Identification de biomarqueurs prédictifs de la survie et de l'effet du traitement dans un contexte de données de grande dimension

Nils Ternes

NNT : 2016SACLS278

THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
L'UNIVERSITE PARIS-SUD

ÉCOLE DOCTORALE N° 570
EDSP Santé publique

Spécialité de doctorat : Santé publique - biostatistiques

Par

M. Nils TERNÈS

Identification de biomarqueurs prédictifs de la survie et de l'effet du traitement dans un contexte de données de grande dimension

Thèse présentée et soutenue à Villejuif, le Mercredi 5 Octobre 2016 :

Composition du Jury :

M. BROËT Philippe	Professeur, Université Paris Saclay	Président
Mme. BOULESTEIX Anne-Laure	Professeur, Université de Munich	Rapporteur
M. ROY Pascal	Professeur, Université Lyon 1	Rapporteur
M. GUEDJ Mickael	Docteur, Pharnext	Examinateur
M. CHIQUET Julien	Docteur, Université Paris Saclay	Examinateur
M. MICHIELS Stefan	Docteur, Université Paris Saclay	Directeur de thèse

Remerciements

Je remercie Pr. Philippe Bröet d'avoir accepté de présider mon jury de thèse. Je remercie également grandement Pr. Anne-Laure Boulesteix et Pr. Pascal Roy d'avoir accepté d'être rapporteurs de ma thèse et d'avoir fait le déplacement pour ma soutenance. Je les remercie pour tous leurs précieux commentaires et suggestions. Enfin, ma gratitude s'adresse également à Mickael Guedj et Julien Chiquet pour avoir accepté d'examiner ce travail.

Je tiens à remercier Stefan Michiels d'avoir dirigé ce travail pendant trois années. Je pense avoir eu beaucoup de chance d'être encadré par un directeur de thèse de cette qualité tant sur le plan professionnel que personnel. Mes craintes autour du doctorat se sont rapidement estompées au fur et à mesure que j'ai appris à te connaitre. Tu as toujours su me guider, m'écouter et tu m'as fait confiance. Ce fut un plaisir de travailler avec toi et je t'en remercie grandement !

Mes remerciements vont obligatoirement à Federico Rotolo pour m'avoir largement épaulé tout au long de ce travail. A mes yeux, tu as été comme un co-directeur étant donné le temps que tu m'as consacré, sans doute plusieurs centaines d'heures. Sans tes conseils et ton aide, je n'aurais sûrement pas pu aboutir à un travail comme celui-ci. En dehors d'un collègue hors pair, tu auras également été un ami, un confident, et je te remercie pour ton écoute.

Je remercie la Fondation Philanthropia Lombard-Odier d'avoir financé ce travail de thèse et Monsieur Luc Guiraud-Guigues, délégué de la fondation, pour avoir suivi l'avancement de celui-ci.

Bien évidemment, je remercie l'ensemble du service de biostatistique et d'épidémiologie de Gustave Roussy pour tout ce qu'ils ont fait pour moi à tous les niveaux. Je remercie Ellen Benhamou pour m'avoir accueilli au sein de son service et Emilie Lanoy pour m'avoir épaulé durant ma première année dans le service. Un grand merci à tous les juniors du service qui y rendent l'ambiance si agréable. Béranger, nous avons été deux à partager ces moments de thèse en même temps. Cela n'a pas toujours été simple mais l'on a su s'écouter et s'aider l'un et l'autre et je t'en remercie.

Un très grand merci à mon bureau si spécial. Ce sont des collègues de bureau comme ça qui m'ont donné l'envie de venir travailler au quotidien. Meumeu, je te remercie pour ton sourire et ta bonne humeur de tous les jours. Margie, je te remercie d'avoir été (trop souvent, je m'en

excuse) mon souffre-douleur et pour ces poképauses (un pokémon est caché dans le manuscrit...). Moon, je te remercie pour m'avoir persécuté (vol de téléphone, affaires par terre, blessure physique et morale). Heureusement que tu sais être sympa parfois ! Nath, merci pour avoir élargi mon vocabulaire aux expressions du 19ème siècle et au verlan,#aFondLesBallons, bigup à toi. Enfin, un profond merci à mes deux compères : Soso et Popo qui ont toujours été là pendant tout ce temps. Sophie, tu as toujours été notre maman du bureau. Je sais que j'ai toujours pu compter sur toi quand j'en avais besoin, toujours à l'écoute. Tu as su (souvent) canaliser mes moments de folie. Matthieu, je te considère comme un frère et je sais que je pouvais également compter sur toi à n'importe quel instant. A l'inverse de Sophie, tu étais là pour entretenir mes moments de folie et pour jouer (on fait un jeu ?). Rares sont ceux qui ont pris le temps de lire l'intégralité de mon manuscrit et vous l'avez même relu à plusieurs reprises. Je ne peux que vous remercier pour tout.

J'adresse, bien entendu, un immense merci à mes deux sœurs, mes parents et mes grands-parents qui ont eu un soutien sans faille à mon égard durant ces années de thèse mais également depuis bien plus longtemps. Ils se sont toujours intéressés à ce que je faisais en essayant de comprendre un maximum de choses, notamment ma grand-mère qui n'aura manqué aucune relecture de mémoires ou manuscrits et cela me touche beaucoup. Un remerciement spécial va à mes parents pour avoir fait le déplacement pour cette soutenance. Je n'ai jamais cessé de dire que ma principale motivation dans cette thèse aura été de pouvoir effectuer ma soutenance devant mes deux parents réunis et qu'ils soient fiers de moi. Je leur dédie cette thèse.

Enfin, outre l'apport scientifique, cette thèse aura également chamboulé ma vie personnelle avec la rencontre d'une personne qui est devenue très chère à mes yeux. Aurélie (bien que je n'aime pas t'appeler par ton prénom), je te remercie pour ta présence et ton soutien au quotidien qui ont été, qui sont, et qui seront toujours si importants pour moi. Tu as été la seule à voir « la partie cachée » de cette thèse, et tu auras été ma force et mon pilier surtout dans les moments très compliqués. Je te dois beaucoup.

Résumé de la thèse

Avec la révolution récente de la génomique et la médecine stratifiée, le développement de signatures moléculaires devient de plus en plus important pour prédire le pronostic (biomarqueurs pronostiques) ou l'effet d'un traitement (biomarqueurs prédictifs) de chaque patient. Cependant, la grande quantité d'information disponible rend la découverte de faux positifs de plus en plus fréquente dans la recherche biomédicale. La présence de données de grande dimension (nombre de biomarqueurs $\gg$ taille d'échantillon) soulève de nombreux défis statistiques tels que la non-identifiabilité des modèles, l'instabilité des biomarqueurs sélectionnés ou encore la multiplicité des tests.

L'objectif de cette thèse a été de proposer et d'évaluer des méthodes statistiques pour l'identification de ces biomarqueurs et l'élaboration d'une prédiction individuelle des probabilités de survie pour des nouveaux patients à partir d'un modèle de régression de Cox. Pour l'identification de biomarqueurs en présence de données de grande dimension, la régression pénalisée lasso est très largement utilisée. Dans le cas de biomarqueurs pronostiques, une extension empirique de cette pénalisation a été proposée permettant d'être plus restrictif sur le choix du paramètre λ dans le but de sélectionner moins de faux positifs. Pour les biomarqueurs prédictifs, l'intérêt s'est porté sur les interactions entre le traitement et les biomarqueurs dans le contexte d'un essai clinique randomisé. Douze approches permettant de les identifier ont été évaluées telles que le lasso (standard, adaptatif, groupé ou encore ridge+lasso), le boosting, la réduction de dimension des effets propres et un modèle implémentant les effets pronostiques par bras. Enfin, à partir d'un modèle de prédiction pénalisé, différentes stratégies ont été évaluées pour obtenir une prédiction individuelle pour un nouveau patient accompagnée d'un intervalle de confiance, tout en évitant un éventuel surapprentissage du modèle.

La performance des approches a été évaluée au travers d'études de simulation proposant des scénarios nuls et alternatifs. Ces méthodes ont également été illustrées sur différents jeux de données, contenant des données d'expression de gènes dans le cancer du sein.

Mots-clés : médecine stratifiée ; données de grande dimension ; régression pénalisée ; biomarqueurs pronostiques ; biomarqueurs prédictifs ; prédiction individuelle.

Abstract

With the recent revolution in genomics and in stratified medicine, the development of molecular signatures is becoming more and more important for predicting the prognosis (prognostic biomarkers) and the treatment effect (predictive biomarkers) of each patient. However, the large quantity of information has rendered false positives more and more frequent in biomedical research. The high-dimensional space (i.e. number of biomarkers $\gg$ sample size) leads to several statistical challenges such as the identifiability of the models, the instability of the selected coefficients or the multiple testing issue.

The aim of this thesis was to propose and evaluate statistical methods for the identification of these biomarkers and the individual predicted survival probability for new patients, in the context of the Cox regression model. For variable selection in a high-dimensional setting, the lasso penalty is commonly used. In the prognostic setting, an empirical extension of the lasso penalty has been proposed to be more stringent on the estimation of the tuning parameter λ in order to select less false positives. In the predictive setting, focus has been given to the biomarker-by-treatment interactions in the setting of a randomized clinical trial. Twelve approaches have been proposed for selecting these interactions such as lasso (standard, adaptive, grouped or ridge+lasso), boosting, dimension reduction of the main effects and a model incorporating arm-specific biomarker effects. Finally, several strategies were studied to obtain an individual survival prediction with a corresponding confidence interval for a future patient from a penalized regression model, while limiting the potential overfit.

The performance of the approaches was evaluated through simulation studies combining null and alternative scenarios. The methods were also illustrated in several data sets containing gene expression data in breast cancer.

Keywords: stratified medicine; high dimensional data; penalized regression; prognostic biomarkers; predictive biomarkers; individual prediction.

Production scientifique

Articles en lien avec la thèse

– Article publié

Ternès N, Rotolo F et Michiels S. *Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models*. Statistics in Medicine **2016**; 35(15):2561-2573. doi : 10.1002/sim.6927.

– Article accepté

Ternès N, Rotolo F, Heinze G et Michiels S. *Identification of biomarker-by-treatment interactions in high-dimensional spaces and survival outcomes*. Biometrical Journal **2016**.

– Article en cours d'écriture

Ternès N, Rotolo F et Michiels S. *Individual prediction of treatment outcome from high-dimensional Cox regression models in randomized controlled trials*.

– Articles annexes publiés

Ternès N, Arnedos M, Koscielny S, Michiels S et Lanoy E. *Statistical methods applied to omics data: predicting response to neoadjuvant therapy in breast cancer*. Current opinion in oncology **2014**; 26(6):576-583. doi : 10.1097/CCO.0000000000000134. (article de Master 2)

Michiels S, Ternès N et Rotolo F. *Statistical controversies in clinical research: prognostic gene signatures are not (yet) useful in clinical practice*. Annals of Oncology **2016**; In Press. doi : 10.1093/annonc/mdw307. (article de revue)

Communications orales

– ISCB 2014, Vienne, Autriche
Ternès N, Rotolo F et Michiels S. An extension of the lasso penalization to reduce false positive selection in high-dimensional Cox models.

– ADELF-EPITER 2014, Nice, France
Ternès N, Rotolo F et Michiels S. Régression pénalisée pour réduire la selection de faux positifs dans un modèle de Cox à haute dimension.

– EpiClin 2015, Montpellier, France
Ternès N, Rotolo F, Heinze G et Michiels S. Identification de biomarqueurs prédictifs dans un modèle de Cox à haute dimension.

– ICTMC 2015, Glasgow, Ecosse
Ternès N, Rotolo F, Heinze G et Michiels S. Prediction of treatment benefit in high-dimensional Cox models via gene signatures in randomized clinical trials.

– SMPGD 2016, Lille, France

Ternès N, Rotolo F, Heinze G et Michiels S. Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces.

Communications affichées

– ISCB 2015, Utrecht, Pays-Bas

Ternès N, Rotolo F, Heinze G et Michiels S. Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces.

– IBC 2016, Victoria, Canada

Ternès N, Rotolo F et Michiels S. Building a high-dimensional Cox regression model for predicting treatment outcome in a randomized controlled trial.

Distinction

– ISCB Student Conference Award 2014, Vienne, Autriche

Autres publications

– Bonvalot S, Ternès N, Fiore M, Bitsakou G, Colombo C, Honoré C, Marrari A, Le Cesne A, Perrone F, Dunant A et Gronchi A. *Spontaneous regression of primary abdominal wall desmoid tumors: more common than previously thought*. Annals of Surgical Oncology **2013**; 20(13):4096-4102. doi : 10.1245/s10434-013-3197-x.

– Deschamps F, Farouil G, Ternès N, Gaudin A, Hakime A, Tselikas L, Teriitehau C, Baudin E, Auperin A et de Baere T. *Thermal ablation techniques: a curative treatment of bone metastases in selected patients?* European Radiology **2014**; 24(8):1971-1980. doi : 10.1007/s00330-014-3202-1.

– Chevance A, Schuster T, Steele R, Ternès N et Platt RW. *Contour plot assessment of existing meta-analyses confirms robust association of statin use and acute kidney injury risk.* Journal of Clinical Epidemiology **2015**; 68(10):1138-1143. doi : 10.1016/j.jclinepi.2015.05.030.

– Dourthe ME, Ternès N, Gajda D, Paci A, Dufour C, Benhamou E et Valteau-Couanet D. *Busulfan-Melphalan followed by autologous stem cell transplantation in patients with high-risk neuroblastoma or Ewing sarcoma: an exposed-unexposed study evaluating the clinical impact of the order of drug administration.* Bone Marrow Transplantation **2016**. In press. doi : 10.1038/bmt.2016.109.

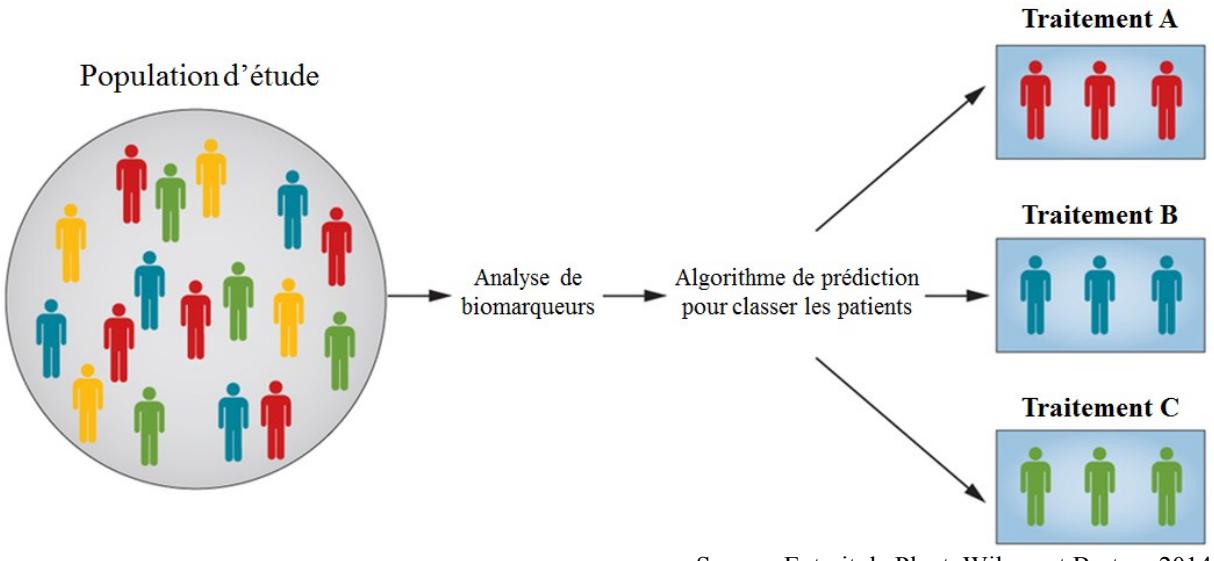
Table des matières

Table des figures

Liste des tableaux

Chapitre 1 Introduction

La médecine stratifiée ou encore de précision correspond, selon la définition de la Haute Autorité de Santé (2014), à « une approche thérapeutique dont l'objectif est de sélectionner les patients auxquels administrer un traitement en fonction d'un marqueur prédictif, afin de ne traiter que la sous-population susceptible de recevoir un bénéfice du traitement » (Figure 1.1).



Source. Extrait de Plant, Wilson et Barton, 2014

Figure 1.1 : Principe de la médecine de précision

Cette approche thérapeutique occupe aujourd'hui une place grandissante dans le monde de la santé, particulièrement en oncologie. Cependant, ce concept est loin d'être nouveau car dès

l'Antiquité, Hippocrate établissait le meilleur traitement pour ses patients en fonction d'un diagnostic basé sur quatre critères : la bile noire, la bile jaune, le flegme et le sang. Néanmoins, grâce aux nouveaux outils tels que la génétique ou les techniques de séquençage à haut débit, il est maintenant possible de séquencer l'intégralité du génome d'une tumeur afin d'en établir une cartographie. Le but étant, *in fine*, de mettre à disposition des médecins cette cartographie pour qu'ils puissent l'intégrer dans leur décision thérapeutique ou encore proposer à des patients d'intégrer des essais cliniques en vue de développer un traitement ciblé. Ces nouvelles données, dites génomiques, peuvent être de différentes natures telles que les mutations, le nombre de copies aberrantes ou encore l'expression des gènes et sont appelées biomarqueurs. Selon la définition officielle de l'institut américain de la santé (*NIH*), un biomarqueur est « une caractéristique mesurable objectivement qui représente un indicateur des processus biologiques normaux ou pathologiques ou de réponse pharmacologique à une intervention thérapeutique » (NIH Biomarkers Definitions Working Group, 2001). Ces biomarqueurs sont généralement classés en deux groupes : les biomarqueurs pronostiques et les biomarqueurs prédictifs (Figure 1.2). Un biomarqueur est considéré comme pronostique s'il permet de discriminer les patients selon leur niveau de risque (e.g. haut ou bas) indépendamment d'un quelconque traitement. Ainsi, la Figure 1.2a montre un exemple de biomarqueur pronostique : pour les patients non traités (courbes pointillées), la survie est différente selon le biomarqueur (courbe verte vs. courbe rouge) traduisant que celui-ci est capable de prédire le cours naturel probable de la maladie (ici, la survie). De plus, l'effet relatif du traitement (courbes pleines vs. courbes pointillées) est similaire pour les deux groupes de risque. Quelques exemples de biomarqueurs pronostiques en oncologie sont l'antigène Ki67 (marqueur de prolifération) ou les récepteurs hormonaux (HER2 et ER). Il existe aussi des signatures pronostiques combinant plusieurs biomarqueurs, c'est le cas par exemple de la signature d'Amsterdam dans le cancer du sein basée sur 70 gènes (*MammaPrint*, van't Veer et al., 2002). Un biomarqueur est considéré comme prédictif ou modificateur de l'effet du traitement si l'effet relatif du traitement varie en fonction de celui-ci. Ainsi, le traitement est par exemple bénéfique uniquement pour les patients ayant un score bas du biomarqueur, alors qu'il est moins bénéfique (Figure 1.2b, interaction quantitative) ou délétère (Figure 1.2c, interaction qualitative) pour les patients ayant un score élevé du biomarqueur.

