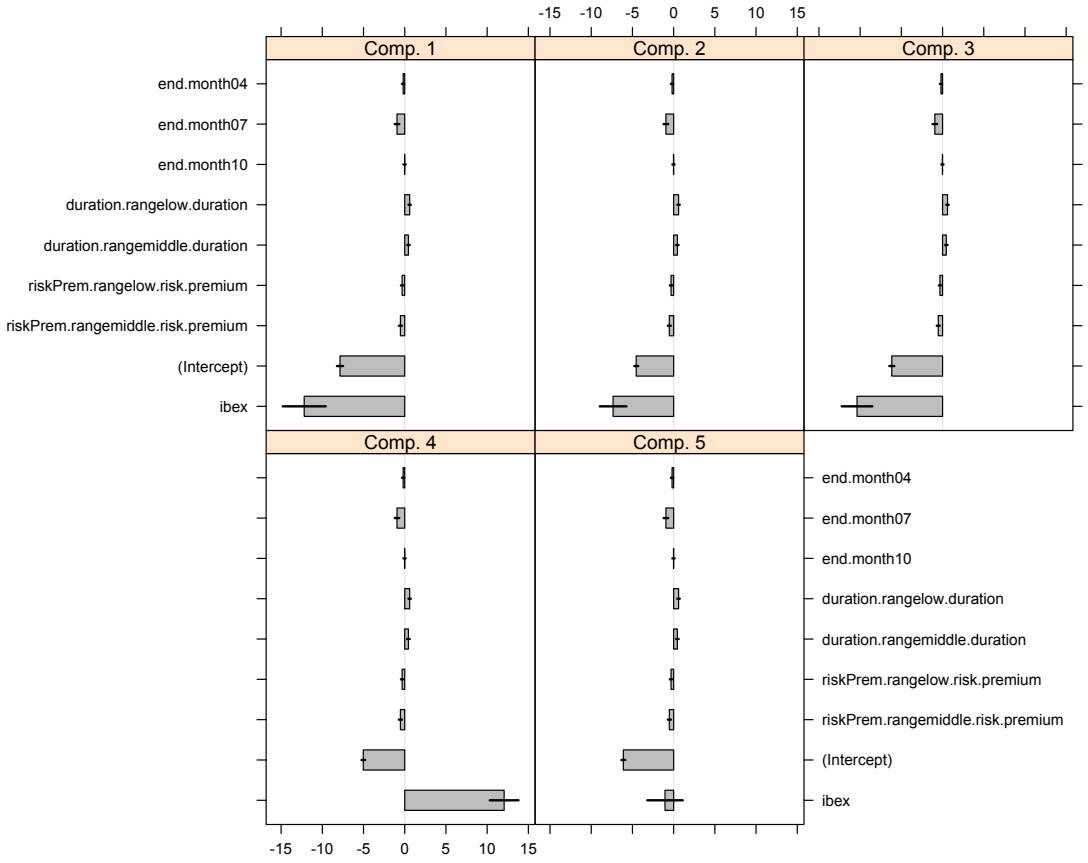


FIGURE C.13 – Coefficients de régression estimés des poids des composantes, produits UC.

FIGURE C.14 – Coefficients de régression des composantes, produits UC.



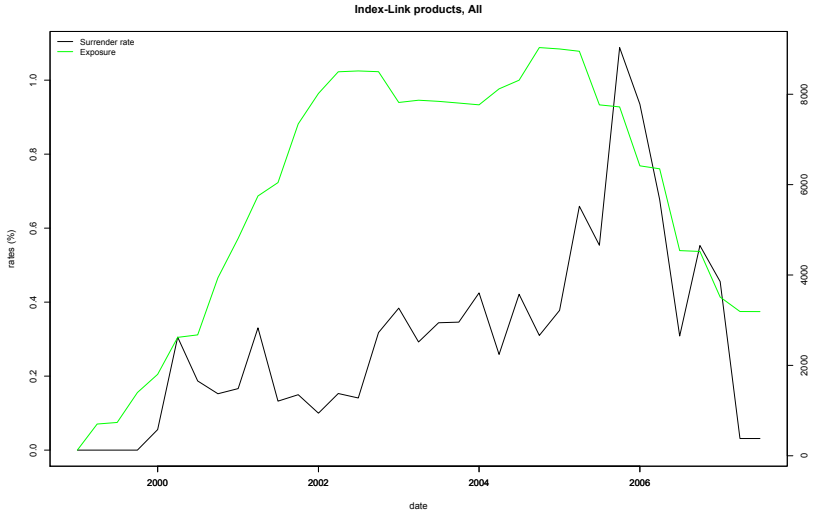
C.4 Famille de produits Index-Link

C.4.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille La première remarque que nous pouvons formuler avec le graphe C.15 est que le taux de rachat est globalement très faible, écrasé par le taux de chute dans le graphe d'origine (c'est la raison pour laquelle nous ne traçons pas le taux de chute ici). La deuxième constatation est que l'exposition d'AXA Seguros à ce type de produit est moindre, en très forte baisse depuis 2005, et que nous n'observons pas clairement de saisonnalité. Le manque de diversification du support d'investissement joue certainement un rôle dans cette statistique, les vendeurs et les souscripteurs connaissant de plus en plus l'importance de cette diversification pour diminuer le risque global, d'où une formule peu attractive. Il n'y a effectivement plus de nouvelle souscription sur ce type de produits depuis début 2005.

Profil des rachats par ancienneté de contrat et taux de rachat par cohorte En un certain sens, le profil des rachats en fonction de l'ancienneté des contrats du graphe C.16

FIGURE C.15 – Exposition et taux de rachat trimestriel du portefeuille de produits Index-Link.



rappelle celui constaté sur les produits en UC. Il en est de même concernant les taux de rachat globaux par cohorte (c'est pourquoi nous regroupons ici ces deux graphiques). Une forme erratique, imprévisible, non-monotone. La différence majeure concerne le comportement des cohortes qui semble davantage lié à l'indice Ibex 35, qui rappelons le s'effondre entre 2000 et 2002, provoquant un niveau moyen de rachat des cohortes supérieur visible sur cette période.

Taux de rachat par date et par ancienneté de contrat La mise en évidence d'une forte hétérogénéité par le graphique C.17 vient confirmer l'ensemble des observations faites précédemment. Il ne se dégage pas de profil précis en fonction de l'ancienneté de contrat, mais la date calendaire (ici entre 2000 et 2002) et donc le contexte économique joue clairement un rôle. La vague de rachat de fin 2005 était déjà observée sur les produits en UC, et ne correspond toujours pas à la chute de l'indice boursier. Nous évoquons des politiques de vente pour expliquer ce fort pic (il semblerait qu'il y ait eu des problèmes avec les réseaux de distribution à cette période mais l'information n'est pas disponible dans la base de données).

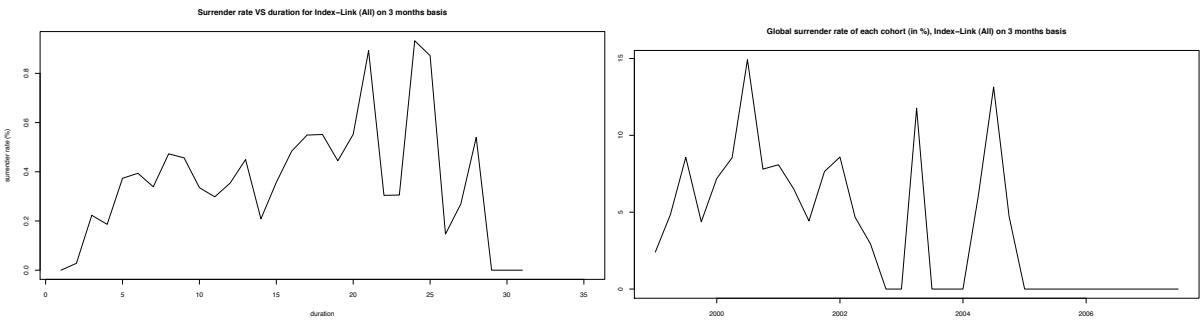
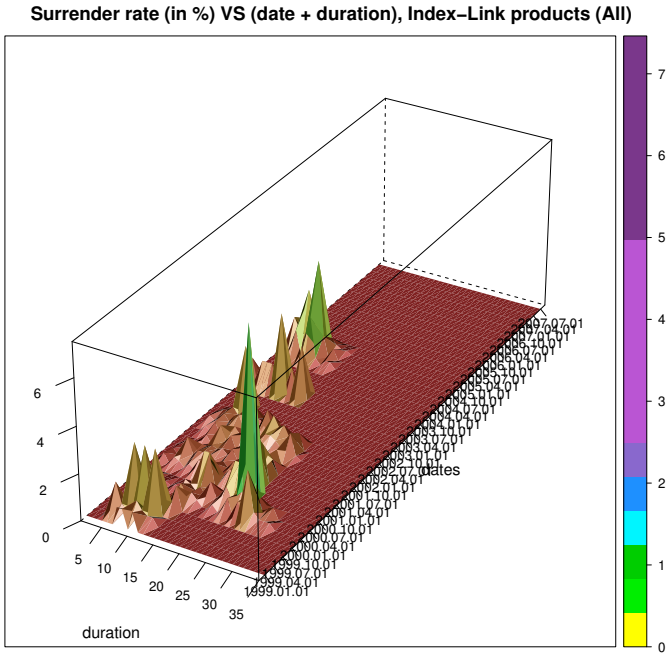


FIGURE C.16 – A gauche : rachat par ancienneté de contrat (en trimestre). A droite : Pourcentage global de rachat par cohorte. Produits Index-Link.

FIGURE C.17 – Profil 3D du taux de rachat par date et par ancienneté de contrat, Index-Link.



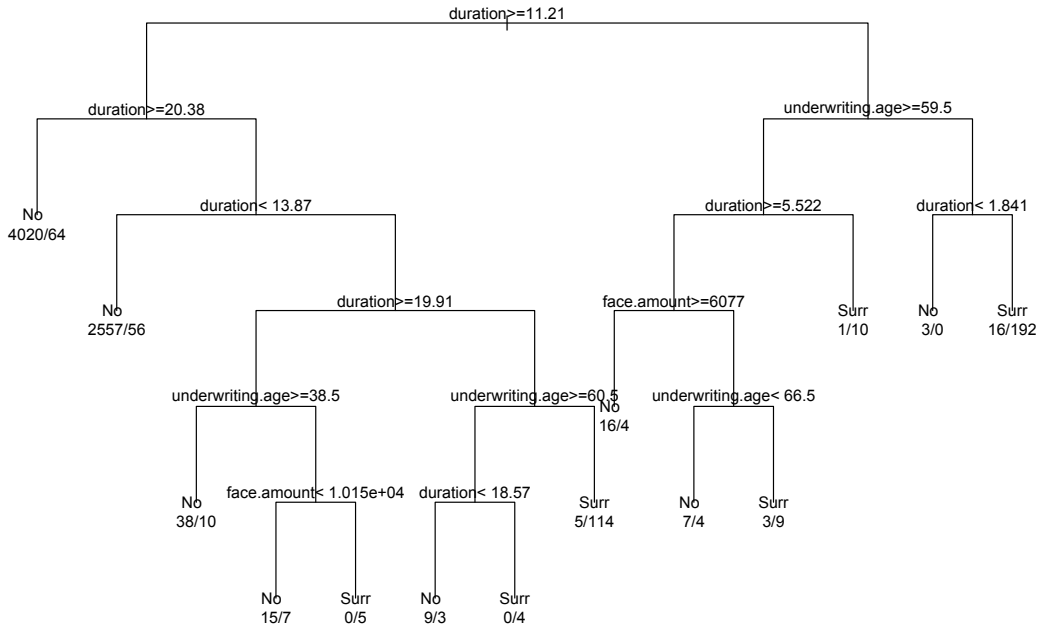
C.4.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre Le classifieur par forêts aléatoires se trompe rarement dans la prévision des rachats lorsque ceux-ci sont effectivement observés (erreur que nous cherchons à minimiser car la plus risquée pour nous), donnant une spécificité rassurante de 99 %. La sensibilité vaut ici 72 % et l'erreur globale de classification est égale à 3,6 %. Le classifieur est très précis sur l'étude statique.

	Rachats non-observés	Rachat observés
Rachats non-prédits	6770	70
Rachats prédits	204	526

Importance des variables explicatives Nous avons choisi cette fois de montrer le classifieur sous forme d'arbre (figure C.18) de classification par la méthode échantillon témoin-échantillon de validation. Nous avons sensiblement le même classement que pour les produits en UC, avec une certaine importance de l'âge de souscription. L'intérêt de cette représentation est la définition de seuils précis, utiles dans un processus de segmentation. Nous retenons donc l'ancienneté du contrat et l'âge de souscription.

FIGURE C.18 – Arbre de classification, donnant l'importance des variables explicatives des produits Index-Link en partant de la racine vers les feuilles.



C.4.3 Boxplot des coefficients du modèle

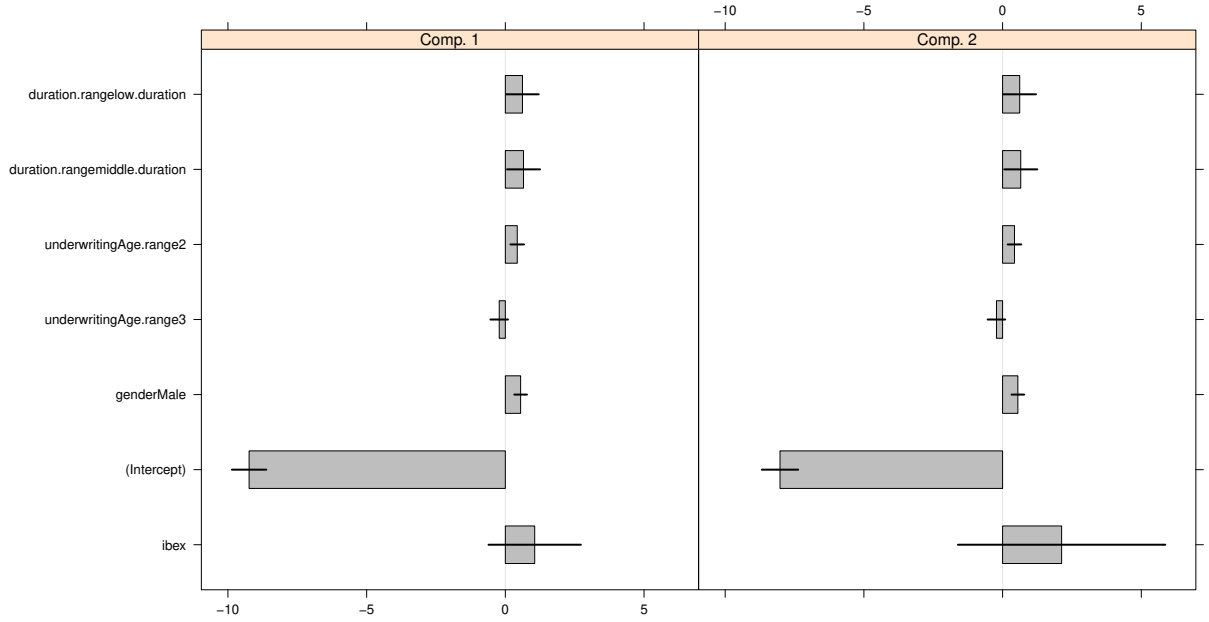
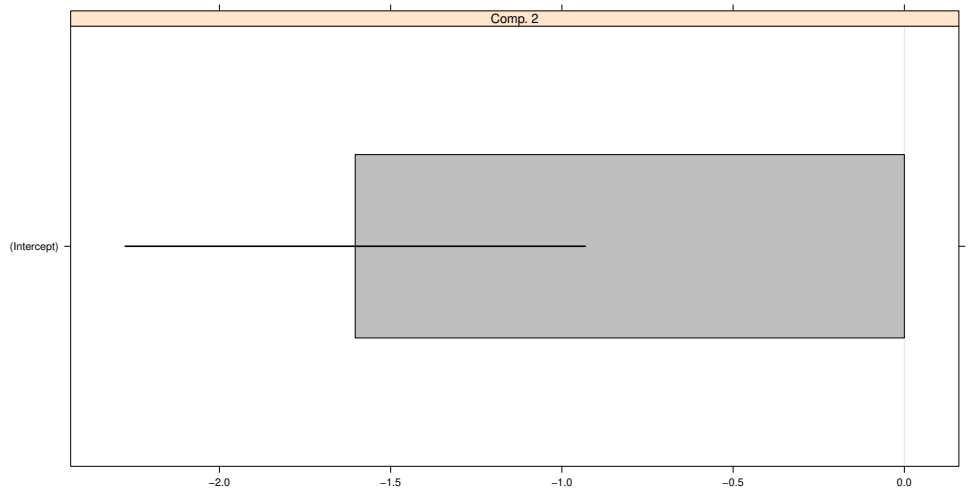


FIGURE C.19 – Coefficients de régression des composantes, produits Index-Link.

FIGURE C.20 – Coefficients de régression estimés des poids des composantes, Index-Link.



C.5 Famille de produits Universal Savings

C.5.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille Comme nous pouvons le constater dans le graphique C.21, les chutes (dont les rachats) n'ont malheureusement été enregistrées dans la base de données qu'à partir de Janvier 2004. L'étude de cette grande famille de produit est intéressante car l'exposition est très importante (elle va jusqu'à 150 000 contrats simultanément en portefeuille). Là encore, la crise financière semble avoir joué un rôle étant donné la chute constatée à partir de 2008. Hormis cette chute, la tendance des rachats était à une croissance légère et continue, peu volatile et présentant une certaine périodicité.

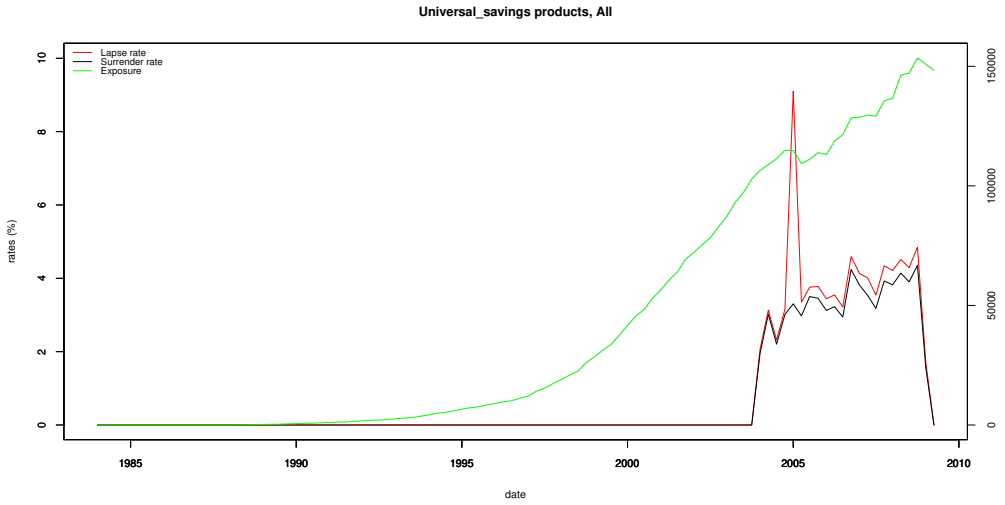
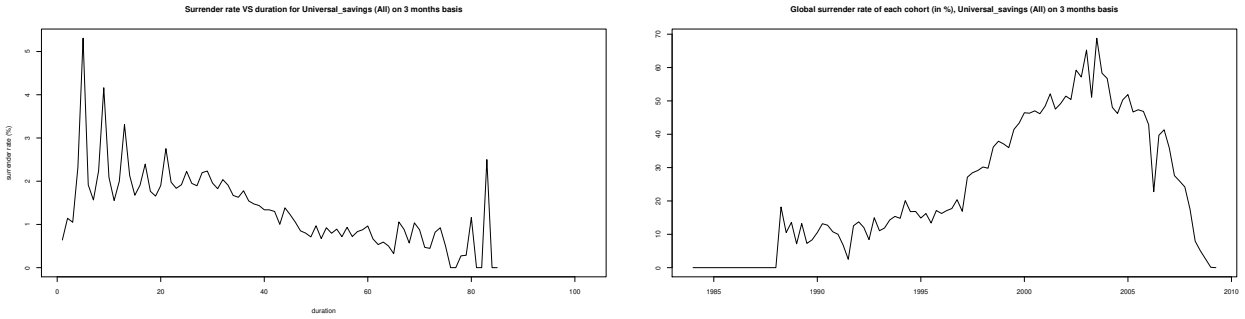


FIGURE C.21 – Exposition et taux de rachat trimestriel du portefeuille, Universal Savings.

FIGURE C.22 – A gauche : rachat par ancienneté de contrat (en trimestre). A droite : Pourcentage global de rachat par cohorte. Produits Universal Savings.



Profil des rachats par ancienneté de contrat et taux de rachat par cohorte En exceptant les premières années pour lesquelles nous retrouvons le même type de profil que pour les produits de pure épargne, la décroissance est plutôt linéaire ensuite, avant de se terminer par un pic dont il semblerait logiquement qu'il soit lié à quelques cohortes particulières. L'autre différence concerne le taux de rachat global par cohorte : bizarrement les cohortes les plus jeunes ont déjà plus racheté que les anciennes (croissance linéaire), ce qui traduit clairement une évolution des moeurs des assurés sur les comportements de rachat. Les enseignements que nous pouvons tirer des graphiques de la figure C.22 sont la nécessité une fois de plus de rendre la variable "ancienneté" catégorielle, et une hétérogénéité entre générations créée par un facteur non-observable. Nous n'affichons pas le taux de rachat par date et par ancienneté de contrat car les conclusions sont identiques aux précédentes.

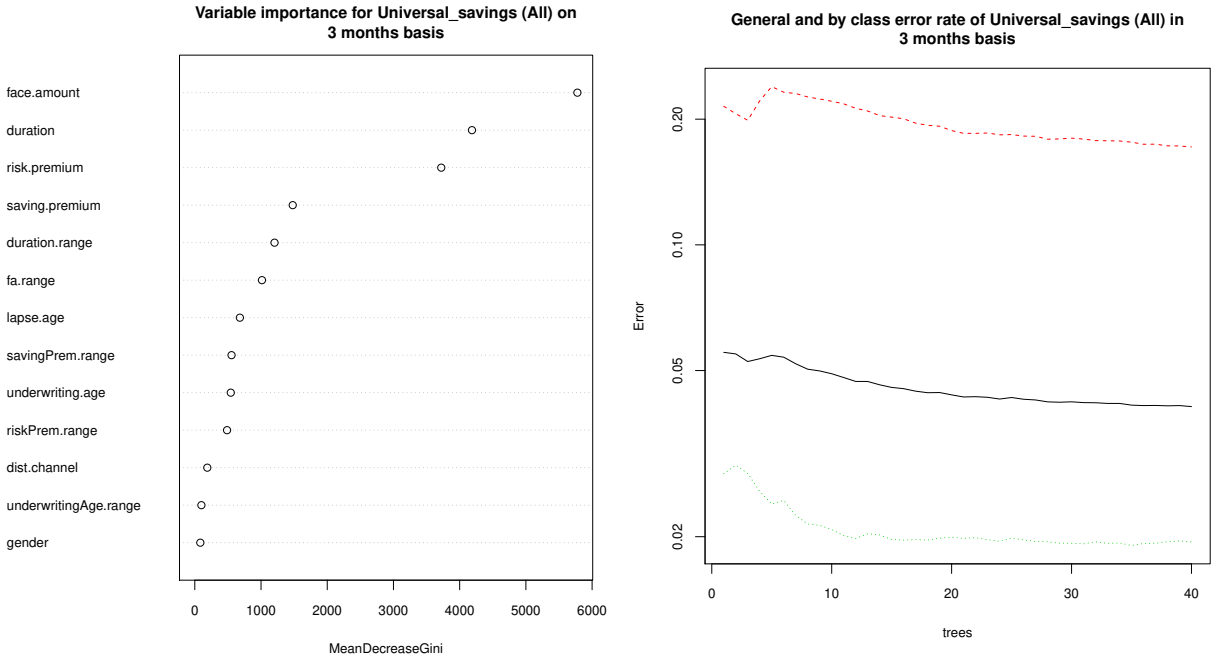
C.5.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre Le processus de classification (sur l'échantillon de validation) résumé dans le tableau ci-dessous donne une erreur de 4.1 %. L'erreur liée à la spécificité (82.8 % de bonnes prévisions) est un peu plus grande que celle de la sensibilité (98.1 %), mais globalement les forêts aléatoires ont un bon pouvoir de prévision ici.