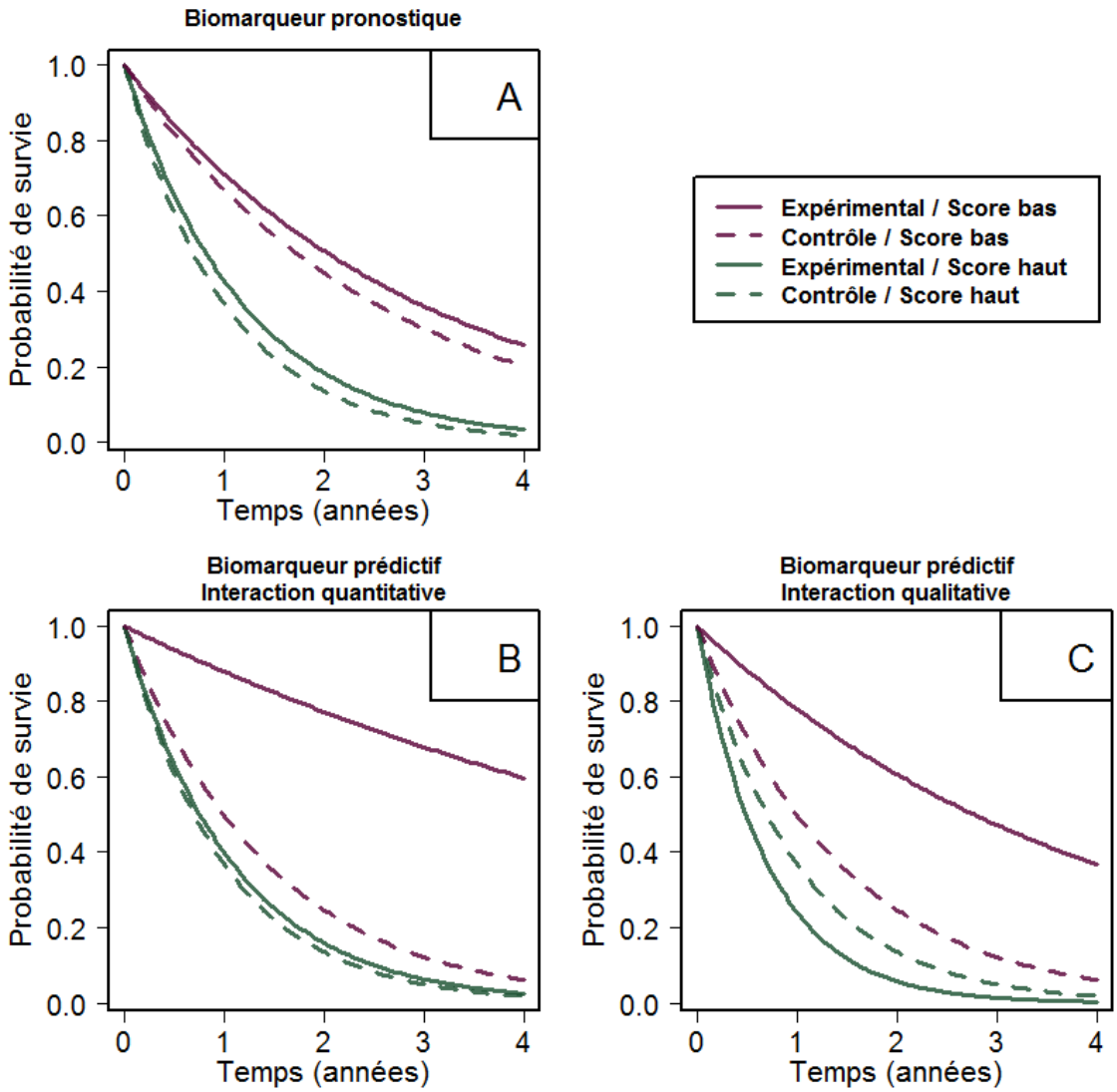


Figure 1.2 : Différents types de biomarqueurs

Pour une interaction quantitative, l'effet du traitement est différent selon le score du biomarqueur mais reste dans la même direction. C'est le cas, par exemple, de la mutation KRAS dans le cancer du côlon (Amado et al., 2008). Bien qu'il ait été montré que cette mutation permet de prédire l'effet des anticorps monoclonaux anti-EGFR (e.g. cetuximab ou panitumumab) sur la survie sans progression, les résultats sont différents en fonction de la mutation KRAS. En effet, le bénéfice de ces anticorps n'a été montré que chez les patients mutés KRAS (Hazard Ratio (HR) = 0,45, Intervalle de Confiance à 95% (IC95%) : 0,34–0,59) et non chez les patients non mutés KRAS (HR = 0,99, IC95% : 0,73–1,36) avec un intervalle de confiance plus large. A l'inverse, la direction de l'effet du traitement est différente selon le score du biomarqueur pour une interaction qualitative, comme par exemple le récepteur EGFR dans le cancer du poumon non à petites cellules (Mok et al., 2009). Bien

que le gefitinib ait montré un bénéfice global par rapport à la combinaison carboplatin plus paclitaxel sur la survie sans progression (HR = 0,74, IC95% : 0,65–0,85), l'effet varie fortement selon le récepteur EGFR. Le bénéfice est bien réel chez les patients ayant une mutation EGFR (HR = 0,48, IC95% : 0,36–0,64), cependant, le traitement semble délétère chez les patients non mutés (HR = 2,85, IC95% : 2,05–3,98). Dans le cas prédictif, il existe aussi des signatures combinant plusieurs biomarqueurs telles que des signatures de 8 et 14 biomarqueurs prédictifs de l'effet du trastuzumab dans le cancer du sein (Perez et al., 2015; Pogue-Geile et al., 2013) ou encore une signature de 84 gènes pour prédire l'effet de l'immunothérapie ciblant MAGE-A3 dans le mélanome ou le cancer du poumon non à petites cellules (Ulloa-Montoya et al., 2013).

Jusqu'à présent, la littérature biomédicale répertorie plus de 150000 articles documentant plusieurs dizaines de milliers de biomarqueurs. Cependant, moins d'une centaine sont actuellement validés dans la pratique clinique (Poste, 2011). En effet, selon les recommandations de l'*ESMO* (*European Society of Medical Oncology*) de 2014, moins de vingt biomarqueurs pronostiques et/ou prédictifs dans le cancer du poumon, du côlon et de la prostate ont un niveau de preuve suffisant (Schneider et al., 2015). De plus, en raison de l'acquisition simple et relativement peu coûteuse d'un nombre très important de données génomiques, il n'a jamais été aussi facile qu'aujourd'hui de générer des faux positifs pouvant avoir de forts impacts scientifiques (MacArthur, 2012). Bien souvent, l'apparition de faux positifs est causée par l'utilisation non appropriée de méthodes statistiques ou encore la mauvaise ou non validation de ces résultats. Ce dernier point a été illustré par Michiels, Koscielny et Hill (2005) grâce à la nouvelle analyse des données provenant des sept plus grandes études publiées visant à prédire le pronostic des patients. Les résultats ont montré que cinq des sept études ne prédisaient le pronostic des patients pas mieux que le hasard. Il devient donc urgent de définir des méthodologies statistiques permettant de limiter le nombre de faux positifs mais aussi de proposer des techniques permettant de valider les biomarqueurs/signatures identifiés. Dans ce contexte de forte expansion des données génomiques et de réflexion autour de la médecine de précision, au moins deux objectifs ont un intérêt particulier : sélectionner les biomarqueurs qui ont réellement un rôle d'un point de vue biologique et sélectionner les patients susceptibles de bénéficier d'une thérapie. Ainsi, au travers des deux premiers axes de mon travail de thèse, je me suis focalisé sur le premier objectif visant à identifier des biomarqueurs prédictifs de la survie communément appelés pronostiques en cancérologie (1er axe, Chapitre 3), ou des biomarqueurs prédictifs de l'effet

d'un traitement (2ème axe, Chapitre 4) tout en limitant la sélection de faux positifs. Pour faire de la sélection de variables, différentes approches ont été proposées dans la littérature et je me suis essentiellement focalisé sur la pénalisation lasso qui, comme nous le verrons, est une approche très populaire notamment en présence de données de grande dimension (Chapitre 2). Dans le cas pronostic, une extension de cette pénalisation appelée lasso-*pcvl* a été proposée pour réduire le nombre de faux positifs sélectionnés sans fortement augmenter le nombre de faux négatifs. Dans le cas prédictif, différents modèles et pénalisations ont été proposés pour sélectionner les interactions entre les biomarqueurs et le traitement en présence de leur effet propre. Le troisième axe de ma thèse avait pour objectif d'obtenir, pour chaque patient, une prédiction individuelle accompagnée d'une mesure d'incertitude permettant d'évaluer s'il est susceptible de bénéficier du traitement (Chapitre 5). Pour cela, différentes stratégies ont été évaluées telles que le choix du modèle, l'utilisation de la technique de validation croisée pour limiter le surapprentissage du modèle ou encore l'utilisation du bootstrap dans la construction des intervalles de confiance. Pour chaque axe, une étude de simulation a été mise en place pour évaluer *in silico* la performance des approches discutées sous différents scénarios. Ces approches ont également été appliquées sur différents jeux de données chez des patientes atteintes d'un cancer du sein et pour lesquelles des données d'expression étaient disponibles : deux bases de données rétrospectives de 523 et 614 patientes et un essai clinique randomisé évaluant l'effet du trastuzumab chez 1574 patientes.

Chapitre 2 Données de grande dimension : limites statistiques et alternatives

2.1 Limites statistiques

Classiquement, dans un modèle de régression, l'objectif est d'estimer des coefficients de régression $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ permettant d'évaluer la force d'association entre des prédicteurs $\boldsymbol{X}$ et un critère de jugement. Dans ce manuscrit, un critère de jugement censuré (i.e. variable de durée, e.g. la survie globale) sera considéré mais l'ensemble du discours est généralisable à d'autres types de critères de jugement (e.g. continus, binaires, etc.). Lorsque l'on s'intéresse à la survie, le modèle de régression le plus couramment utilisé est le modèle à risques proportionnels :

$$h(t, \boldsymbol{X}) = h_0(t)\exp(\boldsymbol{\beta X}) \tag{2.1}$$

avec $h(t, \boldsymbol{X})$ correspondant au risque instantané de survenue d'événement chez un patient ayant le profil $\boldsymbol{X}$ au temps *t*, et $h_0(t)$ le risque de base au même temps. A partir de ce modèle, il est également possible d'estimer la probabilité de survie d'un patient au temps *t* :

$$S(t, \boldsymbol{X}) = \exp\left(-\int_0^t h(u, \boldsymbol{X})\, du\right). \tag{2.2}$$

Généralement, les coefficients de régression $\boldsymbol{\beta}$ sont estimés tels que la fonction de vraisemblance du modèle soit maximale pour les données observées. L'approche d'estimation la plus souvent utilisée est l'approche semi-paramétrique de Cox (Cox, 1972) qui maximise la fonction de log-vraisemblance partielle $l(\boldsymbol{\beta}, \boldsymbol{X})$ qui ne nécessite pas d'estimer la fonction de risque de base $h_0(t)$, et s'écrit :

$$l(\boldsymbol{\beta}, \boldsymbol{X}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta X_i} - \log\left(\sum_{j \epsilon R_i} \exp(\boldsymbol{\beta X_j}) \right) \right] \quad (2.3)$$

avec *n* le nombre total de patients, δ_i l'indicateur de censure (0, si censure et 1, si événement) et R_i l'ensemble des individus encore à risque (i.e. sans événement ni censure) juste avant le temps t_i. Analytiquement, il est possible de trouver une solution unique $\boldsymbol{\beta}$ qui maximise la fonction $l(\boldsymbol{\beta}, \boldsymbol{X})$ et cela uniquement si le nombre de paramètres à estimer *p* est inférieur au nombre d'évènements parmi les observations *n*. Or, avec la révolution récente de la génomique, la quantité d'information disponible (biomarqueurs) a fortement augmenté pour atteindre une situation où $p \gg n$ (on parle alors de données de grande dimension, avec généralement $p > 10^4$ et $n < 10^3$) et donc le modèle devient ainsi non identifiable (i.e. impossibilité de trouver une solution unique pour $\boldsymbol{\beta}$).

En dehors de ce problème de non-identifiabilité du modèle, se pose la question de la sélection de variables. En effet, même si le modèle contenant l'ensemble des biomarqueurs admettait une solution unique pour estimer les coefficients $\boldsymbol{\beta}$, cela aurait peu de pertinence d'un point de vue clinique car aucune interprétation ne pourrait en être tirée. L'objectif clinique est de chercher à sélectionner les biomarqueurs qui ont un véritable effet et de laisser de côté les biomarqueurs qui n'en ont pas. D'autres problématiques telles que la multiplicité des tests (i.e. augmentation du risque de conclure à tort à la significativité d'un test : faux positif) ou encore l'instabilité des coefficients estimés, peuvent également être soulevées dans ce contexte.

2.2 Alternatives possibles

Jusqu'alors, de nombreuses approches ont été proposées et sont utilisées dans la littérature pour le cas où le nombre de biomarqueurs candidats est très élevé. Pour avoir un aperçu, nous

en présentons ici cinq familles dont une, la régression pénalisée, est détaillée plus longuement (section 2.3) car centrale dans ce travail de thèse.

Régression univariée et multivariée. L'approche univariée consiste à étudier indépendamment chaque biomarqueur en évaluant sa force d'association avec le critère de jugement dans un modèle de régression. Les biomarqueurs sont ensuite classés selon leur degré d'association (à partir d'un test de Wald, un score ou encore un score modifié comme proposé par Tusher, Tibshirani et Chu (2001) avec l'algorithme *SAM*) et un seuil est fixé pour identifier un sous-groupe de biomarqueurs (e.g. top k des plus fortes associations). Cependant, la multiplicité des tests (p tests) peut rendre cette approche peu puissante. De plus, bien que très simple à implémenter, cette approche ne tient pas compte de la corrélation entre les biomarqueurs, ce qui peut être une limite importante dans ce contexte de données génomiques. En effet, plusieurs biomarqueurs fortement corrélés peuvent être sélectionnés indépendamment les uns des autres, mais leur combinaison peut être sous-optimale en raison de la redondance d'information. Pour pallier cela, il est possible d'opter d'emblée pour une approche multivariée qui permet de tenir compte de la dépendance entre les biomarqueurs. Bien entendu, il n'est pas possible de considérer un modèle multivarié contenant initialement l'ensemble des biomarqueurs car $p \gg n$. La technique de sélection la plus courante est la sélection pas-à-pas ascendante qui consiste à débuter d'un modèle nul et à y inclure, à chaque itération, le biomarqueur qui y apporte le plus d'information statistique. Un seuil doit également être fixé, comme par exemple le nombre de biomarqueurs à inclure ou une quantité minimale d'ajout d'information, pour arrêter le processus et ainsi faire de la sélection. Une des limites de cette approche est que le modèle retenu dépend fortement des premiers biomarqueurs qui y sont inclus ou que le temps de calcul peut être extrêmement long si p est grand.

Réduction de dimension. Le principe de cette famille de méthodes est de faire une synthèse de l'information contenue dans un grand nombre de variables au travers de nouvelles variables appelées composantes. Ces composantes, indépendantes entre elles, sont des combinaisons linéaires des variables initiales permettant de garder la plus grande part de variabilité possible. Dans un contexte de grande dimension ($p \gg n$), l'estimation d'un nombre relativement petit de composantes (C, avec $C < n \ll p$) permet d'utiliser dans de meilleures conditions des techniques multivariées classiques et ainsi, effectuer de la prédiction. En revanche, les composantes calculées au travers de cette réduction de dimension ne sont pas utilisables d'un

point de vue de sélection de biomarqueurs car celles-ci sont le résultat d'une combinaison de plusieurs, voire l'ensemble des biomarqueurs. Une alternative peut être d'utiliser une technique de sélection de variables en amont de la réduction de dimension (e.g. analyse en composante principale supervisée). Il existe différentes techniques permettant de réaliser une réduction de dimension et certaines d'entre elles sont discutées au cours de ce manuscrit (e.g. analyse en composante principale ou régression des moindres carrés partiels).

Approches causales. On parle de lien de causalité entre deux variables lorsque ce lien ne peut pas être causé par de potentiels facteurs de confusion. Par exemple, dans un essai randomisé, le traitement est alloué aléatoirement chez les patients lui permettant ainsi d'être indépendant des autres covariables (mesurées ou non). Ainsi, l'association entre le traitement et le critère de jugement peut être interprétée comme un effet causal. Dans le cas des données génomiques, il n'est bien entendu pas possible de faire d'intervention sur le génome humain. On se place donc toujours dans une situation observationnelle. Les diagrammes causaux, appelés également graphiques acycliques dirigés (*directed acyclic graphs* ou *DAG*) peuvent être utilisés pour inférer des effets causaux en situation non interventionnelle (Pearl, 2000). Ces graphiques représentent les relations ordonnées entre les variables : une flèche entre deux variables signifie qu'une relation causale entre elles est possible. A l'inverse, une absence de flèche signifie avec certitude qu'il n'y a pas d'effet causal. En théorie, ces diagrammes doivent être exhaustifs de l'ensemble des facteurs de confusion possibles pour évaluer les biais potentiels et estimer les effets causaux. Toutefois, dans un contexte de grande dimension, les relations entre les biomarqueurs sont généralement inconnues et les diagrammes ne peuvent pas être tracés. Des approches causales ont alors été proposées dans ce contexte. C'est le cas de l'*IDA* (signifiant *intervention-calculus when the DAG is absent*, Maathuis, Kalisch et Bühlmann, 2009) qui vise à reconstruire le squelette du *DAG* sous-jacent à partir de tests statistiques (algorithme PC, Kalisch et Bühlmann, 2007) et dans lequel les flèches peuvent être unidirectionnelles ou bidirectionnelles. Ces dernières ne permettent pas d'obtenir un *DAG* unique mais plutôt un ensemble de *DAG* traduisant l'indépendance conditionnelle provenant des données. Les effets causaux de chaque biomarqueur sur le critère de jugement sont estimés pour chaque *DAG* possible et l'effet retenu correspond au plus petit des effets estimés. Cette approche ne permet donc pas de faire directement de la sélection de variables, ou le cas échéant il est nécessaire de fixer un seuil en amont (e.g. effet minimum ou top k des biomarqueurs). Cependant, il a été montré que cette approche a tendance à être très instable en matière de sélection et d'estimation, et des extensions

permettant de stabiliser les résultats ont été proposées. C'est le cas du *causal stability ranking* (*CStaR*, Stekhoven et al., 2012) dont la technique de stabilisation, appelée *stability selection*, sera présentée dans la suite du manuscrit dans un contexte de régression pénalisée (section 3.3.2). Nous avons récemment appliqué ces approches causales chez des patientes atteintes d'un cancer du sein et traitées par letrozole afin d'identifier les gènes les plus fortement associés ou « causales » de la réduction de la prolifération tumorale (Ternès et al., 2014, travail de Master 2). Dans cette application, les approches causales n'apportent pas d'information supplémentaire aux régressions pénalisées et sont très coûteuses en temps de calcul.

2.3 La régression pénalisée

Ce type de régression, souvent utilisé dans le cadre de données de grande dimension, consiste à introduire un terme de pénalisation $p(\cdot)$ dans l'écriture de la vraisemblance du modèle $l(\cdot)$ (vraisemblance partielle pour le modèle de Cox). Cette pénalisation permet d'estimer des coefficients de régression qui sont moins variables mais légèrement biaisés (compromis entre variance et biais). Cette pénalisation est basée sur les coefficients de régression et dépend généralement d'un unique paramètre, positif ou nul, que l'on note λ. La vraisemblance pénalisée l_p s'écrit donc :

$$l_p(\lambda, \boldsymbol{\beta}, \boldsymbol{X}) = l(\boldsymbol{\beta}, \boldsymbol{X}) - p(\lambda, \boldsymbol{\beta}). \quad (2.4)$$

Un des objectifs principaux des régressions pénalisées est de contraindre les coefficients de régression à tendre vers zéro. Ainsi, plus la pénalisation λ est forte, plus la contrainte est forte. A l'inverse, si λ est nul le modèle est non pénalisé. Jusqu'à présent, des pénalisations de différentes natures ont été proposées dans la littérature dont quelques-unes sont présentées ci-dessous par ordre chronologique.

Le ridge. Cette pénalisation a été introduite par Hoerl et Kennard (1970). Elle correspond à la norme L2 des coefficients de régression. En effet, la pénalisation ridge peut s'écrire :

$$p(\lambda, \boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \beta_j^2.$$

Cette pénalisation permet de contraindre les coefficients de régression en les faisant tendre vers zéro au fur et à mesure que λ augmente. Cependant, les coefficients ne sont jamais strictement égaux à zéro. Il n'est donc pas possible de faire de la sélection de variables, ce qui peut être un important inconvénient dans un contexte de données de grande dimension. En effet, si l'objectif est d'obtenir une signature de gènes, il est préférable d'avoir une liste réduite de biomarqueurs. En revanche, la pénalisation ridge a montré de bonnes performances en ce qui concerne la prédiction (Bøvelstad et al., 2007).