	Rachats non-observés	Rachat observés
Rachats non-prédits	11972	2480
Rachats prédits	1700	85840

Importance des variables explicatives Il faut noter que pour l'une des premières fois ce n'est pas l'ancienneté du contrat qui apparait comme le facteur de risque le plus discriminant par rapport au comportement de rachat. La richesse de l'assuré est clef dans le processus de décision au vu des résultats de la figure C.23, suivie de l'ancienneté de contrat et de la prime de risque et d'épargne. Les niveaux de ces différentes primes sont corrélés à la richesse de l'assuré : nous prenons donc uniquement cette dernière variable dans la modélisation. Malgré leur impact relativement faible au vu de ce graphique, le réseau de distribution et l'âge de souscription seront introduits dans la modélisation de la décision de rachat pour gagner en précision.

FIGURE C.23 – Importance des variables explicatives, produit Universal Savings.



C.5.3 Boxplot des coefficients du modèle

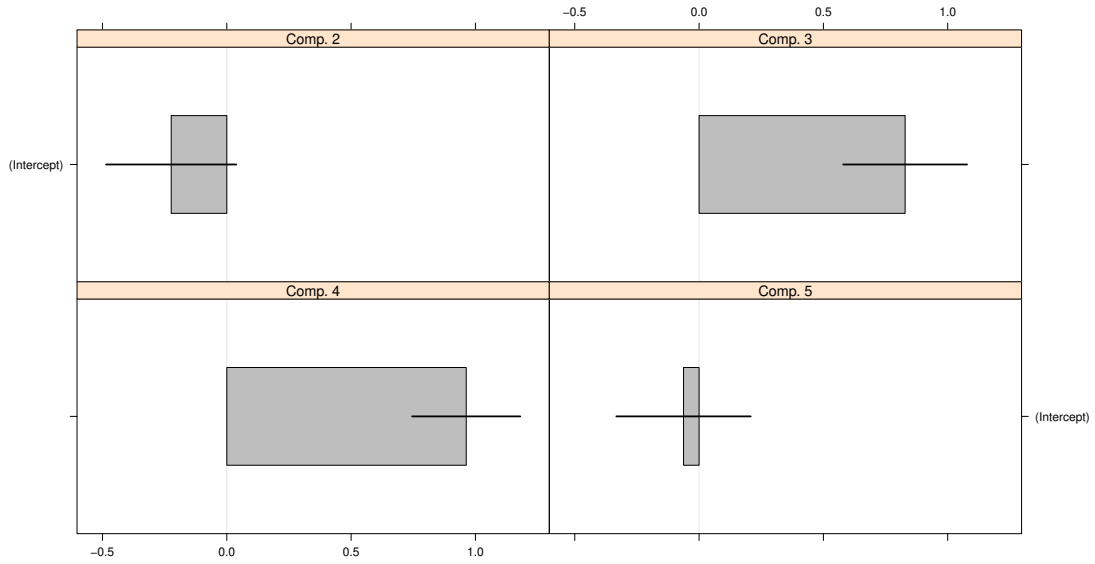
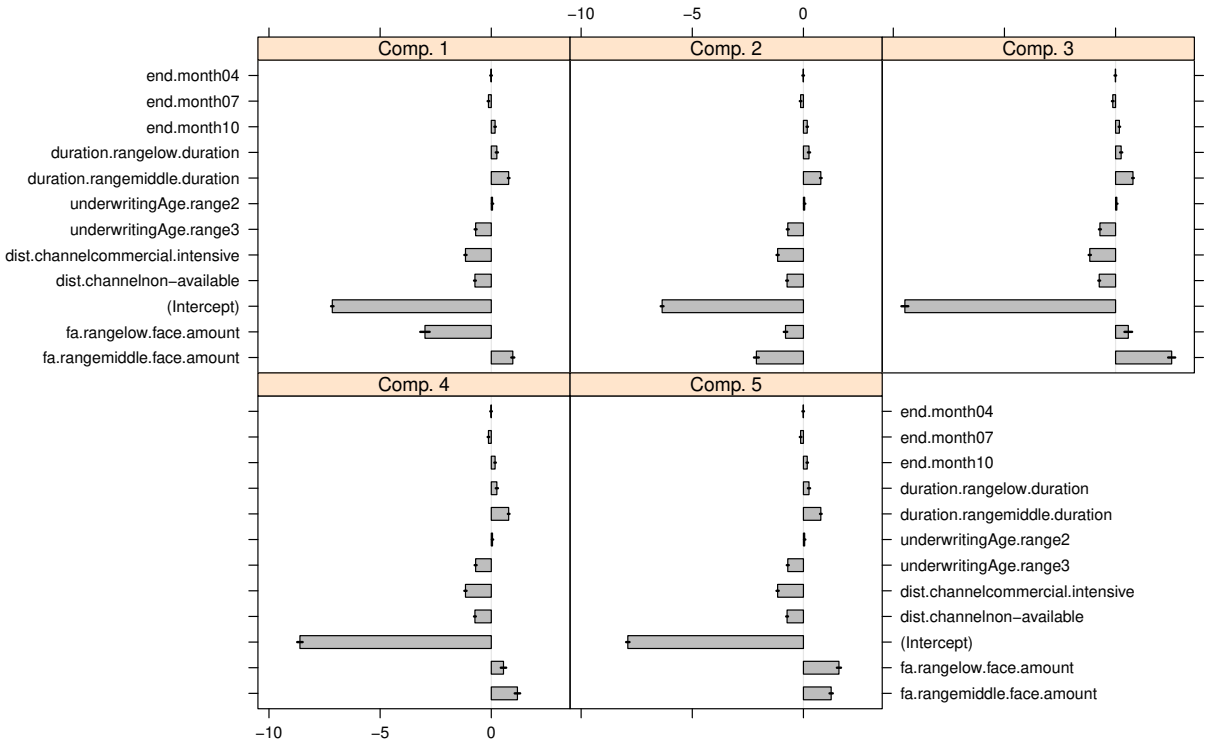


FIGURE C.24 – Coefficients de régression estimés des poids des composantes, produits Universal Savings.

FIGURE C.25 – Coefficients de régression des composantes, produits Universal Savings.



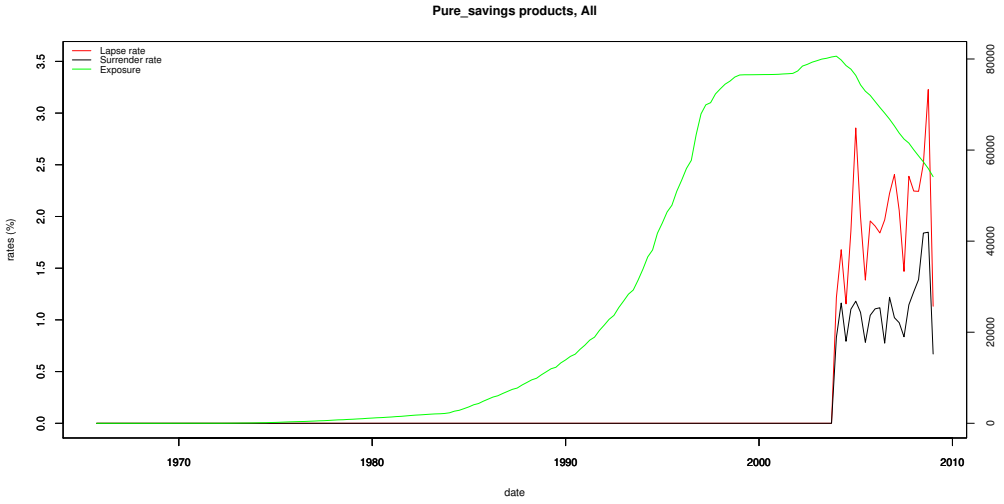
C.6 Famille de produits Pure Savings

C.6.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille Le constat du graphique C.26 est que le taux de rachat semble avoir une saisonnalité (périodicité), avec un taux de rachat qui semble augmenter et baisser à des intervalles de temps réguliers. Sur la fin de la période d'observation, nous remarquons un comportement anormal de la courbe avec un forte hausse suivie d'une baisse brusque et importante. Nous allons voir dans la suite s'il est possible d'expliquer ces mouvements. L'exposition sur cette ligne de produit peut aller jusqu'à 80 000 contrats, ce qui laisse présager des résultats statistiques robustes (qu'ils soient bons ou mauvais).

Profil des rachats par ancienneté de contrat Le profil des rachats en fonction de l'ancienneté des contrats donné par le graphe C.27 semble indiquer une forme en cloche, avec une majorité des assurés qui rachète en moyenne vers le 50ème trimestre (13ème année). Cette forme n'a d'explication ni par les conditions des contrats quant aux rachats, ni par la fiscalité. Cela reste donc une donnée statistique dont il faudra tenir compte lors de l'introduction de la variable "ancienneté" dans la modélisation (sous forme catégorielle de préférence donc).

FIGURE C.26 – Exposition et taux de rachat trimestriel du portefeuille, Pure Savings.



Taux de rachat par cohorte et nouvelles affaires Nous avons décidé d'afficher l'évolution des nouvelles affaires sur ce produit car un phénomène peu courant est à l'origine de l'"hétérogénéité" constatée sur la figure C.28 (à gauche). En fait il n'y a pratiquement plus de nouvelles affaires souscrites fin 1999-début 2000 (entre 1 et 15 contrats par trimestre !), ce qui fait mécaniquement chuter le taux de rachat de ces cohortes lorsque le(s) seul(s) assuré(s) ayant souscrit n'a(ont) pas racheté. Ce phénomène est marginal et ne doit donc pas être interprêté comme des changements de comportements (hétérogènes), de même que le pic de la cohorte fin 2008 (1 seul contrat avait été souscrit et il a été racheté). Comme pour les produits Universal Savings, nous n'affichons pas le graphique 3D du taux de rachat en fonction de l'ancienneté du contrat et de la date car il n'amène pas d'information intéressante supplémentaire.

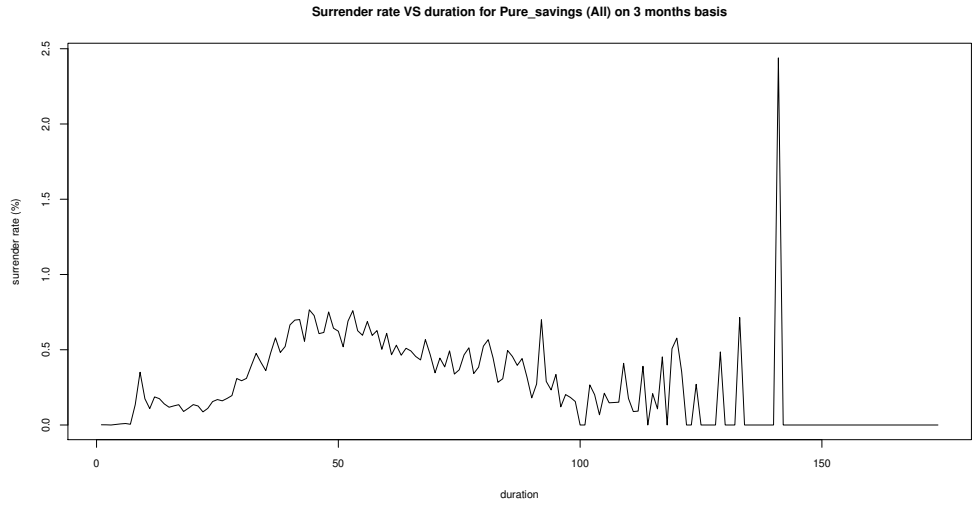
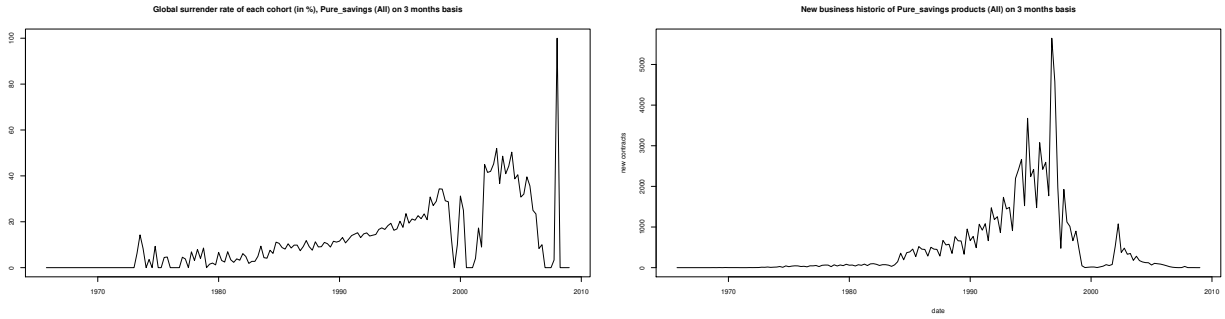


FIGURE C.27 – Rachat par ancienneté de contrat (en trimestre) pour les produits Pure Savings.

FIGURE C.28 – A gauche : taux de rachat global par cohorte. A droite : nouvelles affaires par trimestre. Produits Pure Savings.



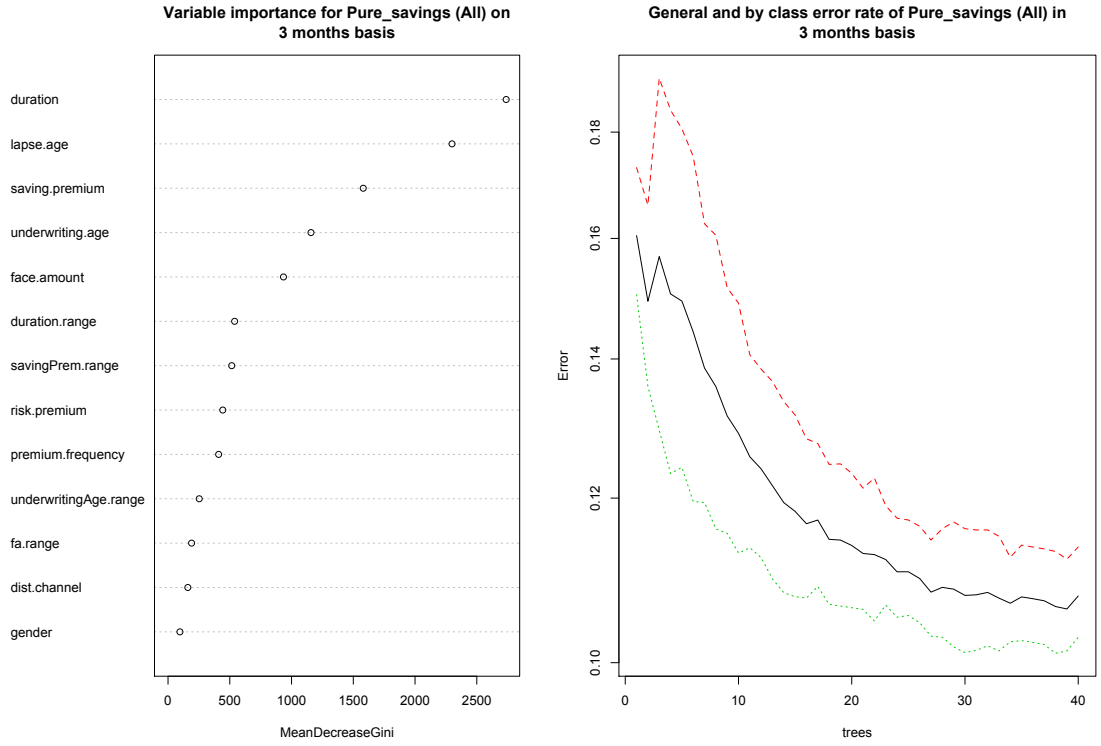
C.6.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre Les critères de performance de la classification des comportements de rachat sur les produits Pure Savings (échantillon de validation) sont quasiment similaires : la spécificité vaut 89.7 % tandis que la sensibilité vaut 88.7 %, pour un taux d'erreur global de mauvaise classification de 10.8 %. La méthode CART apparaît encore comme une bonne alternative de modèle de classification, malgré un taux d'erreur en hausse comparé aux précédentes applications.

	Rachats non-observés	Rachat observés
Rachats non-prédits	11046	1411
Rachats prédits	1602	13944

Importance des variables explicatives Au vu de la figure C.29, les variables explicatives les plus discriminantes sont dans l'ordre décroissant : l'ancienneté de contrat, l'âge au moment du rachat, la prime d'épargne, l'âge de souscription, la richesse... Pour rester en ligne avec les hypothèses de modélisation (variables indépendantes en théorie) et les modélisations des autres produits, nous considérons la saisonnalité, l'ancienneté de contrat et l'âge de souscription (corrélé à l'âge au moment rachat). Ces trois facteurs explicatifs devraient suffire à effectuer nos prévisions de taux.

FIGURE C.29 – Importance des variables explicatives, produit Pure Savings.



C.6.3 Boxplot des coefficients du modèle

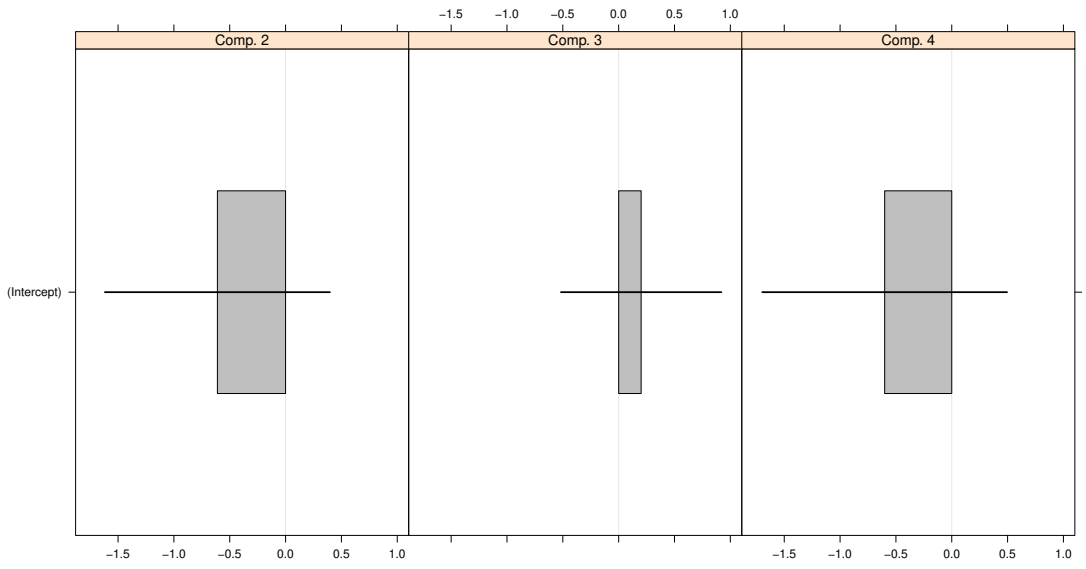
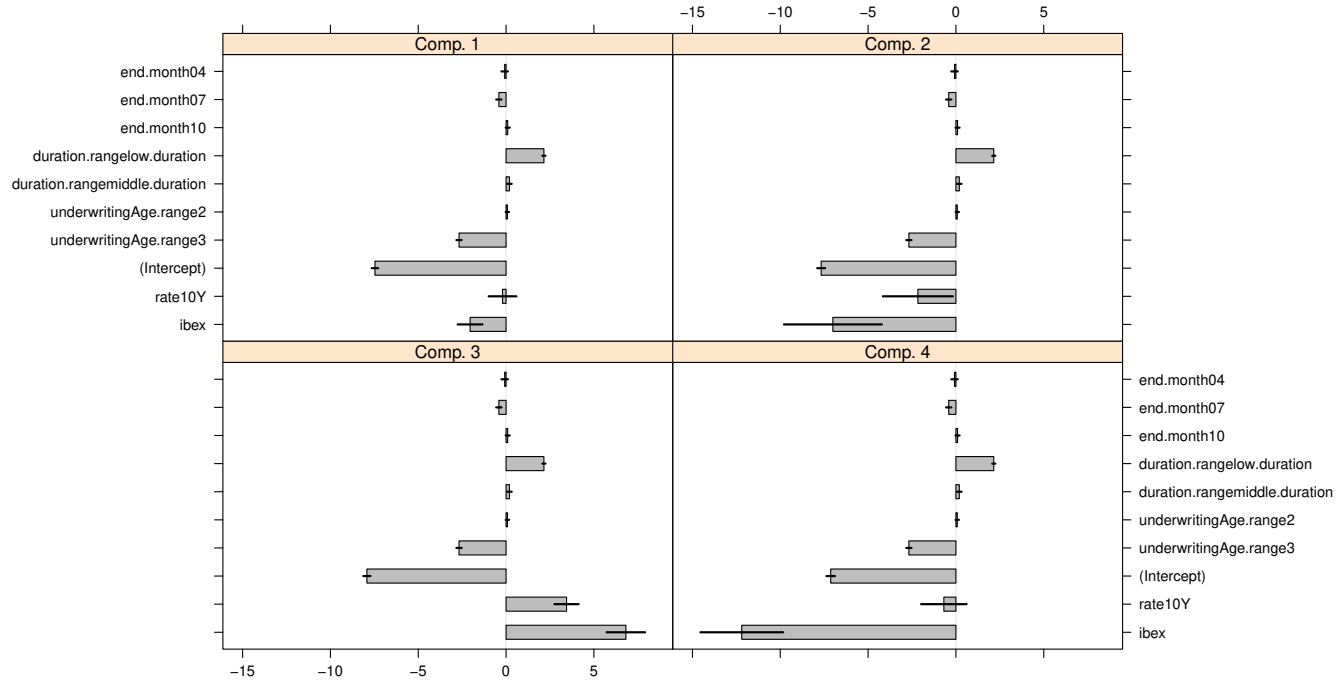


FIGURE C.30 – Coefficients de régression estimés des poids des composantes, Pure Savings.

FIGURE C.31 – Coefficients de régression des composantes, produits Pure Savings.



C.7 Famille de produits "Structured Products"

C.7.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille Nous n'affichons pas le taux de chute car il écrase complètement le taux de rachat, signalant au passage que les comportements de rachat ne sont pas très nombreux sur cette ligne (mais l'exposition est limitée par rapport aux produits dont nous venons de parler). Nous verrons par la suite que le taux de rachat présente en revanche une très forte volatilité, ce qui n'est pas étonnant vu le fonctionnement même de cette famille de produit. Aucune saisonnalité n'est évidente. Toute la difficulté sera donc de considérer les bonnes variables explicatives dans la modélisation.

Profil des rachats par ancienneté de contrat Le profil des rachats par ancienneté de contrat et le taux de rachat par cohorte (graphe C.33), ainsi que le taux de rachat par date et par ancienneté de contrat (graphe C.34) laisse présager une très forte hétérogénéité. Il est difficile de considérer que le taux de rachat est monotone en fonction de l'ancienneté, même si dans ce cas il semble qu'une tendance se dégage (plus d'assurés rachètent plus tard, ce qui peut paraitre surprenant d'ailleurs). En ce qui concerne le comportement des cohortes, il est très imprévisible (notons qu'il n'y a plus de souscription sur ce type de produit depuis début 2005). Nous ne sommes pas surpris de constater qu'il n'existe aucun profil type à dégager de ce type de produit, qui de par sa complexité et le fait qu'il dépende souvent uniquement des marchés rend les comportements aussi hétérogènes qu'imprévisibles. Notre objectif sera donc

FIGURE C.32 – Exposition et taux de rachat trimestriel du portefeuille de produits Structurés.