Le bridge. La régression bridge a été introduite par Frank et Friedman (1993) et correspond à une famille particulière des régressions pénalisées. La pénalisation bridge peut s'écrire :

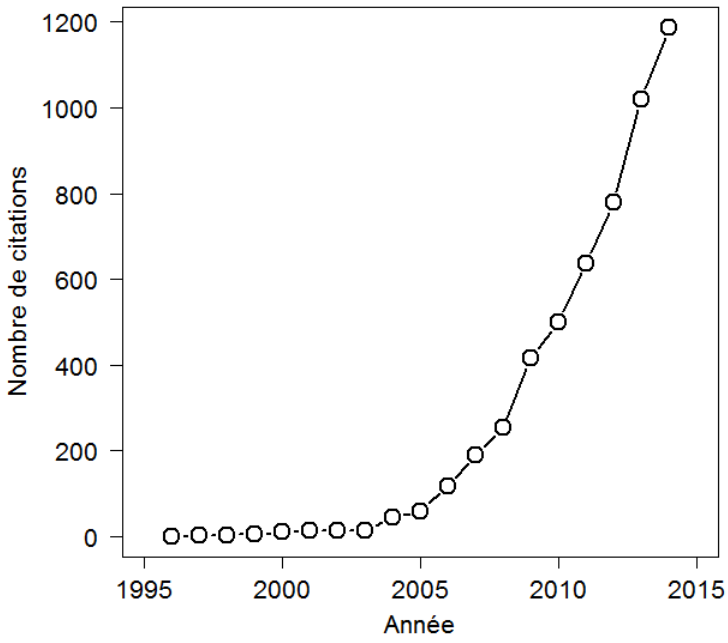
$$p(\lambda, \omega, \boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|^{\omega}$$

avec $\omega > 0$. La combinaison optimale de ω et λ est habituellement réalisée à partir d'une validation croisée et pour une grille prédéfinie de valeurs. Le bridge permet de faire de la sélection de variables pour tout $0 < \omega \leq 1$ puisque certains coefficients deviennent strictement égaux à zéro. Huang, Ma et Zhang (2008) ont montré que dans un contexte de données de grande dimension, le bridge a la propriété *oracle* pour tout $0 < \omega < 1$. La propriété *oracle* décrite par Fan et Li (2001), puis par Fan et Peng (2004), signifie pour un échantillon de taille infinie que (i) l'estimateur identifie correctement les coefficients non nuls avec une probabilité convergeant vers 1 et que (ii) ces coefficients sont asymptotiquement normaux avec les mêmes moyennes et covariances que si les coefficients nuls étaient connus par avance (i.e. modèle non pénalisé contenant uniquement les coefficients non nuls). La pénalisation ridge correspond au cas particulier de la pénalisation bridge où $\omega = 2$.

Le lasso. Tout comme la pénalisation ridge, la pénalisation lasso (ou ***l**east **a**bsolute **s**hrinkage and **s**election **o**perator*) introduite par Tibshirani (1996, 1997) est un cas particulier de la pénalisation bridge avec $\omega = 1$ et correspond à la norme L1 des coefficients de régression. On note donc cette pénalisation :

$$p(\lambda, \boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|.$$

Comme l'indique son acronyme, le lasso permet de contraindre les coefficients en les faisant tendre vers zéro au fur et à mesure que λ augmente (***s**hrinkage*), jusqu'à même devenir strictement égaux à zéro pour des valeurs élevées de λ, ce qui correspond à de la sélection de variables (***s**election*).



Source : base de données Web of science

Figure 2.1 : Nombre de citations de l'article de référence de la pénalisation lasso par année

Ainsi, le nombre de biomarqueurs restant dans le modèle (p_λ) pour un certain λ tend à décroître au fur et à mesure que ce paramètre augmente. Il s'agit d'une pénalisation très largement utilisée dont la mise en œuvre ne cesse de croître au cours de ces dernières années (Figure 2.1). Une rétrospective de l'évolution de cette pénalisation a d'ailleurs été faite récemment par son auteur (Tibshirani, 2011). Dans cette thèse, nous nous sommes focalisés sur cette pénalisation pour effectuer de la sélection de biomarqueurs pronostiques et/ou prédictifs. Il a cependant été montré que le lasso est sous-optimal, notamment en raison de son non-respect de la propriété *oracle* ou encore de ses lacunes en présence de fortes corrélations puisqu'il a tendance à ne sélectionner arbitrairement qu'un seul biomarqueur parmi l'ensemble des biomarqueurs corrélés. Ces points seront discutés dans ce manuscrit.

Le scad. La pénalisation scad (ou ***s**moothly **c**lipped **a**bsolute **d**eviation*, Fan et Li, 2001) a été proposée pour satisfaire à la propriété *oracle* (contrairement au lasso). Elle peut s'écrire :

$$p(\lambda, \kappa, \boldsymbol{\beta}) = \sum_{j=1}^{p} p_j(\lambda, \kappa, \beta_j) \ avec \ p_j(\lambda, \kappa, \beta_j) = \begin{cases} \lambda|\beta_j|, \ si \ |\beta_j| \leq \lambda \\ -\dfrac{|\beta_j|^2 - 2\kappa\lambda|\beta_j| + \lambda^2}{2(\kappa - 1)}, \ si \ \lambda < |\beta_j| \leq \kappa\lambda \\ \dfrac{(\kappa + 1)\lambda^2}{2}, \ si \ |\beta_j| > \kappa\lambda \end{cases}$$

avec $\kappa > 2$. Il s'agit d'une fonction spline quadratique avec deux nœuds à λ et $\kappa\lambda$ qui permet aux coefficients ayant de fortes valeurs de n'être que très légèrement pénalisés. La pénalisation scad nécessite l'estimation de deux paramètres (κ et λ), ce qui peut avoir des effets négatifs. Plutôt que de chercher la meilleure combinaison (κ, λ), Fan et Li (2001, 2002) recommandent de fixer κ = 3,7 et de n'optimiser que le paramètre λ. Il a néanmoins été montré que le scad est très compliqué à mettre en œuvre d'un point de vue calculatoire. Des processus itératifs ont été proposés pour pallier ce problème, cependant, aucun consensus n'est encore établi et les résultats sont très sensibles à ce choix (Benner et al., 2010).

L'elastic net. Cette pénalisation a été introduite plus tardivement par Zou et Hastie (2005) et correspond à la somme des pénalisations lasso et ridge décrites ci-dessus. On la note :

$$p(\lambda, \lambda_2, \boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

avec respectivement λ et λ_2 les paramètres de pénalisation du lasso et du ridge. La pénalisation elastic net vise à combiner les avantages de ces deux pénalisations. Tout comme le lasso, l'elastic net permet de faire de la sélection de variables. De plus, l'ajout de la pénalisation ridge permet à l'elastic net de sélectionner plus de variables et d'être plus performant que la pénalisation lasso seule en matière de prédiction, notamment dans le cas où les variables sont très corrélées entre elles.

Le *mcp*. Cette pénalisation (ou ***m**inimax **c**oncave **p**enalty*, Zhang, 2010) est très semblable à la pénalisation scad puisqu'il s'agit d'une pénalisation concave ou non convexe ayant pour objectif d'éliminer les variables non importantes du modèle tout en gardant les variables importantes non pénalisées. Tout comme le scad, le mcp garantit la propriété *oracle* mais n'est pas simple à mettre en œuvre. On note cette pénalisation :

$$p(\lambda, \kappa, \boldsymbol{\beta}) = \sum_{j=1}^{p} p_j(\lambda, \kappa, \beta_j) \ \ avec \ \ p_j(\lambda, \kappa, \beta_j) = \begin{cases} \lambda\beta_j - \dfrac{\beta_j^2}{2\kappa}, & si\ \beta_j \leq \kappa\lambda \\ \dfrac{1}{2}\kappa\lambda^2, & si\ \beta_j > \kappa\lambda \end{cases}$$

avec $\kappa > 0$ qu'il est suggéré de fixer à 2,7 par Zhang (2010).

Chapitre 3 Identification de biomarqueurs prédictifs de la survie

Parmi les méthodes de sélection existantes, la pénalisation lasso est sans doute un bon compromis entre simplicité d'utilisation et bonnes propriétés statistiques, ce qui en fait l'une des plus utilisées dans les applications. Cependant, la technique la plus couramment utilisée pour choisir le paramètre de pénalisation – la validation croisée – est connue pour ne pas être suffisamment conservatrice, produisant un nombre élevé de faux positifs. C'est pourquoi, dans la première partie de mon projet de thèse, je me suis focalisé sur le choix de paramètre de pénalisation λ du lasso notamment dans le cadre de l'identification de biomarqueurs prédictifs de la survie, appelés biomarqueurs pronostiques pour la suite de la lecture (Ternès, Rotolo et Michiels, 2016a).

3.1 Estimation du paramètre de pénalisation λ du lasso

Comme dit précédemment, l'ajout d'une pénalisation lasso dans un modèle de régression permet de faire de la sélection de variables de par la norme L1 des coefficients de régression. Le degré de parcimonie ou de complexité du modèle dépend du paramètre de pénalisation λ. Ce paramètre varie de 0 (modèle complet comprenant tous les biomarqueurs) à $+\infty$ (modèle nul ne comprenant aucun biomarqueur). Il est donc important d'estimer correctement ce paramètre afin d'identifier le sous-groupe de biomarqueurs retenus dans le modèle.

Classiquement, ce paramètre λ est estimé à partir d'une validation croisée. Cette technique, introduite par Stone (1974) et basée sur le ré-échantillonnage des données observées, permet d'évaluer l'erreur de prédiction et la qualité de l'ajustement d'un modèle en utilisant des sous-ensembles aléatoires des données. L'utilisation la plus courante de la validation croisée est de commencer par diviser l'échantillon en *K* sous-échantillons de taille comparable (on parle de *K-fold cross-validation*). L'idée de la validation croisée est d'imiter une validation externe qui consisterait à développer un modèle sur des données (échantillon d'apprentissage), puis à l'évaluer sur des données n'ayant pas servi à sa construction (échantillon de validation) pour limiter une quelconque sur-évaluation du modèle. Dans le cas de la validation croisée en *K* sous-échantillons, un échantillon d'apprentissage est constitué à partir de $K - 1$ des sous-échantillons et le sous-échantillon restant est considéré comme un échantillon test. Ce processus est répété *K* fois de façon à ce que chaque sous-échantillon soit considéré une fois comme échantillon test. Pour chaque itération $k \in \{1, \dots, K\}$, une première étape consiste à estimer, à partir des données de l'échantillon d'apprentissage $\boldsymbol{X}_{-k}$, les coefficients de régression $\boldsymbol{\beta}_{-k}$ en maximisant la log-vraisemblance pénalisée l_p. Généralement, ces coefficients sont dans un second temps affectés aux données de l'échantillon de validation $\boldsymbol{X}_k$ afin d'y calculer la log-vraisemblance pénalisée $l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}_k)$ et d'évaluer la contribution de cette log-vraisemblance sur la log-vraisemblance dans l'échantillon complet $\boldsymbol{X}$. En effet, dans un modèle linéaire ou encore logistique, la relation $l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}_k) = l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}) - l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}_{-k})$ est vraie car la contribution d'un individu ne dépend que de ses propres données. Ainsi, les différentes log-vraisemblances sont indépendantes entre elles. L'idée étant que plus la contribution de cette log-vraisemblance est forte et plus le modèle estimé à partir des données d'apprentissage sera de bonne qualité. En revanche, et comme le soulignent Verweij et van Houwelingen (1993, 1994), lorsque l'on s'intéresse au modèle de Cox, la contribution d'un individu *i* dépend de ses données mais également de celles des autres au travers de R_i, l'ensemble des individus à risque juste avant le temps t_i (voir Chapitre 2). Ainsi, les différentes log-vraisemblances présentées ci-dessus ne sont pas indépendantes les unes des autres et $l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}) \neq \sum_{z=1}^{K} l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}_z)$. Par conséquent, il est préférable de calculer la contribution à la log-vraisemblance de l'échantillon de validation comme la différence $l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}) - l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}_{-k})$ plutôt que de directement calculer $l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}_k)$.

A la fin des *K* itérations, la log-vraisemblance par validation croisée (*cross-validated log-likelihood* ou *cvl*) pour un certain paramètre λ correspond à la somme des contributions des *K* échantillons de validation :

$$cvl(\lambda) = \sum_{k=1}^{K} \left(l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}) - l_p(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}_{-k}) \right). \quad (3.1)$$

Le paramètre de pénalisation estimé $\hat{\lambda}_{cvl}$ est celui qui maximise la fonction de log-vraisemblance par validation croisée $\hat{\lambda}_{cvl} = \text{argmax}_{\lambda}\{cvl\}$. Il s'agit donc d'une mesure essentiellement basée sur un critère de prédiction. Par la suite, on parlera de lasso-*cvl* pour désigner l'approche lasso avec l'estimation du paramètre λ maximisant (3.1).

Une dernière interrogation reste le choix du nombre de sous-échantillons *K*. En théorie, ce nombre peut varier de 2 (échantillon de validation correspondant à environ 50% des données) à *n* (échantillon de validation correspondant à une seule observation). Pour ce dernier cas, on parle de *leave-one out cross-validation* (*LOOCV*). Plusieurs auteurs ont montré que le choix du nombre de sous-échantillons *K* est essentiellement un compromis entre le biais et la variabilité de l'estimation des coefficients de régression (Breiman et Spector, 1992 ; Efron, 1983 ; Weiss, 1991). En effet, il est montré que plus l'échantillon de validation est grand (i.e. faible valeur de *K*) plus ce biais sera important. A l'inverse, plus l'échantillon de validation est petit (i.e. grande valeur de *K*) plus cette variabilité sera importante. Breiman et Spector (1992) ont montré ce postulat de façon empirique et ont également montré qu'en matière de sélection de modèle, le choix *K* = 5 affiche de bons résultats. Cependant, il est à noter que ce dernier résultat a été montré dans le contexte d'un modèle linéaire et non pénalisé.

Ainsi, dans ce travail de thèse, le nombre de sous-échantillons *K* a été fixé à 5 et une analyse de sensibilité a été réalisée pour voir l'impact du choix de *K* sur la performance de sélection du lasso-*cvl* (*K* = 3 et 10, Annexe A1). Cette étude préliminaire montre qu'avec *K* = 10, le modèle est plus restrictif qu'avec *K* = 5. Cependant, les différences entre *K* = 5 et 10 sont très faibles et d'un point de vue du temps de calcul, notre choix s'est porté sur *K* = 5. Enfin, il est à noter que le classement des méthodes n'est pas impacté par le choix de *K* = 5 ou 10.

3.2 L'excès de faux positifs

Bien que d'un point de vue théorique le lasso ait une bonne capacité de sélection de variables sous certaines conditions dans le cas $p \gg n$, Meinshausen et Bühlmann (2006) et Zhao et Yu (2006) ont montré que cela était moins vraisemblable d'un point de vue pratique. En effet, le paramètre de pénalisation λ est souvent choisi tel que l'erreur de prédiction soit minimale (*cvl*). Or, les auteurs montrent que cela conduit communément à la sélection d'un modèle

contenant à la fois les vrais prédicteurs mais aussi des faux positifs notamment dans le cas de données de grande dimension avec peu de vrais prédicteurs. Ce résultat a également été observé d'un point de vue empirique par d'autres auteurs. Par exemple, à travers une étude de simulation basée sur n = 200 patients et p = 1000 biomarqueurs, Benner et al. (2010) ont montré que le lasso-*cvl* identifie bien les q véritables biomarqueurs actifs (5 sur q = 5 et 23–25 sur q = 30, avec des grands effets $\boldsymbol{\beta}$ égaux chacun à 1,5) mais aussi un grand nombre de faux positifs (57–68) dans le cadre d'un modèle de Cox. Plus récemment, des résultats similaires ont été montrés par Roberts et Nowak (2014) dans le cadre d'un modèle linéaire. En effet, même dans des situations très favorables telles que n = 3000, p = 100, q = 6 avec des effets $\boldsymbol{\beta}$ variant de 0,2 à 7, le lasso-*cvl* identifie les biomarqueurs actifs (6/6 en moyenne) mais sélectionne aussi de nombreux faux positifs (15,3/94 en moyenne). Nous avons souhaité reproduire ce résultat au travers d'une étude de simulation préliminaire. Les résultats de celle-ci sont présentés en Figure 3.1. Cette figure illustre la qualité de sélection de variables de la pénalisation lasso, du point de vue du taux de faux négatifs (*false negative rate* ou *FNR* = FN/q (Tableau 3.1), i.e. taux de biomarqueurs écartés par le modèle parmi les actifs) et du taux de fausses découvertes (*false discovery rate* ou *FDR* = FP/Q (Tableau 3.1), i.e. taux de biomarqueurs inactifs parmi ceux sélectionnés) pour différentes valeurs du paramètre de pénalisation λ.

Tableau 3.1 : Classification des biomarqueurs

	Actifs	**Inactifs**	Total
Sélectionnés	Vrai Positifs (VP)	Faux Positifs (FP)	Q
Non sélectionnés	Faux Négatifs (FN)	Vrai Négatifs (VN)	$p - Q$
Total	q	$p - q$	p

Pour un exemple, avec p = 100 biomarqueurs, q = 10 actifs et plusieurs tailles d'échantillon n, nous avons représenté la combinaison (*FDR* ; *FNR*) : du modèle complet avec $\lambda = 0$ (croix noire), du modèle avec $\hat{\lambda}_{cvl}$ (point blanc), du modèle nul avec $\lambda_0 = \min\{\lambda | p_\lambda = 0\}$ (point noir), ainsi que pour 100 valeurs de λ comprises entre $\hat{\lambda}_{cvl}$ et λ_0 (trait plein) pour 100 réplications. Le premier constat issu de la Figure 3.1 est que, lorsque n est suffisamment élevé (ici, $n \geq 250$), la pénalisation lasso (point blanc) permet d'identifier correctement les biomarqueurs actifs (*FNR* faible voire nul lorsque n est grand) en complément d'un grand nombre de faux positifs (*FDR* très élevé, généralement supérieur à 0,5).

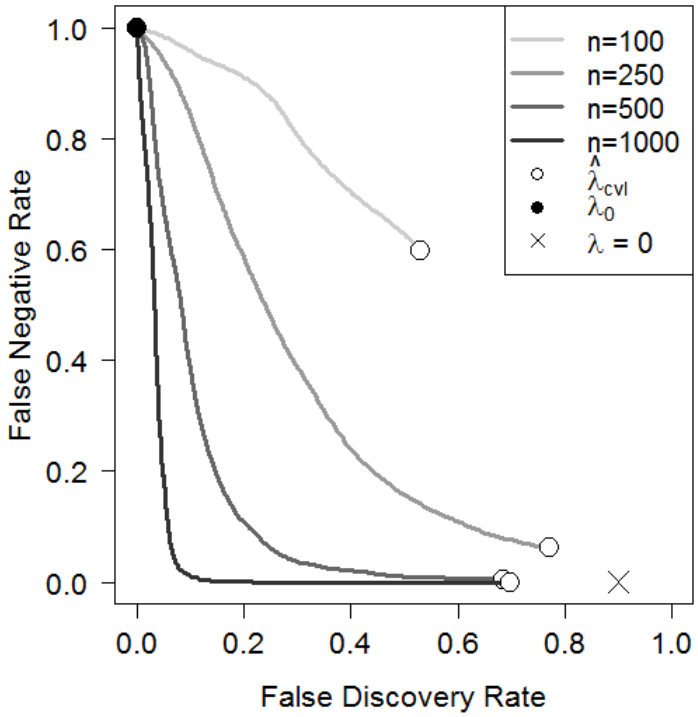


Figure 3.1 : Taux de faux négatifs (*FNR*) en fonction du taux de fausses découvertes (*FDR*) pour des λ variant du modèle ($\hat{\lambda}_{cvl}$) au modèle nul (λ_0) avec p = 100 et q = 10

Ce résultat est en accord avec la littérature. Le second constat concerne la relation entre le *FNR* et le *FDR* pour différentes valeurs de λ comprises entre $\hat{\lambda}_{cvl}$ et λ_0. Comme on peut le voir graphiquement, cette relation semble convexe et de plus en plus prononcée au fur et à mesure que le nombre d'observations n augmente. Cela signifie qu'une augmentation modérée de $\hat{\lambda}_{cvl}$ peut permettre de réduire considérablement le *FDR* sans augmentation forte du *FNR*. En d'autres termes, cela signifie que pour une telle augmentation, les biomarqueurs écartés sont principalement des faux positifs et non des vrais positifs. En revanche, une augmentation excessive du paramètre λ, dépassant le « point de rupture », conduira à une forte augmentation du *FNR* avec une réduction négligeable du *FDR*.

3.3 Améliorations possibles

Pour pallier cet excès de faux positifs sélectionnés, plusieurs méthodes peuvent être proposées. Au vu de la Figure 3.1, des méthodes basées sur l'ajout d'une pénalisation supplémentaire semblent être adaptées pour augmenter la valeur du paramètre de pénalisation λ afin d'être plus restrictif sur la liste des biomarqueurs sélectionnés. Ces pénalisations sont résumées dans le Tableau 3.2. Des variantes du lasso ont également été proposées dans la

littérature pour améliorer les capacités du lasso standard et nous en présentons deux : le lasso adaptatif et le *stability selection*.

3.3.1 Ajout d'une pénalisation pour être plus restrictif sur la sélection

Dans l'optique de choisir un paramètre de pénalisation λ permettant une sélection plus conservatrice des biomarqueurs, plusieurs pénalisations $pen(\lambda)$ ont été rappelées par Müller et Welsh (2010). Dans cet article, les auteurs s'intéressent à pénaliser directement la fonction de perte (ici, la log-vraisemblance). Dans notre cas, la fonction à pénaliser est la log-vraisemblance par validation croisée (*cvl*). Ainsi, la nouvelle fonction à maximiser pour estimer λ est : $cvl(\lambda) - pen(\lambda)$. Selon les auteurs, les pénalisations les plus simples sont de la forme : $pen(\lambda) = \theta_\lambda f_\lambda(p_\lambda)$ où θ_λ correspond à un multiplicateur et $f_\lambda(p_\lambda)$ correspond à une fonction du nombre p_λ de paramètres de régression non nuls restant dans le modèle pour un certain λ. L'idée étant que $pen(\lambda)$ diminue au fur et à mesure que p_λ diminue, et donc que λ augmente.

Tableau 3.2 : Extensions du lasso

Méthode	**Estimation de $\boldsymbol{\lambda}$**	$\boldsymbol{pen(\lambda)}$	
		θ_λ	$f_\lambda(p_\lambda)$
lasso-*cvl*	$\text{argmax}_\lambda\{cvl\}$	–	–
lasso-*AIC*	$\text{argmax}_\lambda\{cvl - pen\}$	2	p_λ
lasso-*RIC*	$\text{argmax}_\lambda\{cvl - pen\}$	2	$\log(p_\lambda)p_\lambda$
lasso-*BIC*	$\text{argmax}_\lambda\{cvl - pen\}$	$\log(n)$	p_λ
lasso-*HQIC*	$\text{argmax}_\lambda\{cvl - pen\}$	$2\log\log(n)$	p_λ
lasso-*eBIC*	$\text{argmax}_\lambda\{cvl - pen\}$	$\log(n) + 2\log(p)$	p_λ
lasso-*AICC*	$\text{argmax}_\lambda\{cvl - pen\}$	2	$\frac{p_\lambda + 1}{n - p_\lambda - 2}$
lasso-*pcvl*	$\text{argmax}_\lambda\{cvl - pen\}$	$\frac{cvl_{\hat{\lambda}_{cvl}} - cvl_{\lambda_0}}{p_{\hat{\lambda}_{cvl}}}$	p_λ
lasso-*1se*	$\text{argmax}_\lambda\{cvl \geq cvl(\hat{\lambda}_{cvl}) - 1\text{SE}\}$	–	–
pct. lasso	$q_{0,95}(\hat{\lambda}_{cvl,1}, \ldots, \hat{\lambda}_{cvl,r})$	–	–

Légende. pct. : *percentile*, *n* : taille de l'échantillon, *p* : nombre de biomarqueurs, $q_{0,95}$: 95ème percentile de la distribution.

Un premier choix pour cette fonction de p_λ est de choisir simplement $f_\lambda(p_\lambda) = p_\lambda$. Dans ce cas, on retrouve le critère d'information d'Akaike (*Akaike information criterion* ou *AIC*, Akaike, 1974) lorsque θ_λ est fixé à 2. Le multiplicateur peut également dépendre du nombre de patients *n*. Les principaux exemples sont les critères d'information bayésiens (*Bayesian*

information criterion ou *BIC*, Schwartz, 1978) ou de Hannan et Quinn (*Hannan and Quinn information criterion* ou *HQIC*, Hannan et Quinn, 1979) où θ_λ est respectivement égal à $\log(n)$ et $2\log\log(n)$. Plus récemment, une extension du *BIC* (*extended BIC* ou *eBIC*, Chen et Chen, 2008) a été proposée pour améliorer ses performances, notamment dans le cadre des études d'association pangénomiques (*GWAS*) et correspond à $\theta_\lambda = \log(n) + 2\log(p)$. Enfin, la fonction $f_\lambda(p_\lambda)$ peut ne pas être simplement égale à p_λ. C'est le cas, par exemple, pour le critère d'information de Akaike corrigé (*corrected AIC* ou *AICC*, Sugiura, 1978) où $f_\lambda(p_\lambda) = (p_\lambda + 1)/(n - p_\lambda - 2)$ ou pour le *risk inflation criterion* (ou *RIC*, Foster et George, 1994) avec lequel $f_\lambda(p_\lambda) = \log(p_\lambda)p_\lambda$.

Bien que ces pénalisations supplémentaires permettent d'être plus restrictif sur la sélection des biomarqueurs en choisissant un paramètre de pénalisation λ plus grand, elles sont pour la plupart fixes par rapport à l'échelle de la log-vraisemblance et ne permettent pas de réellement avoir un compromis entre l'exhaustivité du modèle (pour $\hat{\lambda}_{cvl}$) et sa parcimonie (pour λ_0). C'est pourquoi, dans ce premier axe du travail de thèse, nous avons proposé une nouvelle extension empirique appelée *pcvl* pour *penalized cross-validated log-likelihood* et illustrée en Figure 3.2. Comme la plupart des pénalisations énoncées précédemment, nous avons choisi $f_\lambda(p_\lambda) = p_\lambda$ comme fonction du nombre de paramètres de régression non nuls restant dans le modèle pour un certain λ. Concernant le multiplicateur θ_λ, nous avons choisi une quantité permettant de résumer à quel point $\hat{\lambda}_{cvl}$ améliore l'exhaustivité du modèle en matière de vraisemblance et détériore sa parcimonie en comparaison à λ_0 en matière de nombre de paramètres non nuls. Ainsi, θ_λ peut être vu comme le ratio entre $cvl_{\hat{\lambda}_{cvl}} - cvl_{\lambda_0}$ (gain en exhaustivité du modèle) et $p_{\hat{\lambda}_{cvl}} - p_{\lambda_0}$(perte en parcimonie) avec p_{λ_0} = 0 par définition. La pénalisation peut alors s'écrire :

$$pen(\lambda) = \frac{cvl_{\hat{\lambda}_{cvl}} - cvl_{\lambda_0}}{p_{\hat{\lambda}_{cvl}}} \times p_\lambda, \forall\, \lambda \in [\hat{\lambda}_{cvl};\, \lambda_0].$$

Nous avons donc proposé de choisir le λ qui maximise

$$pcvl(\lambda) = cvl(\lambda) - \left(\frac{cvl_{\hat{\lambda}_{cvl}} - cvl_{\lambda_0}}{p_{\hat{\lambda}_{cvl}}} \times p_\lambda\right). \qquad (3.2)$$

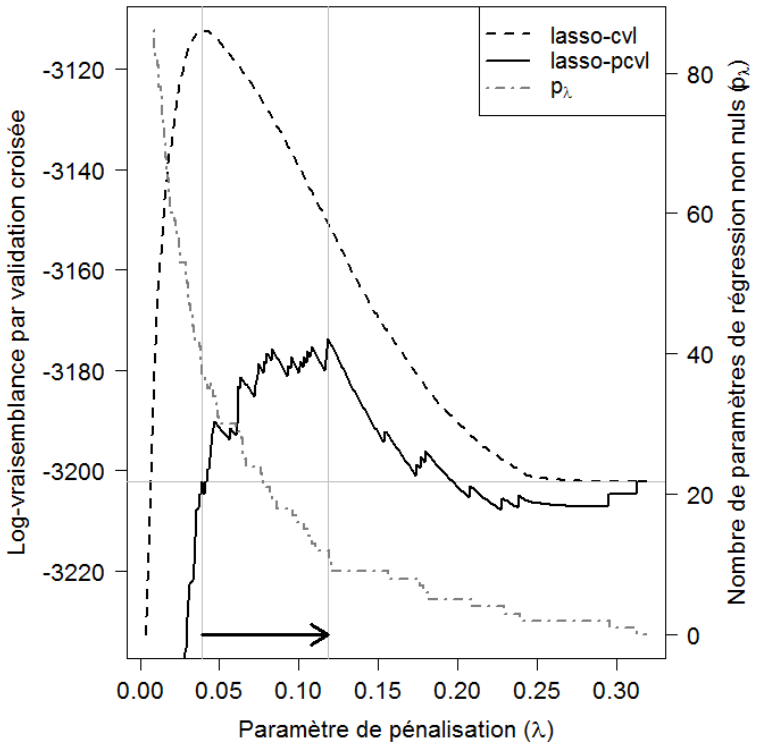


Figure 3.2 : Illustration de la log-vraisemblance par validation croisée classique (*cvl)* ou pénalisée *(pcvl)* et du nombre de paramètres de régression non nuls (p_λ) en fonction de λ à partir d'un jeu de données simulé (n = 500, p = 100, q = 10)

Contrairement aux autres pénalisations présentées jusqu'à présent, notre définition de θ_λ permet de donner un poids à la variation de parcimonie, $p_\lambda / p_{\hat{\lambda}_{cvl}} \in [0, 1]$, qui est à la même échelle que l'amplitude de la *cvl* sur l'étendue $\lambda \in [\hat{\lambda}_{cvl}\,;\,\lambda_0]$. En effet, ce choix de pénalisation implique que $pcvl(\hat{\lambda}_{cvl}) = pcvl(\lambda_0)$ ce qui force le compromis entre la qualité du modèle et sa parcimonie en mettant ces deux extrêmes au même niveau de log-vraisemblance par validation croisée. Enfin, pour le cas particulier où $\hat{\lambda}_{pcvl} = \text{argmax}_\lambda\{pcvl\} = \{\hat{\lambda}_{cvl}\,;\,\lambda_0\}$, nous avons choisi $\hat{\lambda}_{pcvl} = \hat{\lambda}_{cvl}$ afin d'éviter d'être trop conservateur.

Une autre approche empirique (Friedman, Hastie et Tibshirani, 2010 ; Hastie, Tibshirani et Friedman, 2009) dépend, tout comme le lasso-*pcvl*, de l'allure de la fonction *cvl* mais également de sa variabilité. En effet, celle-ci consiste à choisir le paramètre λ en fonction de l'incertitude de la déviance (mesure proportionnelle à la log-vraisemblance par validation croisée). Cette approche, nommée *one-standard error rule*, consiste à choisir le plus grand paramètre λ tel que la déviance soit au plus à un écart type de la déviance minimale. Cette approche vise donc à être, elle aussi, plus contraignante sur le choix du paramètre. Ainsi, plus la variabilité de la vraisemblance est importante, plus la valeur d'un écart type sera élevée et

plus le paramètre λ choisi sera élevé par rapport à celui estimé initialement par le lasso standard $\hat{\lambda}_{cvl}$, et inversement.

Enfin, Roberts et Nowak (2014) ont également montré que le résultat de la sélection du lasso standard peut dépendre (parfois fortement) de l'affectation aléatoire des observations dans les différents sous-groupes lors de la validation croisée. En effet, dans certains cas et notamment lorsque $n \ll p$, l'estimation du paramètre λ peut être très instable et ainsi la liste des biomarqueurs sélectionnés peut grandement varier. Ces auteurs proposent alors d'utiliser ce constat pour tenter d'être plus restrictif sur le nombre de biomarqueurs sélectionnés. La méthode proposée, dite *percentile lasso*, consiste à estimer r fois le paramètre λ pour différentes listes d'affectation des observations aux sous-groupes. On obtient ainsi une liste $\Omega = \{\hat{\lambda}_{cvl,1}, \dots, \hat{\lambda}_{cvl,r}\}$ de paramètres de pénalisation et celui retenu, *in fine*, correspond à un haut percentile (e.g. 0,95) de cette liste : $\hat{\lambda} = \mathrm{q}_{0,95}(\Omega)$. Dans leur article, les auteurs ont choisi r = 100 affectations, cependant ils affirment que pour tout $r \geq 10$ le *percentile lasso* a des performances comparables. Ainsi, dans notre travail, r = 20 a été choisi afin d'éviter des temps de calcul trop longs.

3.3.2 Variantes du lasso

Plusieurs variantes de la pénalisation lasso ont été proposées depuis sa publication originale (Tibshirani, 2011). Dans ce travail de thèse, je me suis intéressé à des variantes du lasso pouvant aider à la réduction du nombre de faux positifs telles que le lasso adaptatif (*adaptive lasso*) ou encore le *stability selection*.

Le principe du lasso adaptatif, proposé par Zou (2006) puis adapté par Zhang et Lu (2007) pour la survie, est identique à celui du lasso standard en introduisant un terme de pondération w_j. Cette pondération a pour but d'attribuer à chaque biomarqueur j une pénalisation qui est spécifique ($\lambda_j = \lambda \times w_j$), contrairement au lasso standard pour lequel la pénalisation est la même pour l'ensemble des biomarqueurs (λ, avec w_j = 1). Le lasso adaptatif vise à discriminer les biomarqueurs selon leur force d'association $\tilde{\beta}_j$ avec le critère de jugement qui est estimée dans une étape préliminaire. On note $w_j = \left|\tilde{\beta}_j\right|^{-1}$. Ainsi plus le biomarqueur a une force d'association $\tilde{\beta}_j$ importante avec le critère de jugement, plus sa pondération w_j sera faible et, par conséquent, plus sa probabilité d'être retenu dans le modèle final sera élevée car λ_j sera faible. On note alors la log-vraisemblance pénalisée

$$l_p(\boldsymbol{\beta}, \boldsymbol{X}) = l(\boldsymbol{\beta}, \boldsymbol{X}) - \lambda \sum_{j=1}^{p} \widetilde{w}_j |\beta_j|.$$

Jusqu'alors, différentes procédures d'estimation des coefficients de régression préliminaires $\widetilde{\boldsymbol{\beta}}$, et donc des poids $\widetilde{\boldsymbol{w}}$, ont été proposées dans la littérature telles que : un modèle multivarié pour le cas $n > p$, une régression ridge, un modèle univarié ou encore une régression lasso préliminaire pour le cas $n \leq p$. Cependant, aucun consensus n'a été établi. Dans l'idée de comparer notre extension aux meilleurs comparateurs possibles, une analyse de sensibilité a été mise en place pour évaluer l'impact du choix de la procédure d'estimation sur la sélection finale des biomarqueurs. Les résultats (Annexe A2) montrent que l'estimation des coefficients de régression $\widetilde{\boldsymbol{\beta}}$ à partir d'une régression lasso préliminaire donne les meilleurs résultats dans notre étude. De plus, ce choix semble pertinent dans le sens où le lasso standard tend à identifier correctement les biomarqueurs actifs avec une grande probabilité malgré des biomarqueurs inactifs supplémentaires. Cela peut être donc vu comme un premier filtrage de variables. Cette procédure a également été adoptée par d'autres auteurs (Benner et al., 2010 ; van de Geer, Bühlmann et Zhou, 2011) et est identique à l'approche *multi-step* lasso à 2 étapes (Bühlmann et van de Geer, 2011).

La seconde variante évaluée est basée sur l'application d'un algorithme de sélection sur des sous-échantillons de données. On nomme cette approche le *stability selection* (Meinshausen et Bühlmann, 2010). Ainsi, lorsque les sous-échantillons sont générés sans remise (jackknife ou *repeated random sampling*), la probabilité de sélection d'un biomarqueur peut être estimée comme étant la proportion de sous-échantillons dans lesquels le biomarqueur a été sélectionné dans le modèle. Comme dit précédemment, un exemple d'utilisation récente de cette approche est le *causal stability ranking* (*CStaR*, Stekhoven et al., 2012) dont l'esprit est de combiner le *stability selection* avec une approche causale (*IDA* : *intervention calculus when the DAG is absent*, Maathuis et al., 2009) connue comme relativement instable (section 2.2). Dans ce projet, le *stability selection* est combiné avec la pénalisation lasso comme algorithme de sélection. Cette approche est basée sur trois paramètres arbitraires : un seuil $\pi_{thr} \in [0\,;1]$ utilisé pour filtrer les variables ayant une basse probabilité de sélection ; un paramètre de fragilité $\zeta \in (0\,;1]$ utilisé pour obtenir un paramètre de pénalisation compris entre λ/ζ et λ pour chaque biomarqueur (i.e. même principe que le lasso adaptatif en dehors du fait que la pénalisation est obtenue aléatoirement et non basée sur des estimations préliminaires) afin

d'introduire une incertitude supplémentaire ; et enfin le nombre attendu d'erreurs de type-I (i.e. faux positifs) noté α_{FWER}. Les paramètres retenus pour ce travail sont : π_{thr} = 0,6, ζ = 1 (i.e. absence de pondération) et $\alpha_{FWER} = 0{,}05 \times p$ correspondant à un nombre attendu de 5% du nombre total de biomarqueurs p comme étant sélectionnés à tort. Pour une compréhension de l'impact de ces différents choix, une analyse de sensibilité a été réalisée (Annexe A3) et montre bien entendu que le critère qui impacte le plus les résultats est le α_{FWER}. Plus celui-ci est faible et plus le modèle final sera restrictif.

3.4 Étude de simulation

Pour évaluer notre extension (lasso-*pcvl*) et la comparer au lasso standard (lasso-*cvl*) ainsi qu'aux dix autres compétiteurs, une étude de simulation a été mise en place. Dans cette étude, nous avons étudié la capacité des méthodes à détecter les vrais biomarqueurs pronostiques (i.e. prédictifs de la survie) dans un modèle de Cox à grande dimension.

3.4.1 Génération des données

Les biomarqueurs $\boldsymbol{X}$ ont été générés à partir d'une loi gaussienne multivariée

$$\boldsymbol{X} \sim N\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \cdots & \sigma_{1p}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{p1}^2 & \cdots & 1 \end{pmatrix}\right),$$

pour permettre une structure de corrélation entre les biomarqueurs. En effet, la covariance entre deux biomarqueurs j et j', notée $\sigma_{jj'}^2$, est égale au produit de leurs écarts types (σ_j et $\sigma_{j'}$) et de leur corrélation ($\rho_{jj'}$) : $\sigma_{jj'}^2 = \rho_{jj'}\sigma_j\sigma_{j'}$. Afin d'être le plus réaliste possible, une structure de corrélation permettant à certains biomarqueurs d'être fortement corrélés et à d'autres d'être totalement indépendants a été choisie. Cette structure de corrélation, que l'on nomme autorégressive par bloc (Pang, Tong et Zhao, 2009), définit m blocs de p_m biomarqueurs pour lesquels la corrélation est égale à $\rho_{jj'} = \rho^{|j-j'|}\ \forall j, j' \in m$. Ainsi, plus un couple de biomarqueurs s'éloigne de la diagonale de la sous-matrice $p_m \times p_m$, plus leur corrélation sera faible. En revanche, pour deux biomarqueurs issus de deux blocs différents leur corrélation sera nulle et ils seront donc indépendants. Enfin, nous avons choisi de générer des biomarqueurs centrés ($\mu_1 = \cdots = \mu_p = 0$) et réduits ($\sigma_1^2 = \cdots = \sigma_p^2 = 1$).

Comme nous nous intéressons à un critère de jugement censuré, nous avons généré des temps de survie *t*. Ces temps ont été générés à partir de valeurs aléatoires issues d'une distribution de Weibull et correspondant à un risque instantané

$$h(t, \boldsymbol{X}) = b^{-a} a t^{a-1} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}),$$

avec a le paramètre de forme et b le paramètre d'échelle. Ainsi, on peut en déduire la fonction de survie

$$S(t, \boldsymbol{X}) = \exp\left(\int_0^t h(u, \boldsymbol{X}) du\right) = \exp(-b^{-a} t^a \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X})),$$

et le temps de survie médian $m(\boldsymbol{X})$ pour un vecteur de covariables $\boldsymbol{X}$ donné

$$m(\boldsymbol{X}) = b(\log 2 \exp(-\boldsymbol{\beta}^T \boldsymbol{X}))^{1/a}.$$

Par conséquent, le paramètre d'échelle de base b_0 pour un profil $\boldsymbol{X_0} = (0, \ldots, 0)^{\mathrm{T}}$ est égal à

$$b_0 = m((0, \ldots, 0)^{\mathrm{T}})(\log 2)^{1/a} = m_0(\log 2)^{-1/a}$$

et pour un patient spécifique $\boldsymbol{X}$, on notera son paramètre d'échelle

$$b(\boldsymbol{X}) = b_0 \times \exp(-\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}/a) = b_0 \mathrm{HR}(\boldsymbol{X})^{-1/a}.$$

Ainsi, il est possible de générer un temps de survie spécifique pour chaque patient à partir d'une distribution de Weibull de paramètre de forme a et de paramètre d'échelle $b(\boldsymbol{X})$:

$$h(t, \boldsymbol{X}) = b(\boldsymbol{X})^{-a} a t^{a-1}.$$

Dans cette étude de simulation, les temps de survie t_e ont été générés avec un risque constant au cours du temps (distribution exponentielle, paramètre de forme a = 1) et une médiane de survie de base $m(\boldsymbol{X_0})$ = 1 an.