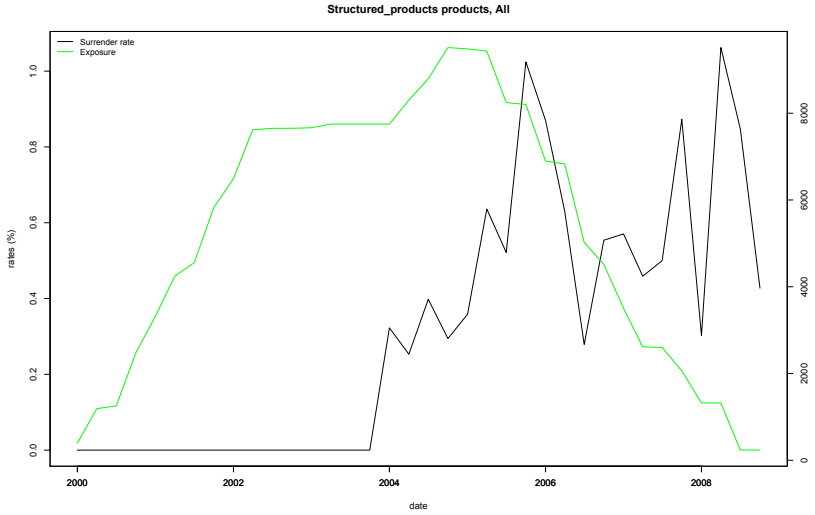
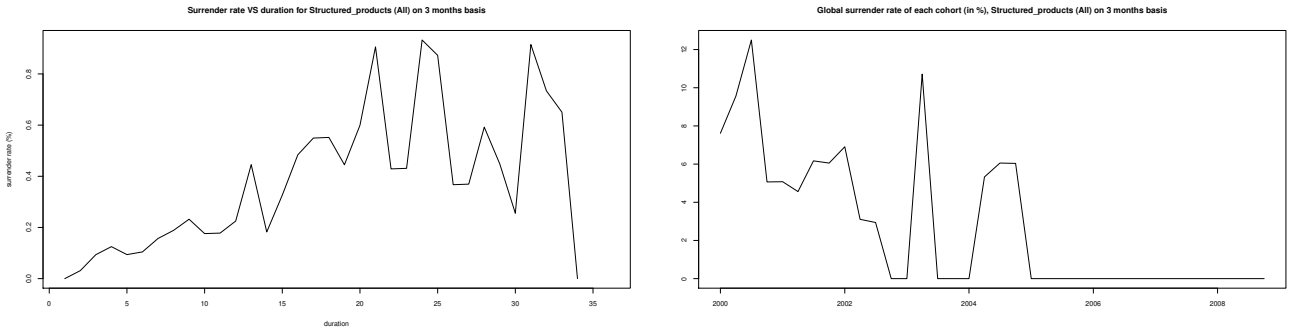


FIGURE C.33 – A gauche : Taux de rachat en fonction de l'ancienneté de contrat. A droite : taux de rachat global par cohorte. Produits structurés.

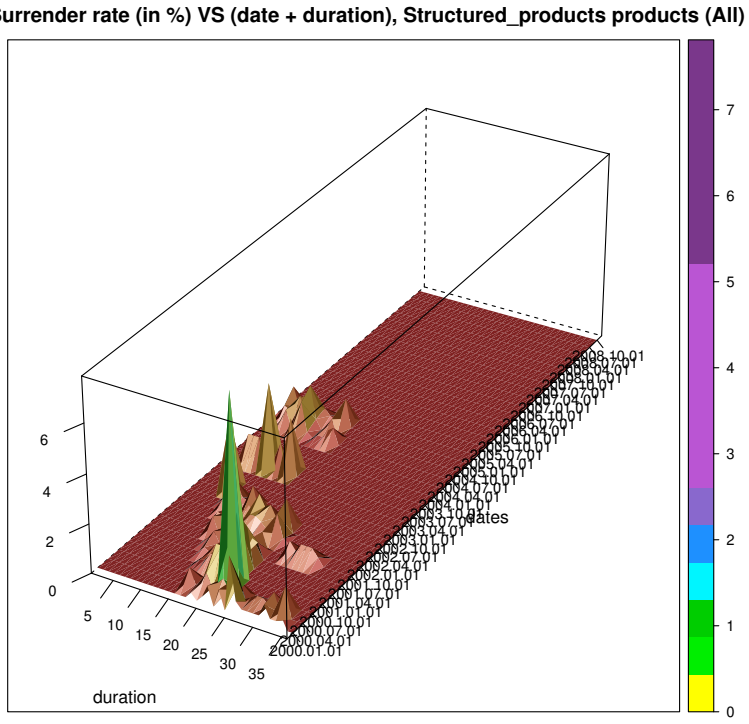


ici de ne pas trop compter sur des effets structurels qui a priori n'aurait qu'un impact dérisoire sur la modélisation finale, et qui risque de "polluer" la modélisation. Ce postulat reste toutefois à vérifier dans l'application.

C.7.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre La classification des comportements de rachat des produits structurés est précise. Le taux d'erreur est seulement de 3,8 %, avec une spécificité de 99,8 % et une sensibilité de 40,3 %. Ces résultats peuvent quand même être trompeur lorsque l'on s'intéresse à ce type de produit en termes de modélisation. En effet, la dépendance aux marchés de cette famille de produit est telle que certaines variables qui apparaissent comme importantes dans ce classifieur peuvent finalement s'avérer inutiles à des fins de projection.

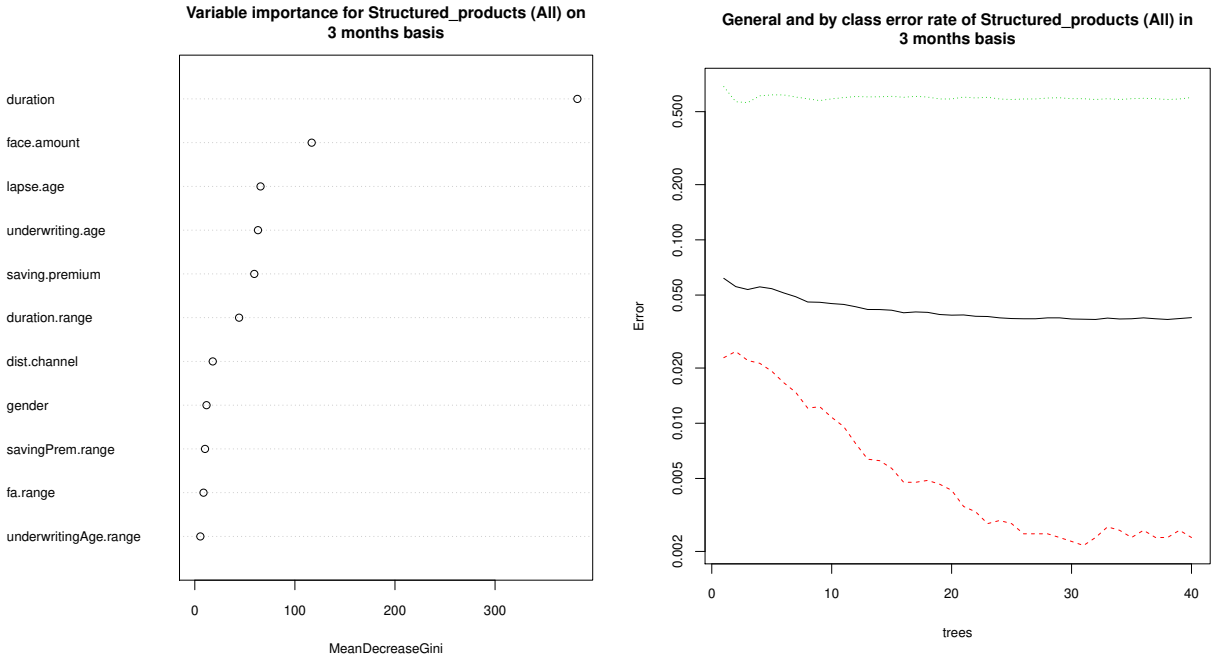
FIGURE C.34 – Profil 3D du taux de rachat par date et par ancienneté de contrat (par trimestre), produit Structurés.



	Rachats non-observés	Rachat observés
Rachats non-prédits	8795	21
Rachats prédits	332	224

Importance des variables explicatives L'ancienneté du contrat, la richesse et l'âge sont les trois variables les plus discriminantes dans le processus de segmentation d'après la figure C.35. Si nous considérons uniquement les variables catégorielles ou catégorisées, l'ancienneté et le réseau de distribution expliquent les décisions de rachat avec les meilleures prévisions. Nous allons voir que finalement aucune de ces variables explicatives n'est considérée dans le modèle prédictif des décisions de rachat, car leur introduction dégradaient clairement sa qualité.

FIGURE C.35 – Importance des variables explicatives, produit Structurés.



C.7.3 Boxplot des coefficients du modèle

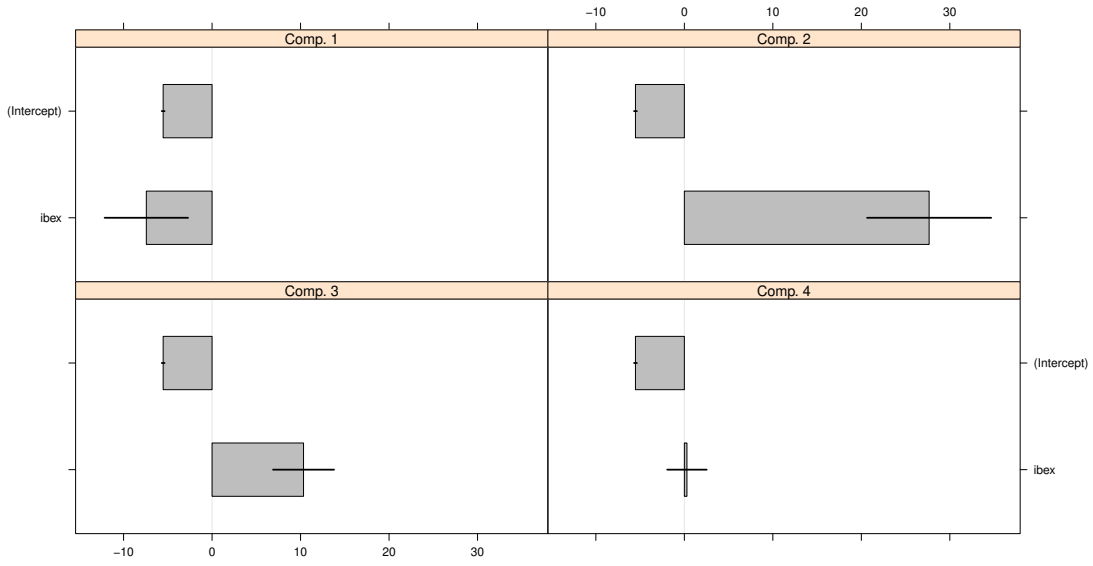


FIGURE C.36 – Coefficients de régression des composantes, produits structurés.

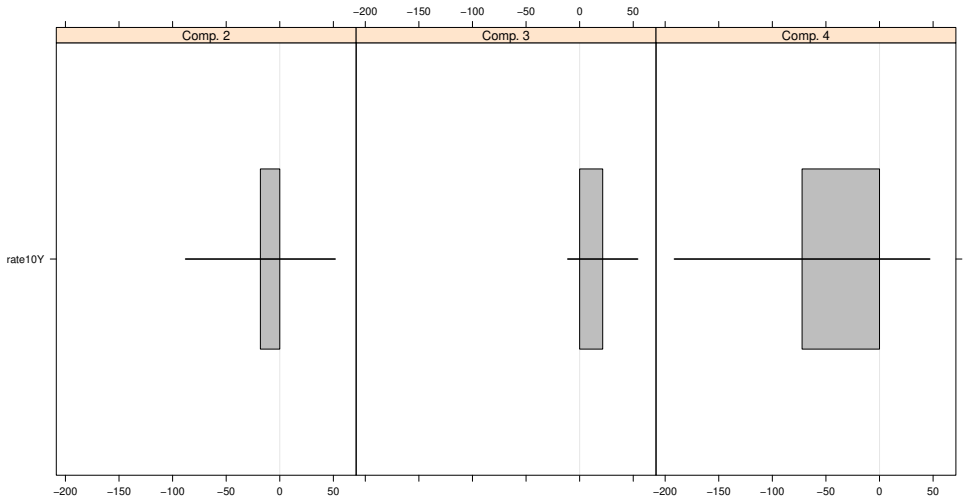


FIGURE C.37 – Coefficients de régression estimés des poids des composantes, produits structurés.