3.4.2 Choix des scénarios

Dans cette étude de simulation, 3^3 = 27 scénarios ont été proposés, faisant varier le nombre de patients *n* (100, 500 et 1000), le nombre de biomarqueurs candidats *p* (10, 100 et 1000) et le nombre de biomarqueurs réellement pronostiques *q*. Pour ce dernier, trois valeurs ont été choisies : un scénario, dit nul, considérant aucun biomarqueur pronostique (q = 0) et deux

scénarios, dits alternatifs, considérant soit un (q = 1) soit plusieurs (q > 1) biomarqueurs pronostiques. Dans cette dernière configuration, le nombre de biomarqueurs réellement pronostiques q, augmentait au fur et à mesure que le nombre total de biomarqueurs candidats p augmentait. Ainsi, pour p = 10, 100 et 1000 biomarqueurs candidats, q = 2 (20%), 10 (10%) et 20 (5%) biomarqueurs pronostiques ont été choisis. Dans tous les cas, la réduction du risque relatif pour l'augmentation d'une unité d'un biomarqueur pronostique q a été fixée à 20%, correspondant à un *HR* égal à 0,8 (i.e. β_j = log(0,8) ≈ -0,22). Pour chaque scénario, 250 réplications ont été effectuées.

Ces scénarios ont été implémentés en considérant une structure de corrélation entre les biomarqueurs et des temps censurés. Pour cela, les données ont été générées avec une structure de corrélation autorégressive par bloc (voir section 3.4.1) avec ρ = 0,6 et des blocs aléatoires de taille variable dépendant du nombre total de biomarqueurs p (i.e. 5 blocs de taille 2 pour p = 10 biomarqueurs, 10 blocs de taille 10 pour p = 100 biomarqueurs et enfin 20 blocs de taille 50 pour p = 1000 biomarqueurs). Les temps de censure t_c ont été générés à partir d'une distribution uniforme (i.e. censure indépendante et non informative) comprise entre 2 et 5 ans correspondant à une étude incluant les patients uniformément sur une période de 3 ans et continuant à les suivre pendant 2 ans avec arrêt de l'étude pour analyse. Ainsi, un patient était considéré comme ayant un évènement à t_e si $t_e \leq t_c$ et à l'inverse, un patient était considéré comme censuré à t_c si $t_e > t_c$. Dans cette étude de simulation, cela s'est traduit par un taux de censure empirique compris entre 11% et 20% selon les scénarios.

Pour finir, différentes analyses de sensibilité ont été réalisées pour étudier l'impact de la corrélation, de la censure ou encore de coefficients de régression $\boldsymbol{\beta}$ variables et basés sur des estimations provenant d'une réelle application. Pour chacune de ces analyses de sensibilité, les 27 scénarios ont également été générés.

3.4.3 Critères d'évaluation

L'objectif principal de cette étude a été d'évaluer la capacité des méthodes à bien identifier les véritables biomarqueurs pronostiques. Par conséquent, des indicateurs basés sur le tableau de contingence (Tableau 3.1) sont proposés.

En effet, notre premier critère d'évaluation est le taux de fausses découvertes (*False Discovery Rate* (*FDR*)) correspondant au taux de biomarqueurs inactifs parmi les Q biomarqueurs sélectionnés par la méthode (Genovese et Wasserman, 2002). On a donc :

$$FDR = \frac{FP}{FP + VP} = \frac{FP}{Q}.$$

Cette mesure, comprise entre 0 et 1, permet de juger de la fiabilité des biomarqueurs identifiés. Il est à noter que dans les scénarios nuls, le nombre de véritables biomarqueurs q est égal à zéro et *a fortiori* VP = 0, par conséquent, le *FDR* est forcément égal à 0 (dans le cas où aucun biomarqueur n'est sélectionné, Q = FP = 0) ou 1 (dans le cas où au moins un biomarqueur est sélectionné, Q = FP ≥ 1). Ainsi, dans ces scénarios, le *FDR* correspond strictement à la définition de l'erreur de type-I (Farcomeni et La, 2007). Le second critère évalué est le taux de faux négatifs (*False Negative Rate* (*FNR*)) et correspond au taux de biomarqueurs réellement pronostiques qui ne sont pas sélectionnés par la méthode (Pawitan et al., 2005). On a donc :

$$FNR = \frac{FN}{FN + VP} = \frac{FN}{q}.$$

Cette mesure, également comprise entre 0 et 1, permet de juger de la puissance de sélection des véritables biomarqueurs pronostiques. En effet, cette mesure peut être interprétée comme l'erreur de type-II et son complémentaire (1 – *FNR* = VP/q) comme la puissance ou encore la sensibilité. Ces deux quantités, le *FDR* et le *FNR*, sont des quantités que l'on cherche à minimiser.

Outre ces quantités, la capacité de sélection des méthodes a également été évaluée d'un point de vue global au travers d'un seul critère mélangeant la capacité de classification des biomarqueurs parmi les actifs (i.e. sensibilité, Se) et les inactifs (i.e. spécificité, Sp). La spécificité correspond au taux de biomarqueurs non sélectionnés parmi les biomarqueurs inactifs (i.e. 1 – taux de faux positifs ou *FPR* = FP/FP+VN = FP/(p – q)). Notre critère, noté G (Kubat et Matwin, 1997), correspond à la moyenne géométrique entre la sensibilité et la spécificité

$$G = \sqrt{\mathrm{Se} \times \mathrm{Sp}} = \sqrt{(1 - FNR) \times (1 - FPR)},$$

et est compris entre 0 et 1 et doit être maximisé.

Enfin, en complément de la capacité de sélection des méthodes, nous avons souhaité évaluer la capacité de prédiction des méthodes. En effet, ces deux critères ne sont pas obligatoirement

liés car une méthode peut sélectionner un biomarqueur faussement positif qui est fortement corrélé à un vrai et ainsi avoir une bonne capacité de prédiction malgré une mauvaise sélection. A l'inverse, un biomarqueur pronostique peut être bien sélectionné, mais son rôle prédictif de la survie peut être négligeable si son effet est petit. Dans cette étude, nous nous sommes intéressés à un critère évaluant la capacité de discrimination des méthodes, tel que la statistique de concordance C. Une comparaison d'estimateurs évaluant la discrimination des modèles de survie a été faite par Schmid et Potapov (2012). La statistique la plus intuitive a été proposée par (Harrell et al., 1982 ; Harrell et al., 1984) et cherche initialement à évaluer la concordance entre un critère de jugement classique et un prédicteur $\boldsymbol{\pi}$ (dans notre cas : $\boldsymbol{\pi} = \boldsymbol{X\beta}$). Une extension a depuis été proposée par Harrell pour le cas des variables de durées présentant à la fois une variable de temps et une indicatrice d'évènement (Harrell, Lee et Mark, 1996 ; Pencina et Agostino, 2004) et s'écrit :

$$C^*_{\text{Har}}(\boldsymbol{\pi}) = \frac{\sum_{i<i'}(\mathbf{I}(t_i < t_{i'})\,\mathbf{I}(\hat{\pi}_i > \hat{\pi}_{i'})\delta_i + \mathbf{I}(t_{i'} < t_i)\,\mathbf{I}(\hat{\pi}_{i'} > \hat{\pi}_i)\delta_{i'})}{\sum_{i<i'}(\mathbf{I}(t_i < t_{i'})\delta_i + \mathbf{I}(t_{i'} < t_i)\delta_{i'})}$$

avec i et $i' \in \{1, \ldots, n\}$. Cet estimateur converge vers la probabilité de concordance dans une situation sans censure. En revanche, en présence de censure, C^*_{Har} est biaisé en raison du fait que les paires d'observations dans lesquelles le plus petit temps de survie observé est censuré, sont ignorées. Pour pallier cela, Uno et al. (2011) ont proposé une modification de C^*_{Har} en introduisant une pondération basée sur la probabilité qu'un individu i soit censuré $\hat{S}_i^C(t)$ calculée à partir de l'estimateur de Kaplan Meier. On note cet estimateur, pour un horizon τ fixé :

$$C_{\text{Uno}}(\tau, \boldsymbol{\pi}) = \frac{\sum_{i,i'}\left(\hat{S}_i^C(t)\right)^{-2}\mathbf{I}(t_i < t_{i'}, t_i < \tau)\,\mathbf{I}(\hat{\pi}_i > \hat{\pi}_{i'})\delta_i}{\sum_{i,i'}\left(\hat{S}_i^C(t)\right)^{-2}\mathbf{I}(t_i < t_{i'}, t_i < \tau)\delta_i} \qquad (3.3)$$

dans lequel la technique de pondération par l'inverse de probabilité de censure permet de converger vers la vraie valeur en présence de censure. Dans notre cas, nous avons choisi $\tau = 5$ ans.

Pour tenir compte d'un éventuel suroptimisme dans l'évaluation de la prédiction lié à un surapprentissage du modèle, un échantillon test ou de validation $\boldsymbol{X}^{\text{V}}$ ayant les mêmes caractéristiques que l'échantillon d'apprentissage $\boldsymbol{X}$ a été généré, pour lequel la statistique de concordance a été calculée (i.e. même principe qu'une validation externe). Ainsi, nous avons

calculé le prédicteur $\boldsymbol{\pi}$ comme le produit matriciel entre les données de l'échantillon test $\boldsymbol{X}^{V}$ et les coefficients de régression estimés à partir de l'échantillon d'apprentissage et non de l'échantillon test $\widehat{\boldsymbol{\beta}}$. A noter que les coefficients $\widehat{\boldsymbol{\beta}}$ sont non pénalisés, i.e. estimés dans un nouveau modèle de Cox non pénalisé ne contenant que les variables sélectionnées par le modèle pénalisé. Nous avons choisi de réestimer les coefficients afin d'obtenir une comparabilité entre toutes les approches discutées car une méthode telle que le *stability selection* ne permet pas d'obtenir d'estimation des coefficients des biomarqueurs mais uniquement de savoir s'ils sont sélectionnés ou non. Cependant, nous verrons au cours du Chapitre 5 qu'en matière de prédiction il semble préférable de garder les coefficients pénalisés plutôt que de les réestimer *a posteriori*. Dans cette étude, une analyse de sensibilité a été réalisée en gardant les coefficients pénalisés (Annexe A4) pour évaluer l'impact de ce choix d'estimation des coefficients. Les résultats semblent effectivement indiquer une légère amélioration de la performance de prédiction lorsque l'on conserve les coefficients pénalisés, cependant, le classement des méthodes reste globalement inchangé.

3.4.4 Implémentation

L'implémentation de l'ensemble des méthodes discutées a été réalisée à l'aide du logiciel `R`. Plusieurs packages ont été testés pour implémenter la pénalisation lasso appliquée à des données de survie (Annexe B19) et le package `glmnet` a finalement été retenu (Friedman et al., 2010). A noter que nous avons découvert qu'une calibration du paramètre de pénalisation estimé par la *cvl* est nécessaire pour utiliser correctement ce package (Annexe B20). Pour implémenter le *stability selection*, nous utilisons le package `lol` (*lots of lasso*, Yuan, 2016). Le package `lol` dépend du package `penalized`, il n'est donc pas nécessaire de faire de calibration supplémentaire comme pour le package `glmnet` (Annexe B20). Enfin, le package `survAUC` (Schmid et Potapov, 2012) a été utilisé pour le calcul de la statistique de concordance de Uno.

3.4.5 Résultats

Pour une présentation plus claire des résultats de l'étude de simulation, nous avons choisi de se focaliser sur le lasso standard (lasso-*cvl*), notre extension (lasso-*pcvl*) et les cinq méthodes montrant les meilleurs résultats : le lasso-*1se*, le lasso-*AIC*, le lasso-*RIC*, le lasso adaptatif et le *stability selection*. Concernant les autres méthodes : le lasso-*HQIC* tend à être trop conservateur pour certaines combinaisons (n, p) notamment lorsque le nombre de patients

n'est pas suffisamment important (n = 100 ou 500). Ce constat est d'autant plus prononcé pour les pénalisations lasso-*BIC* et lasso-*eBIC* qui sont plus conservatrices que le lasso-*HQIC* de par leur nature ($pen_{HQIC}(\lambda) < pen_{BIC}(\lambda) < pen_{eBIC}(\lambda) \Rightarrow \hat{\lambda}_{HQIC} \leq \hat{\lambda}_{BIC} \leq \hat{\lambda}_{eBIC}$, Annexe A5). Le lasso-*AICC* montre des résultats proches du lasso-*cvl* standard et ce d'autant plus lorsque n augmente. Ce constat était également attendu au vu de la nature de la pénalisation du lasso-*AICC* ($n \gg p_\lambda \Rightarrow pen_{AICC}(\lambda) = \frac{p_\lambda+1}{n-p_\lambda-2} \to 0 \Rightarrow cvl(\lambda) \approx AICC(\lambda)$). Enfin, bien que le *percentile lasso* réduise confortablement le *FDR* dans le scénario nul comparé au lasso-*cvl*, son intérêt reste limité dans les scénarios alternatifs avec une faible réduction du *FDR* notamment dans le cas où n est grand. En revanche, lorsque n est petit, les écarts semblent plus importants sans doute parce que le lasso-*cvl* est moins stable et, par conséquent, l'étendue des valeurs de λ pour les différentes réplications du *percentile lasso* est plus importante. Malheureusement, dans le cas où n est petit, le lasso-*cvl* rate déjà beaucoup de biomarqueurs actifs et donc le *percentile lasso* est beaucoup trop conservateur. Les résultats de ces cinq méthodes sont présentés succinctement en Annexe A6.

Scénario nul (q = 0). Pour rappel, dans ce scénario aucun biomarqueur n'est réellement actif. Par conséquent, pour un jeu de données, le *FNR* n'est pas calculable et le *FDR* est binaire (0 si le modèle nul est sélectionné, 1 sinon). Ainsi, en moyennant sur plusieurs répétitions, le *FDR* correspond à la probabilité qu'une méthode sélectionne au moins un biomarqueur qui, de fait, est un faux positif. Les résultats pour ce scénario nul sont présentés en Tableau 3.3. Concernant le lasso-*cvl* standard, le *FDR* augmente en moyenne au fur et à mesure que le nombre de biomarqueurs candidats p augmente : variant de 0,32 à 0,38 pour p = 10 biomarqueurs et de 0,47 à 0,49 pour p = 1000 biomarqueurs. Le lasso-*pcvl* et le lasso-*RIC* donnent les mêmes résultats que le lasso-*cvl* en matière de *FDR* signifiant que le lasso-*pcvl* et le lasso-*RIC* sélectionnent au moins un biomarqueur lorsque le lasso-*cvl* en sélectionne au moins un également. Ce résultat était attendu de par leur pénalisation. En effet, pour le lasso-*pcvl* nous forçons l'égalité $pcvl(\hat{\lambda}_{cvl}) = pcvl(\lambda_0)$ et nous privilégions le choix de $\hat{\lambda}_{cvl}$ par rapport à λ_0, ainsi, si $\hat{\lambda}_{cvl} \neq \lambda_0$ et $\text{argmax}_\lambda\{pcvl\} = \{\hat{\lambda}_{cvl}; \lambda_0\}$ alors nous choisirons toujours $\hat{\lambda}_{pcvl} = \hat{\lambda}_{cvl}$. Il n'est donc pas possible de choisir λ_0, correspondant au modèle nul, lorsque $\hat{\lambda}_{cvl} \neq \lambda_0$. Un raisonnement similaire peut être réalisé pour le lasso-*RIC* dont la pénalisation est $pen_{RIC}(\lambda)$=$2\log(p_\lambda)p_\lambda$. En effet, avec $\lambda_1 = \min\{\lambda | p_\lambda = 1\}$ nous avons $pen_{RIC}(\lambda_1) = pen_{RIC}(\lambda_0) = 0$ car log(1) = 0.

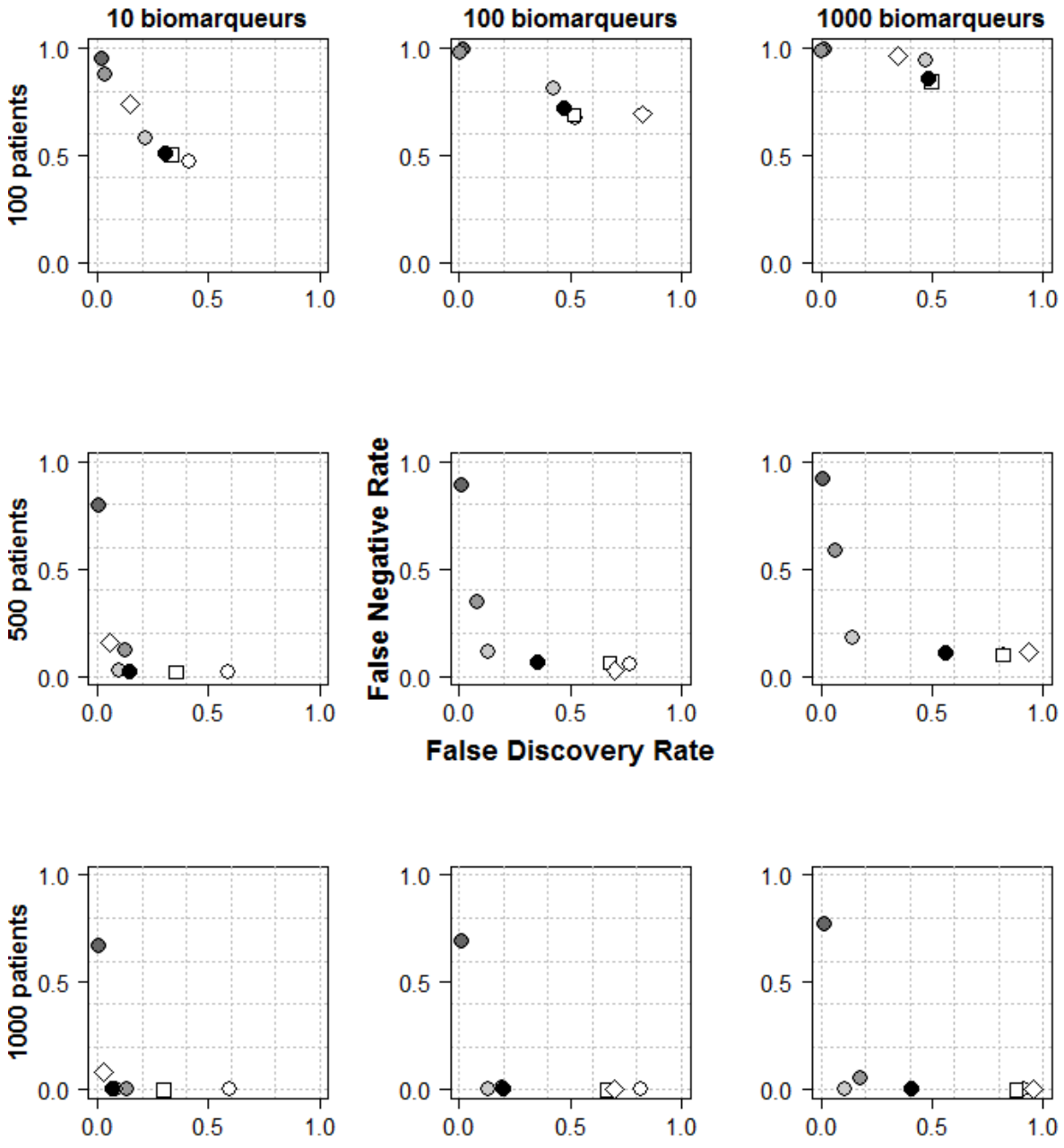
Tableau 3.3 : Taux de fausses découvertes (*FDR*) / nombre de biomarqueurs sélectionnés dans le scénario nul ($q = 0$)

p	n	lasso-*cvl*	lasso-*1se*	lasso-*pcvl*	lasso-*AIC*	lasso-*RIC*	lasso adaptatif	*stability selection*
10	100	0,38 / 1,30	0,00 / 0,00	0,38 / 0,95	0,02 / 0,03	0,38 / 0,40	0,38 / 1,00	0,26 / 0,24
	500	0,32 / 1,11	0,00 / 0,00	0,32 / 0,89	0,02 / 0,02	0,32 / 0,32	0,32 / 0,87	0,29 / 0,28
	1000	0,36 / 1,28	0,00 / 0,00	0,36 / 0,97	0,03 / 0,03	0,36 / 0,37	0,36 / 0,98	0,23 / 0,32
100	100	0,42 / 2,43	0,00 / 0,00	0,42 / 1,69	0,01 / 0,01	0,42 / 0,42	0,42 / 1,96	0,88 / 2,33
	500	0,45 / 2,78	0,00 / 0,00	0,45 / 1,85	0,00 / 0,00	0,45 / 0,45	0,45 / 2,49	0,98 / 3,42
	1000	0,42 / 2,45	0,00 / 0,00	0,42 / 1,65	0,01 / 0,01	0,42 / 0,42	0,42 / 2,22	0,97 / 3,56
1000	100	0,47 / 3,22	0,01 / 0,03	0,47 / 2,11	0,02 / 0,02	0,47 / 0,47	0,47 / 2,69	0,34 / 0,40
	500	0,48 / 4,86	0,00 / 0,00	0,48 / 2,88	0,00 / 0,00	0,48 / 0,48	0,48 / 4,62	1,00 / 13,6
	1000	0,49 / 6,40	0,00 / 0,00	0,49 / 4,10	0,00 / 0,00	0,49 / 0,50	0,49 / 6,15	1,00 / 23,9

Légende. p : nombre de biomarqueurs, n : taille de l'échantillon. Quantités moyennes basées sur 250 réplications.