Annexe D

Espace des paramètres des GLMs

D.1 Mélange de régressions linéaires

D.1.1 Décomposition de la log-vraisemblance L_{cc}

La log-vraisemblance classifiante conditionnelle $\log L_{cc}(\psi_G; y_j)$ peut s'écrire pour une observation y_j comme la somme de deux termes :

$$\log L_{cc}(\psi_G; y_j) = \log\Big(\overbrace{\sum_{i=1}^{G} \overbrace{\pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{(y_j - X_j\beta_i)^2}{\sigma_i^2}\right)}^{A_i}}^{A}\Big) + \underbrace{\sum_{i=1}^{G} \underbrace{\frac{\pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{(y_j - X_j\beta_i)^2}{\sigma_i^2}\right)}{\sum_{k=1}^{G} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2}\frac{(y_j - X_j\beta_k)^2}{\sigma_k^2}\right)}}_{b_i = A_i / \sum_k A_k} \log\left(\frac{\pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{(y_j - X_j\beta_i)^2}{\sigma_i^2}\right)}{\sum_{k=1}^{G} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2}\frac{(y_j - X_j\beta_k)^2}{\sigma_k^2}\right)}\right)}_{B_i = b_i \log b_i}}_{B = \sum_i B_i}.$$

D.1.2 Calcul des limites et des cas critiques

Etudions les limites de A et B (dans le cas unidimensionnel pour simplifier) dans les différentes configurations possibles. Nous devons ainsi calculer et exprimer :

1. $\lim_{\pi_i \to 0^+} \log L_{cc}(\psi_G; y_j)$;
2. $\lim_{\beta_i \to +\infty} \log L_{cc}(\psi_G; y_j)$ et $\lim_{\beta_i \to -\infty} \log L_{cc}(\psi_G; y_j)$;
3. $\lim_{\sigma_i^2 \to 0^+} \log L_{cc}(\psi_G; y_j)$ et $\lim_{\sigma_i^2 \to +\infty} \log L_{cc}(\psi_G; y_j)$.

Remarque importante : dans toute cette annexe, nous calculons les limites en considérant que X_j, y_j sont positifs afin de simplifier l'exposition des résultats (trop de cas différents sinon). De plus, l'étude des limites pour un paramètre donné se fera en considérant les autres paramètres fixes à des valeurs non problématiques.

Calcul de la limite : $\lim\limits_{\pi_i \to 0^+} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Premièrement, il est immédiat que $\lim\limits_{\pi_i \to 0^+} A_i = 0$. Comme $\sum_{i=1}^{G} \pi_i = 1$, tous les poids ne peuvent pas être nuls en même temps, d'où $\lim\limits_{\pi_i \to 0^+} A = K$, avec K une constante.
$\rightarrow$ De plus, nous exploitons le résultat bien connu selon lequel $\lim\limits_{\pi_i \to 0} \pi_i \log \pi_i = 0$. Ainsi, on a $\lim\limits_{\pi_i \to 0^+} B_i = 0$ et $\lim\limits_{\pi_i \to 0^+} B = K^{'}$. Par contre, nous réalisons qu'un problème apparaît pour la dérivée de l'entropie puisqu'elle n'est pas dérivable en 0.

$\Rightarrow$ Finalement, il vient naturellement

$$\lim_{\pi_i \to 0^+} \log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\pi_i \to 0^+} \frac{\partial \log L_{cc}(\psi_G; y_j)}{\partial \pi_i} = \infty.$$

Nota Bene : le raisonnement sera identique pour chaque classe de GLMs, aussi nous ne redévelopperons pas cet argumentaire et considèrerons que la limite lorsque π_i tend vers 0 est toujours problèmatique pour la dérivée de l'entropie.

Calcul de la limite : $\lim\limits_{\beta_i \to +\infty} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Tout d'abord, $\lim\limits_{\beta_i \to +\infty} A_i = 0$. Par conséquent, si tous les β_i $(i = 1, ..., G)$ ne tendent pas vers l'infini en même temps alors $\lim\limits_{\beta_i \to +\infty} \log A = \lim\limits_{\beta_i \to +\infty} \log(A_1 + ... + A_i + ... + A_G) = K^{'}$.
$\rightarrow$ Sachant que $b_i = A_i/A$, il vient directement que $\lim\limits_{\beta_i \to +\infty} b_i = 0$. Nous nous retrouvons dans la configuration où $\lim\limits_{\beta_i \to +\infty} B_i = \lim\limits_{\beta_i \to +\infty} b_i \log b_i = 0$. Ainsi $\lim\limits_{\beta_i \to +\infty} B = K$. En revanche, comme $\lim\limits_{\beta_i \to +\infty} b_i = 0$, la dérivée de l'entropie en 0 explose quand $\beta_i \to +\infty$.

$\Rightarrow$ Finalement, on a

$$\lim_{\beta_i \to +\infty} \log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\beta_i \to +\infty} \frac{\partial \log L_{cc}(\psi_G; y_j)}{\partial \beta_i} = \infty.$$

Calcul de la limite : $\lim\limits_{\beta_i \to -\infty} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Tout d'abord, $\lim\limits_{\beta_i \to -\infty} A_i = 0$. Par conséquent, si tous les β_i $(i = 1, ..., G)$ ne tendent pas vers moins l'infini en même temps alors $\lim\limits_{\beta_i \to -\infty} A = K$.
$\rightarrow$ Par le même raisonnement que précédemment, nous obtenons $\lim\limits_{\beta_i \to -\infty} b_i = 0$ et $\lim\limits_{\beta_i \to -\infty} B = K$.

$\Rightarrow$ Finalement, il vient donc

$$\lim_{\beta_i \to -\infty} \log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\beta_i \to -\infty} \frac{\partial \log L_{cc}(\psi_G; y_j)}{\partial \beta_i} = \infty.$$

Calcul de la limite : $\lim\limits_{\sigma_i^2 \to 0^+} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Lorsque nous calculons $\lim\limits_{\sigma_i^2 \to 0} A_i$, nous avons une forme indéterminée (F.I) de type "$+\infty \times 0$". Pour lever cette F.I., nous pouvons écrire A_i d'une autre manière : en effet après quelques transformations,

$$A_i = \pi_i \exp\left[-\frac{1}{2}\frac{1}{\sigma_i^2}\left((y_j - X_j\beta_i)^2 + \sigma_i^2 \log(2\pi\sigma_i^2)\right)\right].$$

Nous avons donc grâce à cette expression $\lim\limits_{\sigma_i^2 \to 0} A_i = 0$.
Par conséquent, $\lim\limits_{\sigma_i^2 \to 0} \log A = \lim\limits_{\sigma_i^2 \to 0} \log(A_1 + ... + A_i + ... + A_G) = K$ dans le cas général.
$\rightarrow$ Sachant que $b_i = A_i/A$, nous avons encore une fois $\lim\limits_{\sigma_i^2 \to 0} b_i = 0$ et $\lim\limits_{\sigma_i^2 \to 0} B = K^{'}$.

$\Rightarrow$ Finalement,

$$\lim_{\sigma_i^2 \to 0^+} \log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\sigma_i^2 \to 0^+} \frac{\partial \log L_{cc}(\psi_G; y_j)}{\partial \sigma_i^2} = \infty.$$

ATTENTION : il existe des cas particuliers pour lesquels la première des deux limites ci-dessus est fausse. Par exemple, si l'observation y_j vaut la moyenne $X_j\beta_i$ et que la variance tend vers 0, la vraisemblance explose. Nos résultats s'inscrivent dans un cadre général mais n'ont pas vocation à étudier tous les cas particuliers.

Calcul de la limite : $\lim\limits_{\sigma_i^2 \to +\infty} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Tout d'abord, $\lim\limits_{\sigma_i^2 \to +\infty} A_i = 0$. Par conséquent, il vient $\lim\limits_{\sigma_i^2 \to +\infty} A = K$.
$\rightarrow$ Par le même raisonnement que quand $\beta_i \to +\infty$, nous obtenons $\lim\limits_{\sigma_i^2 \to +\infty} b_i = 0$ et $\lim\limits_{\sigma_i^2 \to +\infty} B = K$.

$\Rightarrow$ Finalement, il vient donc

$$\lim_{\sigma_i^2 \to +\infty} \log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\sigma_i^2 \to +\infty} \frac{\partial \log L_{cc}(\psi_G; y_j)}{\partial \sigma_i^2} = \infty.$$

D.2 Mélange de régressions de Poisson

D.2.1 Décomposition de la log-vraisemblance L_{cc}

La log-vraisemblance classifiante conditionnelle $\log L_{cc}(\psi_G; y_j)$ peut s'écrire pour une observation y_j comme la somme de deux termes :

$$\log\Big(\underbrace{\sum_{i=1}^{G}\underbrace{\pi_i\, e^{-e^{X_j\beta_i}}\,\frac{\left[e^{X_j\beta_i}\right]^{y_j}}{y_j!}}_{A_i}}_{A}\Big)+\underbrace{\sum_{i=1}^{G}\underbrace{\frac{\pi_i\, e^{-e^{X_j\beta_i}}\,\frac{\left[e^{X_j\beta_i}\right]^{y_j}}{y_j!}}{\sum_{k=1}^{G}\pi_k\, e^{-e^{X_j\beta_k}}\,\frac{\left[e^{X_j\beta_k}\right]^{y_j}}{y_j!}}}_{b_i}\log\left(\frac{\pi_i\, e^{-e^{X_j\beta_i}}\,\frac{\left[e^{X_j\beta_i}\right]^{y_j}}{y_j!}}{\sum_{k=1}^{G}\pi_k\, e^{-e^{X_j\beta_k}}\,\frac{\left[e^{X_j\beta_k}\right]^{y_j}}{y_j!}}\right)}_{B_i=b_i\log b_i}}_{B=\sum_i B_i}.$$

D.2.2 Calcul des limites et des cas critiques

La loi de Poisson n'ayant qu'un paramètre, seules les limites de β_i en l'infini et son opposé seront d'intérêt ici. L'étude de ces limites équivaut à faire tendre la moyenne de la distribution de Poisson vers les extrêmes de son domaine de définition (0 et l'infini).

Calcul de la limite : $\lim\limits_{\beta_i\to-\infty}\log L_{cc}(\psi_G; y_j)$

→ Tout d'abord, $\lim\limits_{\beta_i\to-\infty} A_i = 0$. Par conséquent, si tous les β_i $(i = 1, ..., G)$ ne tendent pas vers moins l'infini en même temps alors $\lim\limits_{\beta_i\to-\infty} A = K$.
→ Par le même raisonnement que précédemment, nous obtenons $\lim\limits_{\beta_i\to-\infty} b_i = 0$ et $\lim\limits_{\beta_i\to-\infty} B = K$.

$$\Rightarrow \text{Finalement,} \quad \lim_{\beta_i\to-\infty}\log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\beta_i\to-\infty}\frac{\partial\log L_{cc}(\psi_G; y_j)}{\partial\beta_i} = +\infty.$$

Calcul de la limite : $\lim\limits_{\beta_i\to+\infty}\log L_{cc}(\psi_G; y_j)$

→ Nous avons une forme indéterminée "$0\times+\infty$" dans la limite $\lim\limits_{\beta_i\to+\infty} A_i$. Pour lever cette indétermination, observons que A_i peut aussi s'écrire après quelques calculs

$$A_i = \frac{\pi_i}{y_j!}\exp\left[e^{X_j\beta_i}\left(\frac{X_j\beta_i}{e^{X_j\beta_i}}-1\right)\right].$$

Ainsi, nous avons le résultat $\lim\limits_{\beta_i\to+\infty} A_i = 0$.
→ Toujours par le même raisonnement, $\lim\limits_{\beta_i\to+\infty} b_i = 0$ et $\lim\limits_{\beta_i\to+\infty} B = K$.
⇒ Finalement, on a

$$\lim_{\beta_i\to+\infty}\log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\beta_i\to+\infty}\frac{\partial\log L_{cc}(\psi_G; y_j)}{\partial\beta_i} = +\infty.$$

D.3 Mélange de régressions logistiques

D.3.1 Décomposition de la log-vraisemblance L_{cc}

La log-vraisemblance classifiante conditionnelle $\log L_{cc}(\psi_G; y_j)$ peut s'écrire pour une observation y_j comme la somme de deux termes :

$$\log L_{cc}(\psi_G; y_j) = \log \Big(\underbrace{\sum_{i=1}^{G} \underbrace{\pi_i \frac{e^{X_j\beta_i}}{1+e^{X_j\beta_i}}}_{A_i}}_{A} \Big) + \underbrace{\sum_{i=1}^{G} \underbrace{\frac{\pi_i \frac{e^{X_j\beta_i}}{1+e^{X_j\beta_i}}}{\sum_{k=1}^{G} \pi_k \frac{e^{X_j\beta_k}}{1+e^{X_j\beta_k}}}}_{b_i} \log \left(\frac{\pi_i \frac{e^{X_j\beta_i}}{1+e^{X_j\beta_i}}}{\sum_{k=1}^{G} \pi_k \frac{e^{X_j\beta_k}}{1+e^{X_j\beta_k}}} \right)}_{B_i = b_i \log b_i}}_{B = \sum_i B_i}$$

D.3.2 Calcul des limites et des cas critiques

La loi Binomiale comporte deux paramètres. Cependant, l'un des deux correspond à l'exposition du portefeuille, et est donc fixé. N'ayant qu'un paramètre variable, seules les limites de β_i en l'infini et son opposé seront étudiées : ceci équivaut à faire tendre la probabilité de rachat vers les extrêmes de son domaine de définition (0 et 1).