Ainsi, on a généralement $RIC(\lambda_1) \geq RIC(\lambda_0)$ et nous choisissons donc toujours λ_1 plutôt que le modèle nul avec λ_0. Cela se produit bien car pour la majorité des scénarios nuls, le *FDR* est égal au nombre moyen de biomarqueurs sélectionnés pour le lasso-*RIC*. Cela signifie que lorsque le lasso-*cvl* identifie au moins un biomarqueur, le lasso-*RIC* ne sélectionne qu'un seul biomarqueur en augmentant le paramètre de pénalisation jusqu'à λ_1 sans jamais parvenir à λ_0 pour les raisons présentées ci-dessus. Le lasso-*pcvl* réduit également le nombre de faux positifs sélectionnés bien que dans une moindre mesure. Le lasso adaptatif obtient, lui aussi, les mêmes résultats que le lasso-*cvl* en ce qui concerne les *FDR* et ne réussit pas à réduire considérablement le nombre moyen de biomarqueurs sélectionnés contrairement au lasso-*pcvl* et au lasso-*RIC*. Le lasso-*1se* et le lasso-*AIC* affichent d'excellents résultats pour ces scénarios nuls avec des *FDR* moyens proches de zéro voire nuls pour l'ensemble des combinaisons (n, p). Enfin, à l'inverse, le *stability selection* a énormément de difficulté à sélectionner le modèle nul, montrant des *FDR* très élevés notamment lorsque p augmente (globalement supérieur à 0,80 lorsque $p \geq 100$). De plus, le nombre de faux positifs sélectionnés est lui aussi élevé (par exemple : 23,9 faux positifs sélectionnés en moyenne pour la combinaison $n = p = 1000$).

Scénario alternatif 1 (q = 1). Dans ce scénario, le lasso-*cvl* réussit à identifier le seul biomarqueur actif se traduisant par des *FNR* proches de zéro, excepté lorsque la taille d'échantillon n est faible ($n = 100$; *FNR* = 0,47, 0,68 et 0,84 pour p = 10, 100 et 1000 biomarqueurs), mais également de nombreux faux positifs se traduisant par des *FDR* élevés compris entre 0,41 et 0,92 (Tableau 3.4, Figure 3.3). On retrouve donc le résultat mis en évidence au cours de l'étude de simulation préliminaire (Figure 3.1). L'ensemble des méthodes réduit le *FDR* à l'exception du *stability selection* dans certains cas.



Légende. losange : *stability selection*, carré : lasso adaptatif, rond : lasso-*cvl* et extensions (blanc : lasso-*cvl*, gris clair : lasso-*RIC*, gris : lasso-*AIC*, gris foncé : lasso-*1se*, noir : lasso-*pcvl*). Quantités moyennes basées sur 250 réplications.

Figure 3.3 : Taux de faux négatifs (*FNR*) en fonction du taux de fausses découvertes (*FDR*) dans le scénario alternatif 1 ($q = 1$)

Comme dit précédemment, le lasso-*cvl* a des difficultés à identifier le biomarqueur actif lorsque la taille d'échantillon n est faible ($FNR \geq 0{,}47$). Par conséquent, les extensions proposées, plus conservatrices, ont d'autant plus de mal à identifier ce biomarqueur bien que le lasso-*pcvl* et le lasso adaptatif limitent l'inflation du *FNR* (augmentation variant de +0,01 à +0,04). Lorsque la taille d'échantillon est suffisamment large ($n = 500$ ou 1000), le lasso-*pcvl*

et le lasso-*RIC* sont les méthodes qui fonctionnent le mieux en réduisant considérablement le *FDR* (réduction allant de -0,61 à -0,31 et -0,81 à -0,50 respectivement) avec une augmentation du *FNR* très faible voire nulle (jusqu'à +0,00 et +0,06 respectivement). Le lasso adaptatif n'augmente pas non plus le *FNR* en comparaison au lasso-*cvl*. Cependant la diminution de *FDR* est plus réduite que pour le lasso-*pcvl* et le lasso-*RIC*, et s'amenuise au fur et à mesure que p augmente (réduction variant de -0,30 pour $p = 10$ à -0,01 pour $p = 1000$). Le lasso-*AIC* réduit également le *FDR* mais avec, en contrepartie, une forte augmentation du *FNR* notamment lorsque n n'est pas suffisamment large ($n = 500$, augmentation du *FNR* de +0,10, +0,28 et +0,46, pour $p = 10$, 100 et 1000 respectivement). Le *stability selection* semble avoir des résultats qui dépendent du nombre de biomarqueurs candidats p. En effet, bien que comparable au lasso-*cvl* en matière de *FNR*, le *FDR* du *stability selection* est fortement réduit pour $p = 10$ (réduction allant de -0,55 à -0,53), légèrement réduit pour $p = 100$ (réduction allant -0,12 à -0,07) et légèrement augmenté pour $p = 1000$ (augmentation allant de +0,04 à +0,14) en comparaison au lasso-*cvl*. Enfin, le lasso-*1se* est extrêmement conservateur dans ce scénario avec des *FNR* constamment supérieurs à 0,67. Une explication de ce résultat est donnée dans la présentation des résultats des analyses de sensibilité plus tardivement dans cette section. Concernant les résultats du critère combiné G (i.e. moyenne géométrique de la sensibilité et de la spécificité, Tableau 3.5), ils suivent la même tendance que les résultats présentés ci-dessus. Pour rappel, G dépend du *FNR* ($= \mathrm{FN}/q$) et du *FPR* ($= \mathrm{FP}/(p - q)$) et pas du *FDR*. Ainsi, lorsque $q \ll p$ (et donc $q \ll p - q$), un faux négatif (impactant le *FNR* à la hauteur de $1/q$) pénalise beaucoup plus G qu'un faux positif (impactant le *FPR* à la hauteur de $1/(p - q)$). Par conséquent, les méthodes les moins conservatrices (*FDR* élevé et *FNR* bas), telles que le *stability selection*, montrent de bons résultats au travers de ce critère. Selon ce critère, le lasso-*pcvl* a de meilleures performances que le lasso-*cvl* dans 7 cas sur 9 (amélioration de +0,002 à +0,173) mais fonctionne moins bien lorsque la taille d'échantillon n est trop faible (réduction de -0,038 à -0,011).

Scénario alternatif 2 ($q > 1$). Dans ce scénario, le lasso-*cvl* affiche, une nouvelle fois, des *FDR* relativement élevés (compris entre 0,38 et 0,83) et des *FNR* très bas lorsque n est suffisamment large ($n = 500$ ou 1000, $FNR \leq 0{,}04$) (Tableau 3.6, Figure 3.4). A l'inverse, les *FNR* sont très importants lorsque la taille d'échantillon n est trop faible ($n = 100$, $FNR \geq 0{,}38$). Lorsque n augmente, le lasso-*pcvl* réduit grandement le *FDR* (réduction comprise entre -0,48 et -0,03) avec une très légère augmentation du *FNR* ($\leq +0{,}10$).

Tableau 3.4 : Taux de fausses découvertes et taux de faux négatifs (*FDR*/*FNR*) dans le scénario alternatif 1 ($q = 1$)

p	n	lasso-*cvl*	lasso-*1se*	lasso-*pcvl*	lasso-*AIC*	lasso-*RIC*	lasso adaptatif	*stability selection*
10	100	0,41 / 0,47	0,02 / 0,95	0,31 / 0,50	0,03 / 0,88	0,22 / 0,58	0,33 / 0,50	0,15 / 0,74
	500	0,59 / 0,02	0,00 / 0,80	0,15 / 0,02	0,13 / 0,12	0,10 / 0,02	0,35 / 0,02	0,05 / 0,16
	1000	0,60 / 0,00	0,00 / 0,67	0,07 / 0,00	0,13 / 0,00	0,09 / 0,00	0,30 / 0,00	0,03 / 0,08
100	100	0,52 / 0,68	0,02 / 0,99	0,47 / 0,72	0,00 / 0,98	0,43 / 0,81	0,52 / 0,69	0,83 / 0,69
	500	0,77 / 0,06	0,01 / 0,89	0,36 / 0,06	0,08 / 0,34	0,13 / 0,11	0,68 / 0,06	0,70 / 0,03
	1000	0,82 / 0,00	0,01 / 0,69	0,20 / 0,00	0,19 / 0,01	0,13 / 0,00	0,66 / 0,00	0,70 / 0,00
1000	100	0,50 / 0,84	0,01 / 0,99	0,49 / 0,86	0,00 / 0,99	0,47 / 0,94	0,50 / 0,84	0,35 / 0,96
	500	0,83 / 0,10	0,01 / 0,92	0,56 / 0,10	0,06 / 0,58	0,14 / 0,18	0,82 / 0,10	0,93 / 0,12
	1000	0,92 / 0,00	0,01 / 0,77	0,41 / 0,00	0,18 / 0,06	0,11 / 0,00	0,88 / 0,00	0,96 / 0,00

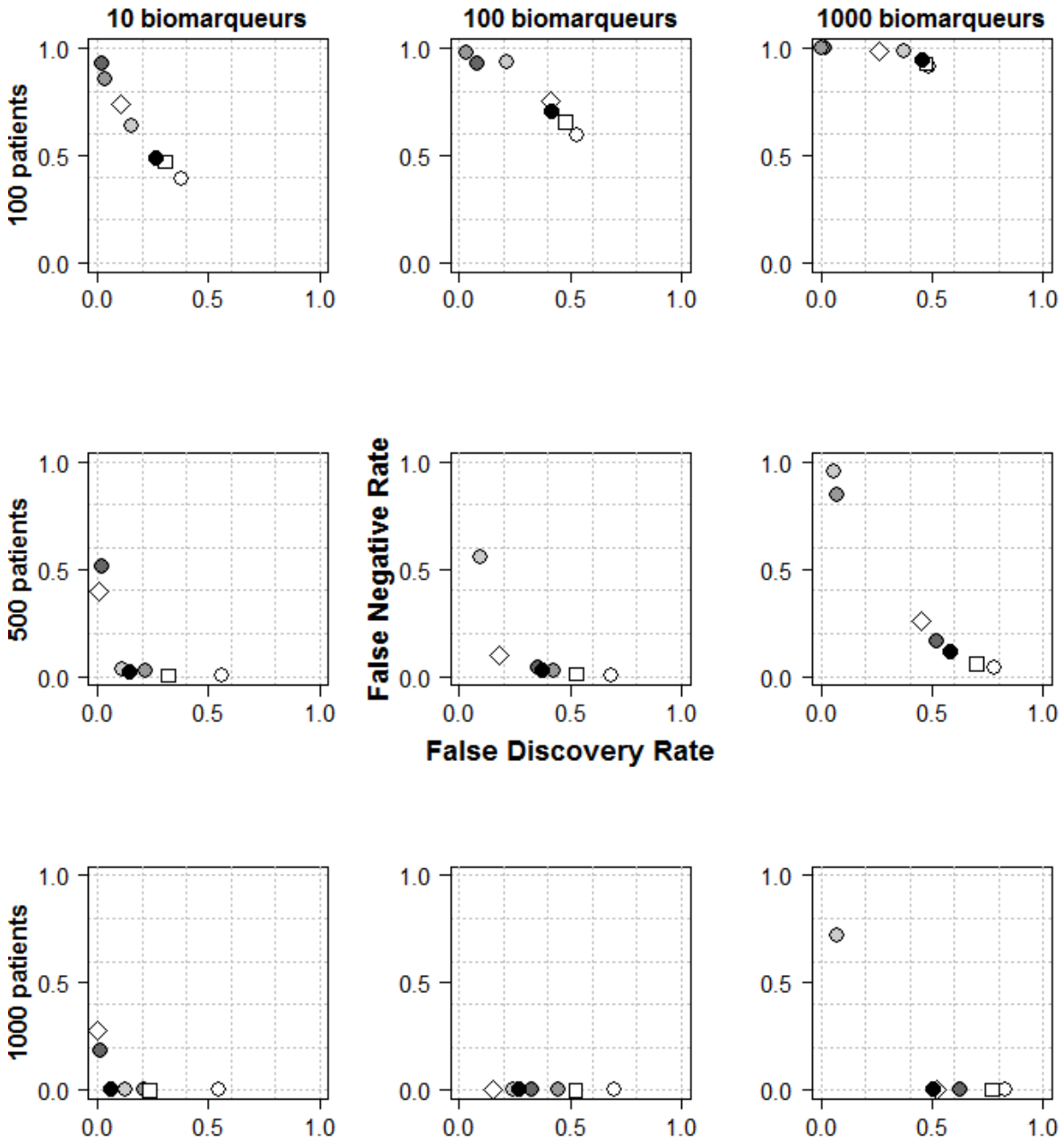
Légende. p : nombre de biomarqueurs, n : taille de l'échantillon. Quantités moyennes basées sur 250 réplications.

Tableau 3.5 : Moyenne géométrique de la sensibilité et de la spécificité (*G*) dans le scénario alternatif 1 ($q = 1$)

p	n	lasso-*cvl*	lasso-*1se*	lasso-*pcvl*	lasso-*AIC*	lasso-*RIC*	lasso adaptatif	*stability selection*
10	100	0,436	0,044	0,455	0,122	0,418	0,451	0,261
	500	0,804	0,204	0,956	0,863	0,965	0,897	0,834
	1000	0,818	0,328	0,991	0,976	0,990	0,923	0,916
100	100	0,314	0,008	0,276	0,020	0,188	0,305	0,304
	500	0,908	0,112	0,933	0,655	0,887	0,907	0,956
	1000	0,959	0,312	0,997	0,990	0,999	0,968	0,983
1000	100	0,155	0,008	0,144	0,012	0,060	0,155	0,036
	500	0,892	0,080	0,894	0,416	0,820	0,893	0,878
	1000	0,990	0,232	0,999	0,944	1,000	0,991	0,988

Légende. p : nombre de biomarqueurs, n : taille de l'échantillon. Quantités moyennes basées sur 250 réplications.

Cette amélioration est d'autant plus prononcée lorsque le nombre de biomarqueurs candidats p est faible ou modéré. A l'inverse, le lasso-*AIC* et le lasso-*RIC* montrent, cette fois-ci, de très mauvais résultats dans certains cas. Ces deux pénalisations sont trop strictes et tendent à ne pas sélectionner de nombreux biomarqueurs actifs lorsque le nombre de biomarqueurs candidats p est important pour une taille d'échantillon n modérée ($n = 500$, $FNR = 0{,}84$ et 0,95 respectivement) voire même large pour le lasso-*RIC* ($n = 1000$, $FNR = 0{,}72$). En effet, le multiplicateur θ_λ fixe du lasso-*AIC* (égal à 2) et dépendant de p_λ pour le lasso-RIC ($\theta_\lambda = 2\log(p_\lambda)$) semble trop important pour ces situations (Annexe A5). Le lasso adaptatif se comporte comme précédemment ($q = 1$), il permet de réduire le *FDR* mais cette réduction est de plus en plus faible au fur et à mesure que p augmente (réduction variant de -0,31 pour $p = 10$ à -0,02 pour $p = 1000$).



Légende. losange : *stability selection*, carré : lasso adaptatif, rond : lasso-*cvl* et extensions (blanc : lasso-*cvl*, gris clair : lasso-*RIC*, gris : lasso-*AIC*, gris foncé : lasso-*1se*, noir : lasso-*pcvl*). Quantités moyennes basées sur 250 réplications.

Figure 3.4 : Taux de faux négatifs (*FNR*) en fonction du taux de fausses découvertes (*FDR*) dans le scénario alternatif 2 ($q > 1$)

Le *stability selection*, quant à lui, semble donner de meilleurs résultats que dans le cas où $q = 1$. En effet, en présence de plusieurs biomarqueurs actifs, le *stability selection* réduit constamment le *FDR* sans forte augmentation du *FNR*, à l'exception du cas où $p = 10$ (*FNR* = 0,75, 0,38 et 0,31 pour $n = 100$, 500 et 1000). Enfin, le lasso-*1se* fonctionne beaucoup mieux dans ce scénario que dans le premier scénario alternatif. Malgré une pénalisation trop forte

pour un faible nombre de biomarqueurs (p = 10, *FNR* = 0,93, 0,51 et 0,18 pour n = 100, 500 et 1000), le lasso-*1se* réduit correctement le *FDR* (réduction variant de -0,38 à -0,21) sans forte augmentation du *FNR* (≤ +0,13) lorsque la taille de l'échantillon et le nombre de biomarqueurs sont suffisamment larges (n = 500 ou 1000, p = 100 ou 1000), Une nouvelle fois, la moyenne géométrique *G* (Tableau 3.7) permet de résumer l'information précédente en un seul critère et en attribuant plus de poids à une augmentation de *FNR* par rapport à une diminution de *FDR*. Cette fois-ci, le lasso-*pcvl* a de meilleures performances que le lasso-*cvl* dans 6 cas sur 9 (amélioration de +0,009 à +0,227) mais fonctionne moins bien lorsque la taille d'échantillon n est faible ou que le nombre de biomarqueurs p est important (réduction allant de -0,081 à -0,018).

Tableau 3.6 : Taux de fausses découvertes et taux de faux négatifs (*FDR*/*FNR*) dans le scénario alternatif 2 (q > 1)

q	p	n	lasso-*cvl*	lasso-*1se*	lasso-*pcvl*	lasso-*AIC*	lasso-*RIC*	lasso adaptatif	*stability selection*
2	10	100	0,38 / 0,39	0,02 / 0,93	0,27 / 0,48	0,04 / 0,86	0,15 / 0,64	0,30 / 0,47	0,10 / 0,74
		500	0,56 / 0,00	0,02 / 0,51	0,15 / 0,02	0,22 / 0,03	0,11 / 0,03	0,32 / 0,01	0,01 / 0,40
		1000	0,55 / 0,00	0,01 / 0,18	0,06 / 0,00	0,21 / 0,00	0,12 / 0,00	0,24 / 0,00	0,00 / 0,28
10	100	100	0,53 / 0,59	0,08 / 0,92	0,42 / 0,71	0,03 / 0,98	0,21 / 0,93	0,48 / 0,65	0,41 / 0,76
		500	0,68 / 0,01	0,35 / 0,04	0,38 / 0,03	0,42 / 0,03	0,10 / 0,56	0,53 / 0,01	0,18 / 0,10
		1000	0,70 / 0,00	0,32 / 0,00	0,27 / 0,00	0,45 / 0,00	0,24 / 0,00	0,53 / 0,00	0,15 / 0,00
20	1000	100	0,48 / 0,92	0,02 / 1,00	0,46 / 0,94	0,00 / 1,00	0,37 / 0,99	0,47 / 0,93	0,26 / 0,99
		500	0,78 / 0,04	0,52 / 0,17	0,58 / 0,11	0,07 / 0,84	0,05 / 0,95	0,70 / 0,06	0,45 / 0,26
		1000	0,83 / 0,00	0,62 / 0,00	0,51 / 0,00	0,52 / 0,00	0,07 / 0,72	0,77 / 0,00	0,52 / 0,01

Légende. q : nombre de biomarqueurs actifs, p : nombre de biomarqueurs, n : taille de l'échantillon. Quantités moyennes basées sur 250 réplications.

Tableau 3.7 : Moyenne géométrique de la sensibilité et de la spécificité (*G*) dans le scénario alternatif 2 (q > 1)

q	p	n	lasso-*cvl*	lasso-*1se*	lasso-*pcvl*	lasso-*AIC*	lasso-*RIC*	lasso adaptatif	*stability selection*
2	10	100	0,535	0,085	0,544	0,164	0,470	0,541	0,337
		500	0,744	0,506	0,953	0,919	0,956	0,893	0,693
		1000	0,760	0,821	0,987	0,950	0,974	0,924	0,790
10	100	100	0,539	0,124	0,458	0,045	0,210	0,506	0,461
		500	0,864	0,944	0,949	0,937	0,590	0,921	0,938
		1000	0,852	0,969	0,977	0,950	0,981	0,922	0,988
20	1000	100	0,204	0,005	0,164	0,001	0,058	0,189	0,053
		500	0,944	0,894	0,926	0,194	0,215	0,946	0,855
		1000	0,947	0,980	0,988	0,988	0,411	0,963	0,986

Légende. q : nombre de biomarqueurs actifs, p : nombre de biomarqueurs, n : taille de l'échantillon. Quantités moyennes basées sur 250 réplications.

Prédiction. Pour rappel, la statistique de concordance de Uno a été utilisée pour évaluer la capacité de prédiction des méthodes en matière de discrimination. Un horizon de $\tau = 5$ ans a été choisi et une analyse de sensibilité a été réalisée pour $\tau = 1$, 2, 3 et 4 ans. Aucune différence n'a été observée entre les différents horizons dans ces simulations (résultats non montrés). Pour le vrai modèle (i.e. modèle contenant uniquement les véritables biomarqueurs actifs), cette statistique augmente au fur et à mesure que n et q augmentent : de 0,521 pour n = 100 et q = 1 à 0,606 pour n = 1000 et q = 20 (Tableau 3.8). Ces résultats permettent de se rendre compte de l'impact de l'oubli de biomarqueurs actifs (i.e. faux négatifs) ou de la sélection de biomarqueurs inactifs (i.e. faux positifs). C'est le cas, par exemple, pour le lasso-*1se* dans les scénarios avec q = 1 où l'on note une diminution de la statistique de Uno, ou encore, le lasso-*AIC* ou lasso-*RIC* dans le scénario avec n = 500, p = 1000 et q = 20 où la statistique de Uno chute dramatiquement (de 0,605 pour le vrai modèle à 0,520 et 0,527 respectivement). Un autre exemple peut être mentionné avec le lasso-*cvl* qui identifie trop de faux positifs dans le scénario avec $n = p$ = 1000 et q = 20 (*FDR* = 0,83 ⇒ nombre de faux positifs moyen = 98) et pour lequel la statistique de Uno est de 0,589 comparé à 0,606 pour le vrai modèle bien que l'ensemble des biomarqueurs actifs soit sélectionné (*FNR* = 0).

Tableau 3.8 : Statistique de concordance de Uno dans les scénarios alternatifs ($q > 0$)

q	*p*	*n*	lasso-*cvl*	lasso-*1se*	lasso-*pcvl*	lasso-*AIC*	lasso-*RIC*	lasso adaptatif	*stability selection*	vrai modèle†
1	10	100	0,511	0,501	0,510	0,503	0,510	0,511	0,507	0,521
		500	0,523	0,506	0,525	0,522	0,525	0,523	0,523	0,526
		1000	0,523	0,508	0,524	0,524	0,524	0,524	0,522	0,525
	100	100	0,506	0,500	0,505	0,501	0,503	0,506	0,505	0,525
		500	0,517	0,503	0,521	0,517	0,523	0,517	0,519	0,525
		1000	0,519	0,508	0,523	0,523	0,524	0,520	0,520	0,524
	1000	100	0,503	0,500	0,503	0,500	0,504	0,503	0,501	0,523
		500	0,511	0,501	0,516	0,509	0,518	0,511	0,510	0,524
		1000	0,514	0,506	0,521	0,522	0,524	0,514	0,511	0,525
2	10	100	0,520	0,504	0,520	0,506	0,519	0,520	0,514	0,532
		500	0,533	0,518	0,535	0,534	0,535	0,534	0,526	0,536
		1000	0,535	0,529	0,535	0,535	0,535	0,535	0,529	0,536
10	100	100	0,540	0,512	0,539	0,506	0,524	0,540	0,541	0,572
		500	0,572	0,576	0,576	0,575	0,552	0,573	0,576	0,580
		1000	0,578	0,581	0,581	0,580	0,581	0,579	0,581	0,582
20	1000	100	0,518	0,501	0,516	0,500	0,511	0,517	0,508	0,591
		500	0,577	0,584	0,585	0,520	0,527	0,580	0,583	0,605
		1000	0,589	0,597	0,600	0,600	0,547	0,591	0,599	0,606

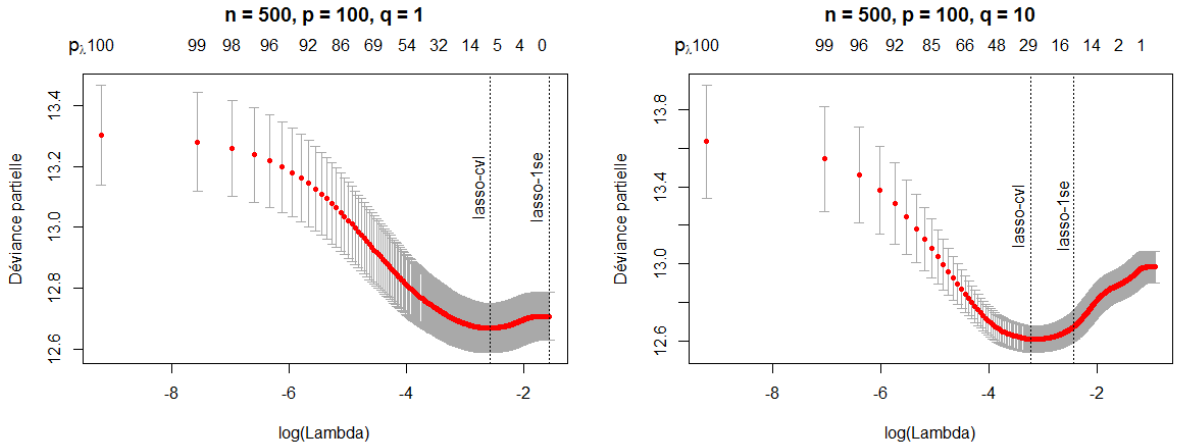
Légende. q : nombre de biomarqueurs actifs, p : nombre de biomarqueurs, n : taille de l'échantillon, † modèle contenant uniquement les biomarqueurs actifs. Quantités moyennes basées sur 250 réplications.

En moyenne, sur l'ensemble des scénarios alternatifs, le lasso-*pcvl* est la méthode qui a la capacité de prédiction la plus proche du vrai modèle (écart de 0,012), suivi par le lasso adaptatif (0,014), le lasso-*cvl* (0,015) et le *stability selection* (0,016). Pour rappel, les prédictions des patients sont calculées à partir de coefficients de régression non pénalisés pour une comparabilité exhaustive entre toutes les méthodes. Les résultats obtenus à partir des coefficients pénalisés sont présentés en Annexe A4, bien que légèrement meilleurs.

Analyses de sensibilité. L'étude de simulation présentée ci-dessus a été réalisée pour des données présentant une structure de corrélation entre biomarqueurs, des temps censurés et des tailles d'effets fixes pour les biomarqueurs actifs. Nous avons souhaité évaluer l'impact de ces différents paramètres.

Concernant la corrélation, nous avons précédemment dit que la pénalisation lasso perdait de sa qualité lorsque les données étaient corrélées. Nous avons donc souhaité quantifier ce résultat dans notre étude. Les résultats, présentés en Annexe A7, suggèrent que le lasso-*cvl* n'est pas réellement impacté par la présence ou non de corrélation entre les biomarqueurs en matière de capacité de sélection (*FDR* variant de -0,08 à +0,09 et *FNR* variant de -0,07 à +0,03 entre l'étude de simulation présentée ci-dessus et l'étude de simulation sans corrélation entre biomarqueurs). Les autres méthodes ne sont également pas impactées.

Concernant la censure, nous avons quantifié son impact sur la capacité de sélection du lasso-*cvl*. Les résultats, présentés en Annexe A8, suggèrent que le lasso-*cvl* n'est pas fortement impacté par la présence ou non de censure en ce qui concerne la capacité de sélection (*FDR* variant de -0,13 à +0,05 et *FNR* variant de -0,04 à +0,04 entre l'étude de simulation présentée ci-dessus et l'étude de simulation sans censure). A noter que dans notre étude de simulation, le taux de censure empirique est relativement bas ($\sim$ 15%). Il n'est donc pas possible d'extrapoler ces résultats pour des taux de censure plus élevés pour lesquels l'impact serait sans doute plus important et la capacité de sélection moins bonne. Les autres méthodes ne sont également pas impactées à l'exception du lasso-*1se* qui l'est fortement. En effet, lorsque $q = 1$, nous avions noté que le lasso-*1se* était extrêmement conservateur avec une sélection du modèle nul à tort dans de nombreux cas. Ce résultat n'est plus observé dans des scénarios sans censure (*FDR* variant de -0,53 à -0,11 et *FNR* variant de +0,07 à +0,80). On observe le même genre de fluctuation pour les deux autres scénarios ($q = 0$ et $q > 1$). La raison est que le choix de la pénalisation du lasso-*1se* dépend de la précision de la fonction de validation croisée. Ainsi, plus la précision est faible, plus l'écart standard sera grand et plus le choix du λ sera conservateur (Figure 3.5).



Légende. q : nombre de biomarqueurs actifs, p : nombre de biomarqueurs, n : taille de l'échantillon, barre verticale : intervalle de confiance à 95%. Données simulées comme décrit en section 3.4.2.

Figure 3.5 : Illustration de l'approche lasso-*1se* pour différents scénarios

Concernant les tailles d'effets des biomarqueurs actifs, nous avons initialement proposé une situation simple dans laquelle la taille d'effet est commune à l'ensemble des biomarqueurs actifs. Or, dans une application réelle, cela peut sembler peu réaliste et nous avons souhaité évaluer l'impact de ce choix. Nous avons donc généré des tailles d'effets variables choisies aléatoirement dans une distribution uniforme $\beta_j \sim U(-0{,}35, -0{,}10)$ correspondant à l'étendue des tailles d'effets observées dans notre application (voir section 3.5). Les résultats, présentés en Annexe A9, suggèrent que le lasso-*cvl* n'est pas réellement impacté par la variabilité des tailles d'effets des biomarqueurs actifs (*FDR* variant de -0,03 à +0,06 et *FNR* variant de -0,07 à -0,00 entre l'étude de simulation présentée ci-dessus et l'étude de simulation avec des tailles d'effets variables). Ce constat est sans doute dû au fait que l'étendue des tailles d'effets considérées (entre -0,35 et -0,10) se situe autour du β unique considéré dans les scénarios principaux ($\approx$ -0,22). La majorité des autres méthodes ne sont également pas impactées à l'exception des lasso-*AIC*, lasso-*RIC* et lasso-*BIC* qui le sont pour quelques scénarios. En effet, dans le scénario n = 500, p = 1000 et q = 20, le lasso-*AIC* réduit fortement le *FNR* lorsque les tailles d'effets sont variables (réduction de -0,34). Le constat est le même pour les lasso-*RIC* et lasso-*BIC* dans le scénario $n = p$ = 1000 et q = 20 avec une réduction du *FNR* de -0,48 et -0,49 respectivement dans le scénario avec des tailles d'effets variables.

3.5 Application

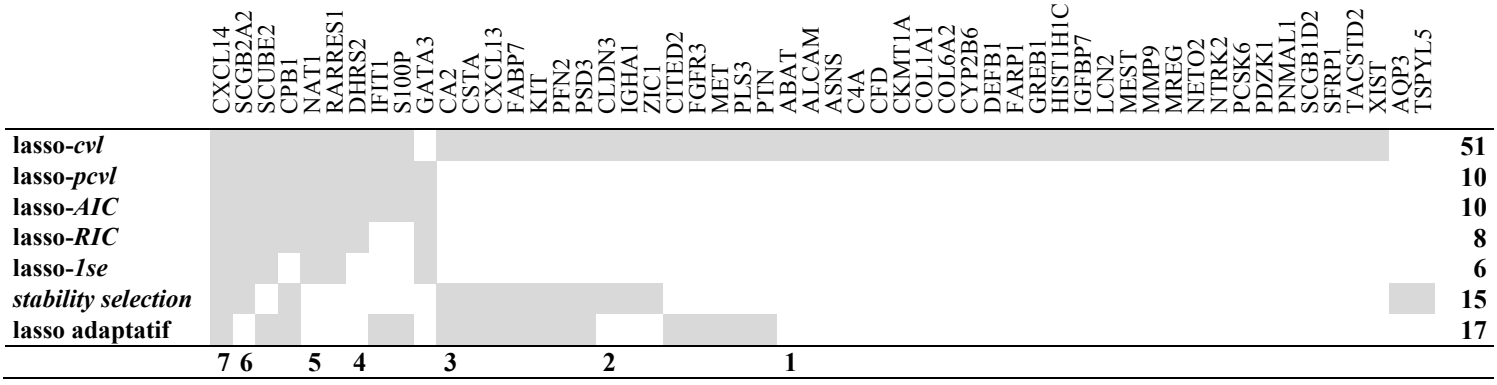
Les méthodes ont également été appliquées à des données réelles provenant d'une base publique (*Gene Expression Omnibus*, Barrett et al., 2005 ; Davis et Meltzer, 2007). Il s'agit

d'une base de données correspondant à n = 523 patientes atteintes d'un cancer du sein et pour lesquelles des données d'expression sont disponibles (puce Affymetrix). La majorité des tumeurs ont un statut HER2 négatif (93%), ER négatif (54%) et sont de grade élevé (59% de grade III). La survie sans récidive à distance à 1 et 5 ans est estimée à 91% [IC95% : 88%–93%] et 72% [IC95% : 68%–77%]. L'objectif de cette application a été d'identifier des biomarqueurs pronostiques de la survie sans récidive à distance.

Avant d'effectuer l'analyse, une étape préliminaire de normalisation des données et de filtrage des variables a été effectuée. Concernant la normalisation, la technique *fRMA* (i.e. *frozen robust multiarray analysis,* Mccall, Bolstad et Irizarry, 2010) a été utilisée pour rendre les différentes puces comparables. Cette technique tient compte d'informations « gelées » obtenues sur de grandes bases de données publiques disponibles (i.e. effet et variabilité de chaque probe) qui sont utilisées pour normaliser et résumer les nouvelles données. Concernant le nombre de biomarqueurs, celui-ci était trop important initialement (p = 22277 gènes). Ainsi, un filtrage préliminaire a été utilisé. Pour ne pas avoir de critère de filtrage basé sur l'association avec le critère de jugement, le choix a été de ne sélectionner que les gènes ayant une variabilité suffisamment élevée entre les patientes, c'est-à-dire un écart interquartile supérieur à 1. Il s'agit d'une technique de filtrage couramment utilisée dans ce contexte (Gentleman et al., 2006) qui permet de s'assurer que les différences entre patients soient peu impactées par du bruit aléatoire. Finalement, p = 1703 gènes standardisés ont été retenus selon ce critère pour l'analyse.

Les résultats de l'analyse, présentés en Tableau 3.9, montrent une grande variabilité entre les méthodes quant au nombre de biomarqueurs sélectionnés. Le lasso standard (lasso-*cvl*) est celui qui identifie le plus de biomarqueurs avec 51 sélectionnés.

Tableau 3.9 : Liste des biomarqueurs pronostiques de la survie sans récidive à distance dans le cancer du sein sélectionnés par les méthodes



Etant donné qu'il s'agit d'une application, il n'est pas possible de classer les gènes en « faux positifs » ou « vrais positifs ». Cependant, il est intéressant de noter que des gènes tels que SCUBE2, NAT1 ou encore ABAT ont déjà été identifiés dans la littérature comme étant pronostiques chez des patientes atteintes d'un cancer du sein (Cheng et al., 2009 ; Jansen et al., 2015 ; Tiang, Butcher et Minchin, 2015).

Les extensions pénalisées réduisent, comme attendu, le nombre de biomarqueurs sélectionnés : 10 pour le lasso-*pcvl* et le lasso-*AIC*, 8 pour le lasso-*RIC* et 6 pour le lasso-*1se.* Les biomarqueurs sélectionnés par ces extensions ne sont pas obligatoirement inclus dans la liste des biomarqueurs sélectionnés par le lasso-*cvl*, comme par exemple le gène GATA3 (Mehra et al., 2005). En effet, bien que le nombre de variables tende à décroître au fur et à mesure que la pénalisation augmente, il est possible que certaines variables apparaissent dans le modèle à partir d'une certaine pénalisation. Ce constat reste très marginal et n'est observé généralement qu'en cas de forte corrélation entre variables. Concernant les deux variantes du lasso que nous avons étudiées (lasso adaptatif et *stability selection*), bien que l'intersection des biomarqueurs sélectionnés par ces deux méthodes et par le lasso-*cvl* soit forte (respectivement 17/17 et 13/15), cela est moins le cas pour le lasso-*pcvl* (respectivement 5/17 et 3/15).

Pour tenter d'avoir un élément de comparaison entre les méthodes, nous avons comptabilisé le nombre de fois où ces biomarqueurs ont été inclus dans des signatures de gènes connues comme étant pronostiques dans le cancer du sein à partir de la base de données *GeneSigDB* (Culhane et al., 2012). Par exemple, les gènes discutés jusqu'alors tels que SCUBE2, S100P, ABAT ou encore GATA3 sont respectivement répertoriés dans 53, 35, 29 et 59 signatures publiées. Parmi les 51 gènes identifiés par le lasso-*cvl*, les 9 également identifiés par le lasso-*pcvl* sont répertoriés dans plus (test de Wilcoxon, p-value = 0,02) de signatures (médiane : 22, étendue : 14–53) que ceux non identifiés par le lasso-*pcvl* (médiane : 16, étendue : 6–31).

3.6 Conclusion

Dans un contexte de données de grande dimension, la régression lasso est largement utilisée pour estimer les coefficients d'un modèle en effectuant simultanément une sélection de variables. Comme toute régression pénalisée, elle dépend d'un paramètre de pénalisation (λ)

pour lequel aucun consensus n'est établi quant à son estimation. Très souvent, ce paramètre est déterminé à partir de la technique de validation croisée (*cvl*). Cependant, de récentes publications ont montré que ce choix tend à sélectionner un nombre important de faux positifs. Or, ces faux positifs sont une préoccupation majeure car ils conduisent à réduire la fiabilité des résultats de recherche. Dans ce travail de thèse, nous avons observé que lorsque la taille d'échantillon est suffisamment importante, la pénalisation lasso permet de : (i) bien sélectionner les vrais biomarqueurs actifs mais en sélectionnant également de nombreux faux positifs et (ii) bien discriminer les biomarqueurs actifs par rapport aux biomarqueurs inactifs (i.e. en moyenne ces derniers sont écartés du modèle pour des pénalisations plus faibles par rapport aux actifs). Ainsi, nous avons suggéré une pénalisation supplémentaire (*pcvl*), basée sur l'allure de la *cvl*, pour sélectionner moins de biomarqueurs qui sont donc essentiellement des faux positifs. Dans la littérature, d'autres pénalisations existent mais ne dépendent pas de la fonction à pénaliser (ici la *cvl*) et peuvent ainsi être inefficaces dans certains cas. Comme ces autres pénalisations, la *pcvl* est une extension empirique qui ne satisfait pas la propriété *oracle* (voir section 2.3).