Calcul de la limite : $\lim\limits_{\beta_i \to -\infty} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Tout d'abord, $\lim\limits_{\beta_i \to -\infty} A_i = 0$. Par conséquent, si tous les β_i $(i = 1, ..., G)$ ne tendent pas vers moins l'infini en même temps alors $\lim\limits_{\beta_i \to -\infty} A = K$.

$\rightarrow$ Par le même raisonnement que précédemment, nous obtenons $\lim\limits_{\beta_i \to -\infty} b_i = 0$ et $\lim\limits_{\beta_i \to -\infty} B = K$.

$$\Rightarrow \text{Finalement,} \quad \lim_{\beta_i \to -\infty} \log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\beta_i \to -\infty} \frac{\partial \log L_{cc}(\psi_G; y_j)}{\partial \beta_i} = +\infty.$$

Calcul de la limite : $\lim\limits_{\beta_i \to +\infty} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Nous avons une forme indéterminée "$\frac{+\infty}{+\infty}$" dans la limite $\lim\limits_{\beta_i \to +\infty} A_i$. Pour lever cette indétermination, observons que A_i peut aussi s'écrire après factorisation

$$A_i = \pi_i \left(\frac{1}{1+e^{-X_j\beta_i}} \right).$$

Ainsi, on a immédiatement : $\lim\limits_{\beta_i \to +\infty} A_i = \pi_i$.

$\rightarrow$ D'autre part, $\lim\limits_{\beta_i \to +\infty} b_i = K$ et $\lim\limits_{\beta_i \to +\infty} B = K'$.

$\Rightarrow$ Finalement, on a

$$\lim_{\beta_i \to +\infty} \log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\beta_i \to +\infty} \frac{\partial \log L_{cc}(\psi_G; y_j)}{\partial \beta_i} = K'.$$

D.4 Mélange de régressions Gamma

Pour étudier les limites de mélanges de cette classe de GLMs, il est nécessaire de rappeler quelques propriétés liées à cette distribution de probabilité. Plus précisément, nous donnons ici quelques notions sur la fonction Gamma qui intervient dans la densité de la loi Gamma. Ainsi, nous avons $\forall z \in \mathbb{R}^{+*}$

$\begin{cases} \Gamma(z+1) = z\Gamma(z), \\ \Gamma(z+1) = z! \quad \text{(pour des entiers).} \end{cases}$

Au voisinage de l'infini, il existe un développement limité de *Gamma* : $\forall z \in v(+\infty)$,

$$\Gamma(z) = z^{z-\frac{1}{2}} e^{-z} \sqrt{2\pi} \left[1 + \frac{1}{12z} + \frac{1}{288z^2} + o(\frac{1}{z^3}) \right].$$

Nous pouvons donc dériver un simple équivalent de cette expression.

D.4.1 Décomposition de la log-vraisemblance L_{cc}

La log-vraisemblance classifiante conditionnelle $\log L_{cc}(\psi_G; y_j)$ peut s'écrire pour une observation y_j comme la somme de deux termes :

$$\log L_{cc}(\psi_G; y_j) = \log \Big(\overbrace{\sum_{i=1}^{G} \overbrace{\frac{\pi_i}{\Gamma(\nu_i)} \, (\nu_i X_j \beta_i)^{\nu_i} \, y_j^{\nu_i - 1} \, e^{-\nu_i X_j \beta_i \, y_j}}^{A_i}}^{A} \Big) +$$

$$\underbrace{\underbrace{\sum_{i=1}^{G} \underbrace{\frac{\frac{\pi_i}{\Gamma(\nu_i)} \, (\nu_i X_j \beta_i)^{\nu_i} \, y_j^{\nu_i - 1} \, e^{-\nu_i X_j \beta_i \, y_j}}{\sum_{k=1}^{G} \frac{\pi_k}{\Gamma(\nu_k)} \, (\nu_k X_j \beta_k)^{\nu_k} \, y_j^{\nu_k - 1} \, e^{-\nu_k X_j \beta_k \, y_j}}}_{b_i} \log \left(\frac{\frac{\pi_i}{\Gamma(\nu_i)} \, (\nu_i X_j \beta_i)^{\nu_i} \, y_j^{\nu_i - 1} \, e^{-\nu_i X_j \beta_i \, y_j}}{\sum_{k=1}^{G} \frac{\pi_k}{\Gamma(\nu_k)} \, (\nu_k X_j \beta_k)^{\nu_k} \, y_j^{\nu_k - 1} \, e^{-\nu_k X_j \beta_k \, y_j}} \right)}_{B_i = b_i \log b_i}}_{B = \sum_i B_i}.$$

D.4.2 Calcul des limites et des cas critiques

Calcul de la limite : $\lim\limits_{\nu_i \to 0^+} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Tout d'abord, $\lim\limits_{\nu_i \to 0^+} A_i = 0$. Par conséquent, si tous les ν_i $(i = 1, ..., G)$ ne tendent pas vers moins 0 en même temps alors $\lim\limits_{\nu_i \to 0^+} A = K$.

$\rightarrow$ Toujours en raisonnant de la même façon, nous obtenons $\lim\limits_{\nu_i \to 0^+} b_i = 0$ et $\lim\limits_{\nu_i \to 0^+} B = K$.

$\Rightarrow$ Finalement, il vient donc

$$\lim_{\nu_i \to 0^+} \log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\nu_i \to 0^+} \frac{\partial \log L_{cc}(\psi_G; y_j)}{\partial \nu_i} = +\infty.$$

Calcul de la limite : $\lim\limits_{\nu_i \to +\infty} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Nous avons une forme indéterminée "$+\infty \times 0$" dans la limite $\lim\limits_{\nu_i \to +\infty} A_i$. Pour lever cette indétermination, reformulons A_i en nous servant de l'équivalent de la fonction Γ en $+\infty$. Après quelques calculs,

$$A_i = \frac{\pi_i}{\sqrt{2\pi}} \exp\left[\nu_i \left(\log(X_j\beta_i y_j) + \frac{1}{2}\frac{\log \nu_i}{\nu_i} + 1 - \frac{\log y_j}{\nu_i} - X_j\beta_i y_j\right)\right].$$

Avec cette écriture, la limite diffère suivant le signe de $\log(X_j\beta_i y_j) - X_j\beta_i y_j + 1$:
$\begin{cases} \text{si } \log(X_j\beta_i y_j) - X_j\beta_i y_j + 1 > 0 \text{ alors } \lim\limits_{\nu_i \to +\infty} A_i = +\infty, \\ \text{si } \log(X_j\beta_i y_j) - X_j\beta_i y_j + 1 < 0 \text{ alors } \lim\limits_{\nu_i \to +\infty} A_i = 0. \end{cases}$

$\rightarrow$ Dans le cas où $\lim\limits_{\nu_i \to +\infty} A_i = +\infty$, il faut calculer la limite $\lim\limits_{\nu_i \to +\infty} B_i$ pour s'assurer qu'il n'y a pas de forme indéterminée pour trouver $\lim\limits_{\nu_i \to +\infty} L_{cc}(\psi_G, y_j)$. Pour cela, nous reformulons b_i de la manière suivante :

$$b_i = 1 - \frac{\sum\limits_{\substack{k=1 \\ k\neq i}}^{G} \pi_k \exp\left[\nu_k \log(X_j\beta_k y_j) + \frac{1}{2}\log \nu_k + \nu_k - \log y_j - \nu_k X_j \beta_k y_j\right]}{\sum_{k=1}^{G} \pi_k \exp\left[\nu_k \log(X_j\beta_k y_j) + \frac{1}{2}\log \nu_k + \nu_k - \log y_j - \nu_k X_j \beta_k y_j\right]}.$$

Donc,
$\begin{cases} \text{si } \log(X_j\beta_i y_j) - X_j\beta_i y_j + 1 > 0 \text{ alors } \lim\limits_{\nu_i \to +\infty} b_i = 1 \quad \Rightarrow \quad \lim\limits_{\nu_i \to +\infty} B_i = 0, \\ \text{si } \log(X_j\beta_i y_j) - X_j\beta_i y_j + 1 < 0 \text{ alors } \lim\limits_{\nu_i \to +\infty} b_i = +\infty \quad \Rightarrow \quad \lim\limits_{\nu_i \to +\infty} B_i = +\infty. \end{cases}$

$\Rightarrow$ Finalement, il vient donc

$$\lim_{\nu_i \to +\infty} \log L_{cc}(\psi_G; y_j) = +\infty.$$

Calcul de la limite : $\lim\limits_{\beta_i \to +\infty} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Nous avons une forme indéterminée "$+\infty \times 0$" dans la limite $\lim\limits_{\beta_i \to +\infty} A_i$. Nous reécrions A_i différemment pour lever cette indétermination :

$$A_i = \frac{\pi_i}{\Gamma(\nu_i)} y_j^{\nu_i - 1} \exp\left[\beta_i \left(\frac{\nu_i \log(\nu_i X_j \beta_i)}{\beta_i} - \nu_i X_j y_j\right)\right].$$

La limite diffère suivant le signe de $\nu_i X_j y_j$. Nous trouvons donc : $\begin{cases} \text{si } \nu_i X_j y_j > 0 \text{ alors } \lim\limits_{\beta_i \to +\infty} A_i = 0, \\ \text{si } \nu_i X_j y_j < 0 \text{ alors } \lim\limits_{\beta_i \to +\infty} A_i = +\infty. \end{cases}$

$\rightarrow$ Dans le cas où $\lim\limits_{\beta_i \to +\infty} A_i = +\infty$, il faut étudier la limite $\lim\limits_{\beta_i \to +\infty} B_i$. Les rôles de ν_i et β_i étant quasi-symétriques, nous retrouvons le même type de résultat que précédemment.
$\Rightarrow$ Ainsi, il vient donc finalement

$$\lim_{\beta_i \to +\infty} \log L_{cc}(\psi_G; y_j) = +\infty.$$

Calcul de la limite : $\lim\limits_{\beta_i \to -\infty} \log L_{cc}(\psi_G; y_j)$

Idem que le cas de $\lim\limits_{\beta_i \to +\infty} \log L_{cc}(\psi_G; y_j)$, sauf que les calculs montrent que les résultats sont inversés par rapport au signe de $\nu_i X_j y_j$. Ainsi, $\begin{cases} \text{si } \nu_i X_j y_j > 0 \text{ alors } \lim\limits_{\beta_i \to -\infty} A_i = +\infty, \\ \text{si } \nu_i X_j y_j < 0 \text{ alors } \lim\limits_{\beta_i \to -\infty} A_i = 0. \end{cases}$

Pour l'étude de B_i, nous avons : $\begin{cases} \text{si } \nu_k X_j y_j > 0 \text{ alors } \lim\limits_{\beta_i \to -\infty} b_i = 1, \\ \text{si } \nu_k X_j y_j < 0 \text{ alors } \lim\limits_{\beta_i \to -\infty} b_i = +\infty. \end{cases}$

$$\Rightarrow \text{Finalement, } \lim_{\beta_i \to -\infty} \log L_{cc}(\psi_G; y_j) = +\infty.$$

D.5 Mélange d'Inverses Gaussiennes

D.5.1 Décomposition de la log-vraisemblance L_{cc}

La log-vraisemblance classifiante conditionnelle $\log L_{cc}(\psi_G; y_j)$ peut s'écrire pour une observation y_j comme la somme de deux termes :

$$\log L_{cc}(\psi_G; y_j) = \log\Big(\overbrace{\sum_{i=1}^{G} \frac{\pi_i}{\sqrt{2\pi\sigma_i^2 y_j^3}} \exp\big(-\frac{1}{2}\frac{(y_j - \sqrt{\frac{1}{X_j\beta_i}})^2}{\frac{1}{X_j\beta_i}\sigma_i^2 y_j}\big)}^{A_i}\Big)^{A} + \underbrace{\underbrace{\sum_{i=1}^{G} \frac{\frac{\pi_i}{\sqrt{2\pi\sigma_i^2 y_j^3}} \exp\left(-\frac{1}{2}\frac{\left(y_j - \sqrt{\frac{1}{X_j\beta_i}}\right)^2}{\frac{1}{X_j\beta_i}\sigma_i^2 y_j}\right)}{\underbrace{\sum_{k=1}^{G} \frac{\pi_k}{\sqrt{2\pi\sigma_k^2 y_j^3}} \exp\left(-\frac{1}{2}\frac{\left(y_j - \sqrt{\frac{1}{X_j\beta_k}}\right)^2}{\frac{1}{X_j\beta_k}\sigma_k^2 y_j}\right)}_{b_i}} \log\left(\frac{\frac{\pi_i}{\sqrt{2\pi\sigma_i^2 y_j^3}} \exp\left(-\frac{1}{2}\frac{\left(y_j - \sqrt{\frac{1}{X_j\beta_i}}\right)^2}{\frac{1}{X_j\beta_i}\sigma_i^2 y_j}\right)}{\sum_{k=1}^{G} \frac{\pi_k}{\sqrt{2\pi\sigma_k^2 y_j^3}} \exp\left(-\frac{1}{2}\frac{\left(y_j - \sqrt{\frac{1}{X_j\beta_k}}\right)^2}{\frac{1}{X_j\beta_k}\sigma_k^2 y_j}\right)}\right)}_{B_i = b_i \log b_i}}_{B = \sum_i B_i}.$$

D.5.2 Calcul des limites et des cas critiques

Calcul de la limite : $\lim\limits_{\beta_i \to +\infty} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Nous remarquons par un simple calcul que $\lim\limits_{\beta_i \to +\infty} A_i = 0$, donc $\lim\limits_{\beta_i \to +\infty} A = K$.

$\rightarrow$ De plus, $\lim\limits_{\beta_i \to +\infty} B_i = \lim\limits_{\beta_i \to +\infty} b_i \log b_i = \lim\limits_{\beta_i \to +\infty} \frac{A_i}{A} \log \frac{A_i}{A} = 0$ car $A_i \to 0$ et $A \to K$.

On a donc $\lim_{\beta_i \to +\infty} B = K^{'}$, et la dérivée de l'entropie explose puisqu'elle n'est pas dérivable en 0 (or $\lim_{\beta_i \to +\infty} b_i = 0$).

$\Rightarrow$ Finalement, il vient

$$\lim_{\beta_i \to +\infty} \log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\beta_i \to +\infty} \frac{\partial \log L_{cc}(\psi_G; y_j)}{\partial \beta_i} = +\infty.$$

Calcul de la limite : $\lim_{\beta_i \to -\infty} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Tout d'abord, $\lim_{\beta_i \to -\infty} A_i = +\infty$. Il faut donc regarder comment se comporte B_i pour être fixé sur une éventuelle forme indéterminée dans la limite de la log-vraisemblance classifiante conditionnelle.
$\rightarrow$ La limite de B_i débouche sur une forme indéterminée, que nous levons en écrivant le terme b_i différemment :

$$b_i = 1 - \frac{\sum\limits_{\substack{k=1 \\ k \neq i}}^{G} \pi_k \exp\left[\frac{1}{\sigma_k^2}\left(-\frac{1}{2}\sigma_k^2 \log(2\pi\sigma_k^2 y_j^3) - \frac{1}{2}\frac{\left(y_j - \sqrt{\frac{1}{X_j\beta_k}}\right)^2}{\frac{y_j}{X_j\beta_k}}\right)\right]}{\sum_{k=1}^{G} \pi_k \exp\left[\frac{1}{\sigma_k^2}\left(-\frac{1}{2}\sigma_k^2 \log(2\pi\sigma_k^2 y_j^3) - \frac{1}{2}\frac{\left(y_j - \sqrt{\frac{1}{X_j\beta_k}}\right)^2}{\frac{y_j}{X_j\beta_k}}\right)\right]}.$$

Ainsi, $\lim_{\beta_i \to -\infty} b_i = 1 \Rightarrow \lim_{\beta_i \to -\infty} B_i = 0$. Nous avons donc $\lim_{\beta_i \to -\infty} B = K$.
Finalement

$$\lim_{\beta_i \to -\infty} \log L_{cc}(\psi_G; y_j) = +\infty.$$

Calcul de la limite : $\lim_{\sigma_i^2 \to 0^+} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Premièrement, remarquons que $\lim_{\sigma_i^2 \to 0^+} A_i = 0$ est une forme indéterminée. De ce fait il est possible de reformuler A_i pour éviter celle-ci, ce qui donne

$$A_i = \pi_i \exp\left[\frac{1}{\sigma_i^2}\left(-\frac{1}{2}\sigma_i^2 \log(2\pi\sigma_i^2 y_j^3) - \frac{1}{2}\frac{\left(y_j - \sqrt{\frac{1}{X_j\beta_i}}\right)^2}{\frac{y_j}{X_j\beta_i}}\right)\right].$$

Dans cette expression, il est relativement immédiat que deux cas se distinguent :
$$\begin{cases} \text{si } \dfrac{y_j}{X_j\beta_i} > 0 \text{ alors } \lim\limits_{\sigma_i^2 \to 0} A_i = 0, \\ \text{si } \dfrac{y_j}{X_j\beta_i} < 0 \text{ alors } \lim\limits_{\sigma_i^2 \to 0} A_i = +\infty. \end{cases}$$

$\rightarrow$ Reste donc à vérifier que la limite de B_i diverge dans le premier cas et converge dans le deuxième pour éviter une nouvelle forme indéterminée. En reprenant le précédent développement de b_i, nous trouvons :
$$\begin{cases} \text{si } \dfrac{y_j}{X_j\beta_i} > 0 \text{ alors } \lim\limits_{\sigma_i^2 \to 0} b_i = +\infty \Rightarrow \lim\limits_{\sigma_i^2 \to 0} B_i = +\infty, \\ \text{si } \dfrac{y_j}{X_j\beta_i} < 0 \text{ alors } \lim\limits_{\sigma_i^2 \to 0} b_i = 1 \Rightarrow \lim\limits_{\sigma_i^2 \to 0} B_i = 0. \end{cases}$$

Les résultats correspondent aux limites espérées, et nous avons donc au final :

$$\lim_{\sigma_i^2 \to 0^+} \log L_{cc}(\psi_G; y_j) = +\infty.$$

Calcul de la limite : $\lim_{\sigma_i^2 \to +\infty} \log L_{cc}(\psi_G; y_j)$

$\rightarrow$ Tout d'abord, $\lim_{\sigma_i^2 \to +\infty} A_i = 0$. Par conséquent, il vient $\lim_{\sigma_i^2 \to +\infty} A = K$.

$\rightarrow$ De plus, $\lim_{\sigma_i^2 \to +\infty} B_i = \lim_{\sigma_i^2 \to +\infty} b_i \log b_i = \lim_{\sigma_i^2 \to +\infty} \frac{A_i}{A} \log \frac{A_i}{A} = 0$.
On a donc $\lim_{\sigma_i^2 \to +\infty} B = K$, et la dérivée de l'entropie explose puisqu'elle n'est pas dérivable en 0 (or $\lim_{\sigma_i^2 \to +\infty} b_i = 0$).

$\Rightarrow$ Finalement, il vient

$$\lim_{\sigma_i^2 \to +\infty} \log L_{cc}(\psi_G; y_j) = K \quad \text{et} \quad \lim_{\sigma_i^2 \to +\infty} \frac{\partial \log L_{cc}(\psi_G; y_j)}{\partial \sigma_i^2} = +\infty.$$

Annexe E

Outil informatique - RExcel

FIGURE E.1 – Entrée des données et tâches par l'utilisateur.

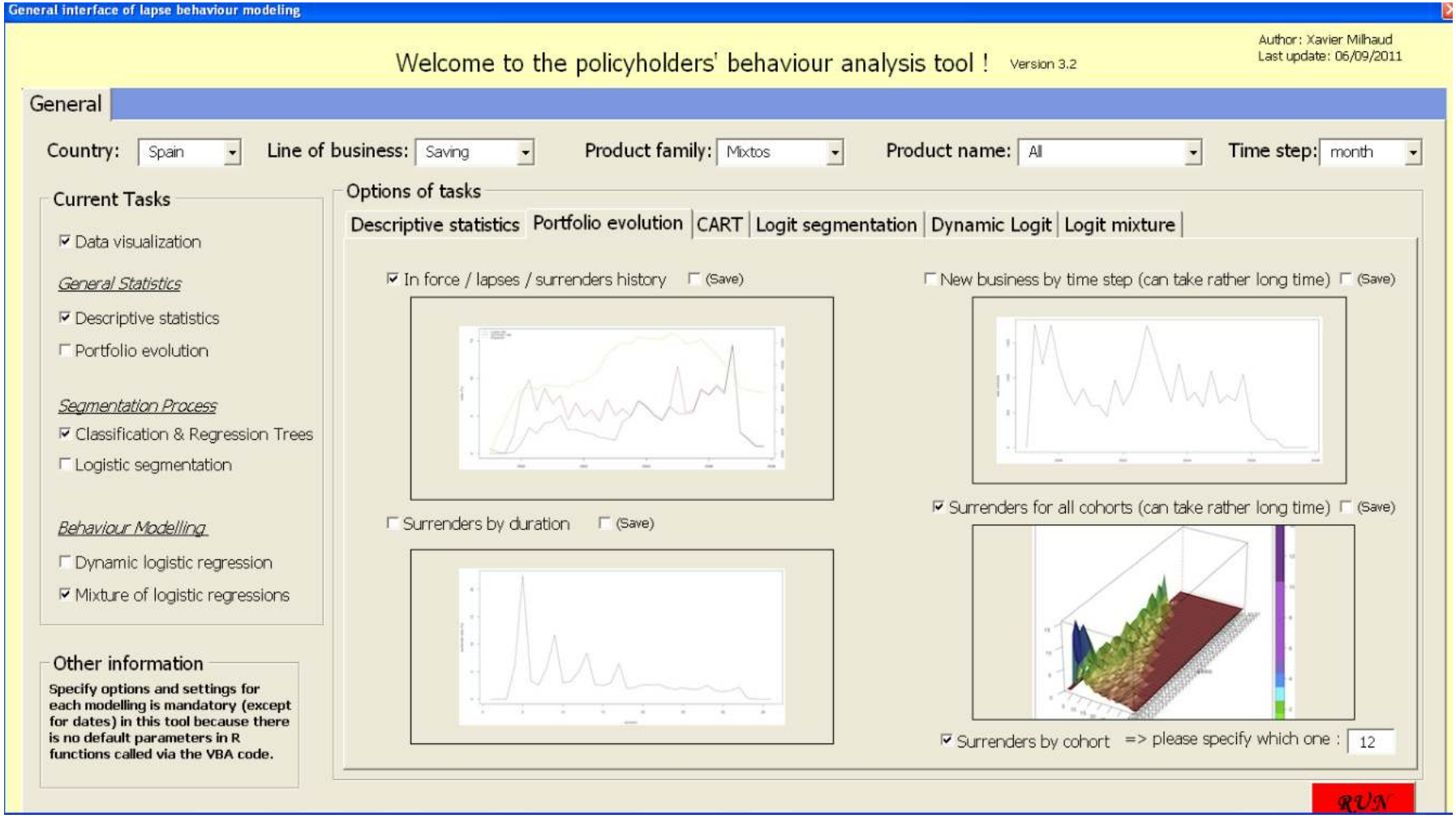


FIGURE E.2 – Exemple d'interface utilisateur.

FIGURE E.3 – Génération de la visualisation des données.

FIGURE E.4 – Génération des résultats des statistiques descriptives.

FIGURE E.5 – Génération de l'évolution du portefeuille, hors graphe.

FIGURE E.6 – Exposition des résultats de l'algorithme CART, hors graphe.

FIGURE E.7 – Génération des résultats des mélanges logistiques, hors graphe.