Pour montrer les bonnes performances de notre extension, nous l'avons comparée à de nombreuses approches au travers d'une étude de simulation couvrant de nombreux scénarios. Les résultats suggèrent que notre extension permet de réduire, parfois considérablement, le nombre de faux positifs sélectionnés sans diminuer fortement la sélection des biomarqueurs actifs. De plus, dans les différents scénarios considérés, notre extension n'est jamais trop conservatrice. Dans cette étude, un autre résultat notable est que la pénalisation lasso n'est pas fortement impactée par la présence de corrélation contrairement à ce qui est dit dans la littérature. Cela peut être lié à la structure de corrélation que nous avons choisie. Enfin, nous avons évalué deux variantes du lasso (lasso adaptatif et *stability selection*) qui satisfont la propriété *oracle*, mais qui n'affichent pas de très bons résultats en pratique sur des échantillons de tailles finies. De plus, ces méthodes nécessitent de faire des choix supplémentaires sur certains paramètres (e.g. estimation des poids pour le lasso adaptatif ou encore nombre de faux positifs attendus pour la *stability selection*) pour lesquels aucun consensus n'est établi mais qui peuvent avoir des impacts importants sur les résultats. Par exemple, nous avons montré dans cette étude que l'utilisation de modèles univariés ou multivariés soumis à la pénalisation ridge pour estimer les poids du lasso adaptatif donnait de mauvais résultats, particulièrement dans les scénarios nuls (Annexe A2). Une explication possible est qu'en principe le lasso adaptatif utilise des poids pour amplifier les différences

d'effets entre les biomarqueurs. Cependant, bien que cela puisse améliorer la puissance de la méthode dans les scénarios alternatifs, cela peut être un inconvénient dans les scénarios nuls pour lesquels les différences aléatoires sont amplifiées artificiellement conduisant à augmenter la sélection de faux positifs. A notre connaissance, la capacité de sélection de cette approche n'avait jamais été évaluée dans les scénarios nuls.

Nous avons également illustré ces méthodes sur des données réelles chez des patientes ayant un cancer du sein. Bien entendu, il n'est pas possible d'évaluer la performance des méthodes au travers de cette application car nous ne sommes pas en mesure d'identifier les gènes réellement actifs. Néanmoins, cette illustration permet d'apprécier la forte disparité des gènes sélectionnés par les différentes méthodes.

Dans ce travail de thèse, nous nous sommes essentiellement focalisés sur la régression lasso. Cependant, de nombreuses autres variantes que celles considérées sont proposées dans la littérature. Un bref résumé a été proposé par Tibshirani (2011) pour montrer l'évolution de la régression lasso et le développement de ses variantes au cours du temps, telles que par exemple le lasso groupé (*group lasso*, Yuan et Lin, 2006, que nous aborderons au Chapitre 4), le lasso fusionné (*fused lasso,* Tibshirani et Saunders, 2005) ou encore le lasso graphique (*graphical lasso,* Yuan et Lin, 2007). D'autres variantes sont disponibles mais n'ont pas été présentées dans cet article. C'est le cas par exemple du lasso bootstrapé (*bolasso*, Bach, 2008), du lasso détendu (*relaxed lasso*, Meinshausen, 2007), du lasso non paramétrique *cosso* (*component selection and smoothing operator*, Lin et Halabi, 2013) ou encore à partir de procédures de permutations (Sabourin, Valdar et Nobel, 2015). Il serait sans doute très intéressant d'avoir un travail de revue plus large permettant de comprendre les avantages et inconvénients de ces différentes variantes et de savoir, d'un point de vue pratique, comment les implémenter de manière cohérente.

En conclusion de ce travail, nous suggérons d'appliquer notre extension *pcvl*, simple à implémenter, pour estimer le paramètre de pénalisation λ de la régression lasso lorsque l'on cherche à minimiser le nombre de faux positifs tout en limitant les faux négatifs.

Chapitre 4 Identification de biomarqueurs prédictifs de l'effet du traitement

Dans mon second projet de thèse, l'objectif était de s'intéresser aux approches permettant d'identifier des biomarqueurs visant à prédire l'effet d'un traitement pour un patient donné (Ternès et al., 2016b). En cancérologie, on nomme généralement ces biomarqueurs comme étant prédictifs ou modificateurs de l'effet du traitement. Ces biomarqueurs peuvent être utilisés pour sélectionner une sous-population à traiter, et idéalement, pour anticiper le risque individuel au cours du temps selon la décision thérapeutique.

4.1 Point de vue statistique

En 2005, Rothwell a mis en évidence le fait que la meilleure approche pour identifier un biomarqueur prédictif est de tester son interaction avec le traitement. Ainsi, une approche qui pourrait être utilisée pour tester des interactions est d'implémenter le modèle suivant contenant le traitement, les biomarqueurs et leur interaction :

$$h(t, \boldsymbol{T}, \boldsymbol{X}) = h_0(t) \exp\left(\alpha \boldsymbol{T} + \sum_{j=1}^{p} \beta_j \boldsymbol{X_j} + \sum_{j=1}^{p} \gamma_j \boldsymbol{X_j T} \right) \quad (4.1)$$

avec α, β_j et γ_j, les coefficients de régression traduisant respectivement l'effet du traitement $\boldsymbol{T}$ entre le bras contrôle et le bras expérimental, l'effet du biomarqueur standardisé $\boldsymbol{X_j}$ et l'interaction entre ces deux $\boldsymbol{X_jT}$, pour chaque biomarqueur j = 1, ..., p. En d'autres termes, la première somme de (4.1) correspond à une composante pronostique représentant l'effet propre des biomarqueurs (i.e. estimant le pronostic général d'un patient indépendamment de l'effet traitement) et la seconde somme de (4.1) correspond à une composante prédictive estimant l'effet du traitement sur le critère de jugement selon les valeurs des biomarqueurs pour un patient donné. Dans ce modèle, nous codons le traitement $\boldsymbol{T}$ en -0,5 (pour le bras contrôle) et +0,5 (pour le bras expérimental). Dans ce travail, nous nous sommes focalisés exclusivement sur l'identification d'une signature prédictive, correspondant à la seconde somme de (4.1), visant à prédire l'effet du traitement. Nous ne nous sommes pas intéressés à la partie pronostique (ce qui a été fait précédemment dans le Chapitre 3), ni directement à la prédiction individuelle de la probabilité de survie pour un patient donné (pour cela, voir Chapitre 5).

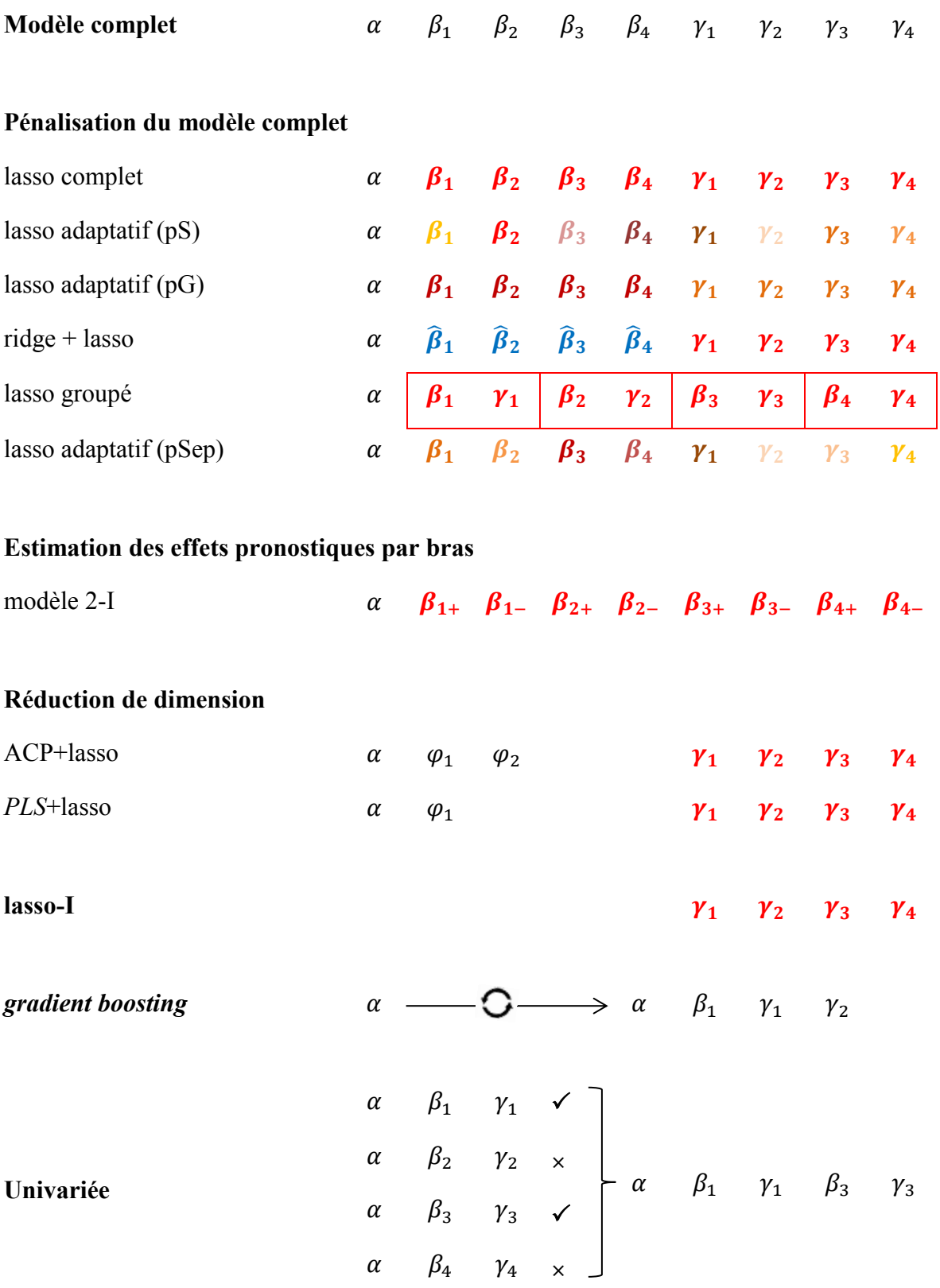
4.2 Approches possibles

Le nombre d'approches possibles pour identifier des biomarqueurs prédictifs est théoriquement infini. Dans ce travail, douze approches ont été proposées et sont décrites ci-dessous. Une schématisation de ces approches est présentée en Figure 4.1, dont l'interprétation se clarifiera tout au long de cette section.

4.2.1 Pénalisation des effets propres et des interactions (5 approches)

La première classe d'approches considérée prévoit d'implémenter le modèle (4.1) et d'y appliquer une pénalisation pour pallier la dimension très importante du modèle (i.e. $2p + 1$, avec $p \gg n$). Cette pénalisation, nommée $p(\lambda, \boldsymbol{\beta}, \boldsymbol{\gamma})$, est appliquée aux effets propres et aux interactions du modèle. Nous avons proposé différentes paramétrisations de cette pénalisation. Dans tous les cas, le traitement (i.e. effet α) est considéré comme une variable non pénalisée.

La première paramétrisation considérée utilise la pénalisation lasso sur l'ensemble des effets propres et des interactions du modèle pour effectuer une sélection de variables et ainsi, identifier des biomarqueurs prédictifs. Cependant, cette approche ne permet pas de garder la structure hiérarchique du modèle, c'est-à-dire que l'effet propre d'un biomarqueur peut être écarté du modèle même si son interaction avec le traitement est encore présente.



Légende. bleu : pénalisation ridge, couleurs chaudes : pénalisation lasso avec ou sans (rouge) pondération, × : biomarqueur non sélectionné, ✓ : biomarqueur sélectionné, ↻ : processus itératif.

Figure 4.1 : Représentation schématique des approches proposées pour identifier des biomarqueurs prédictifs

Bien, Taylor et Tibshirani (2013) ont montré que l'absence d'effet propre pour une interaction peut affecter l'interprétation de $\boldsymbol{\gamma}$ ou encore la calibration du modèle, mais qu'il pourrait s'agir d'un problème mineur lorsque l'on s'intéresse à la sélection de variables. Nous avons appelé cette approche **lasso complet** dans la suite du manuscrit et sa pénalisation peut s'écrire :

$$p(\lambda, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda \left(\sum_{j=1}^{p} |\beta_j| + \sum_{j=1}^{p} |\gamma_j| \right).$$

Dans ce cas, les effets propres ($\boldsymbol{\beta}$) et les interactions ($\boldsymbol{\gamma}$) sont pénalisés équitablement avec le même paramètre de pénalisation λ.

En réalité, il est possible que les effets propres des biomarqueurs et leur interaction avec le traitement aient des tailles d'effets très différentes et, dans ce cas, il n'est pas judicieux de les pénaliser de la même façon. Pour remédier à cela, nous proposons d'ajouter des poids aux différentes variables afin de pondérer leur pénalisation. Cette approche correspond strictement au lasso adaptatif (Zhang et Lu, 2007 ; Zou, 2006) présenté en section 3.3.2. Dans ce contexte d'identification d'interactions, nous estimons les poids dans une étape préliminaire à partir du modèle complet (4.1) soumis à la pénalisation ridge pour les effets propres des biomarqueurs et leur interaction avec le traitement. L'utilisation de la pénalisation ridge permet d'estimer les coefficients de régression $\tilde{\beta}_j^R$ et $\tilde{\gamma}_j^R$ sans pour autant faire de sélection de variables (voir section 2.3 pour plus de détails), tout en restant dans un modèle multivarié qui tient compte des corrélations entre biomarqueurs. Pour l'estimation de ces poids, deux stratégies ont été évaluées : des poids spécifiques par coefficient et des poids groupés par type d'effets. L'utilisation de poids spécifiques correspond exactement au lasso adaptatif standard dans le sens où chaque variable a un poids spécifique correspondant à l'inverse de la valeur absolue du coefficient de régression estimé lors de l'étape préliminaire. On note donc la pénalisation

$$p(\lambda, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda \left(\sum_{j=1}^{p} \frac{1}{|\tilde{\beta}_j^R|} |\beta_j| + \sum_{j=1}^{p} \frac{1}{|\tilde{\gamma}_j^R|} |\gamma_j| \right).$$

La stratégie des poids groupés consiste quant à elle à estimer un poids unique pour l'ensemble des effets propres et un poids unique pour l'ensemble des interactions. Pour chacun, le poids est estimé comme l'inverse de la moyenne des valeurs absolues des coefficients de régression et on note la pénalisation

$$p(\lambda, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda \left(\frac{1}{\tilde{\beta}^R} \sum_{j=1}^{p} |\beta_j| + \frac{1}{\tilde{\gamma}^R} \sum_{j=1}^{p} |\gamma_j| \right), \qquad \tilde{\beta}^R = \frac{1}{p} \sum_{j=1}^{p} |\tilde{\beta}_j^R|, \tilde{\gamma}^R = \frac{1}{p} \sum_{j=1}^{p} |\tilde{\gamma}_j^R|.$$

Ces deux approches adaptatives sont nommées respectivement **lasso-pS** (i.e. poids spécifiques) et **lasso-pG** (i.e. poids groupés) dans la suite du manuscrit.

Les trois approches présentées jusqu'ici (lasso complet ou lassos adaptatifs) proposent d'effectuer une sélection de variables sur les effets propres des biomarqueurs et leur interaction avec le traitement, ce qui peut être sous-optimal en raison du non-respect de la hiérarchie du modèle. Pour corriger cela, nous proposons de soumettre les effets propres à la pénalisation ridge plutôt qu'à la pénalisation lasso. Cela permet alors de garder l'ensemble des effets propres des biomarqueurs dans le modèle final tout en contrôlant un éventuel surajustement. La pénalisation lasso reste néanmoins appliquée aux interactions pour identifier les biomarqueurs prédictifs potentiels. Idéalement, nous obtenons donc un modèle soumis aux pénalisations ridge et lasso sur des coefficients différents. Cette double pénalisation peut s'écrire

$$p(\lambda, \lambda_2, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_2 \sum_{i=1}^{p} \beta_j^2 + \lambda \sum_{j=1}^{p} |\gamma_j|.$$

En pratique, cette optimisation simultanée des deux paramètres de pénalisation peut être très coûteuse en temps de calcul et avoir des effets négatifs sur la généralisation des résultats. Etant donné que l'objectif principal concerne la bonne sélection des interactions, nous accordons plus d'importance à l'optimisation du paramètre de pénalisation du lasso (λ) au détriment d'un choix plus grossier du paramètre λ_2. Ce parti pris a également été choisi par Zou et Hastie (2005) dans un contexte assez similaire lorsque les pénalisations ridge et lasso sont appliquées aux mêmes variables : l'*elastic net*. En effet, les auteurs ont proposé d'estimer de manière plus fine le paramètre λ pour une grille arbitraire de valeurs pour le paramètre λ_2 (par exemple : 0,001, 0,01, 0,1, 1, 10, 100, 1000, etc.). Dans notre cas, cette stratégie reste complexe à mettre en place car, à notre connaissance, seul le package `penalized` sous `R` peut être utilisé pour faire cela en combinant différentes options, mais l'implémentation reste néanmoins fastidieuse. C'est pourquoi nous avons proposé une solution alternative visant à estimer, dans un premier temps, les coefficients de régression $\boldsymbol{\beta}$ dans un modèle contenant uniquement les effets propres des biomarqueurs. Puis, dans un deuxième temps, ces

coefficients sont fixés en *offset* dans le modèle final comprenant le traitement et les interactions entre les biomarqueurs et le traitement, qui sont quant à elles soumises à la pénalisation lasso. Une analyse de sensibilité a été réalisée pour un nombre très réduit de 10 biomarqueurs et les résultats montrent peu de différences entre les deux approches (double optimisation vs. *offset*). Bien entendu, ce résultat n'est pas extrapolable pour une situation de grande dimension mais le temps de calcul extrêmement long de la double optimisation ne nous a pas permis de faire cette comparaison pour un nombre plus important de biomarqueurs. Cette approche est nommée **ridge+lasso** dans la suite du manuscrit.

Enfin, pour clore les approches proposant de pénaliser à la fois les effets propres des biomarqueurs et leur interaction, nous nous sommes intéressés au **lasso groupé** (Yuan et Lin, 2006). Cette approche peut être vue comme un compromis entre le fait de garder l'ensemble des effets propres des biomarqueurs du modèle (comme le ridge+lasso) et le fait de faire de la sélection sur ces effets, quitte à ne pas respecter la contrainte de hiérarchie du modèle (comme le lasso complet ou le lasso adaptatif). En effet, contrairement au lasso standard, le lasso groupé n'effectue pas de sélection de variables au niveau individuel mais uniquement au niveau de groupes spécifiés par avance. En d'autres termes, les variables présentes dans un groupe sont soit toutes inclues soit toutes exclues du modèle pour un certain paramètre λ. Le lasso standard est donc un cas particulier du lasso groupé pour lequel la taille des groupes est égale à 1. Dans notre cas, nous pouvons ainsi considérer p groupes de taille 2 contenant pour chacun l'effet propre du biomarqueur et son interaction avec le traitement. La pénalisation peut s'écrire :

$$p(\lambda, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda \sum_{j=1}^{p} \left\| (\beta_j, \gamma_j)^T \right\|_2 = \lambda \sum_{j=1}^{p} \sqrt{\beta_j^2 + \gamma_j^2}$$

avec $||\cdot||_2$ la norme Euclidienne. Ainsi, le modèle final contiendra obligatoirement autant d'effets propres que d'interactions. Cette pénalisation permet donc de respecter la contrainte de hiérarchie du modèle tout en réduisant fortement sa dimension grâce à la sélection de variables sur les effets propres.

4.2.2 Régression sans effets propres (1 approche)

Récemment, Tian et al. (2014) ont proposé une approche pour estimer les interactions entre les biomarqueurs et le traitement sans avoir à modéliser les effets propres. Les auteurs suggèrent de simplement coder le traitement en $\pm 0{,}5$ et de le multiplier avec les biomarqueurs centrés. On note ces nouvelles variables, dites modifiées, $\boldsymbol{M} = \boldsymbol{XT}$ et on note le modèle

$$h(t, \boldsymbol{M}) = h_0(t) \exp\left(\sum_{j=1}^{p} \gamma_j M_j\right). \tag{4.2}$$

Finalement, ce modèle correspond exactement au modèle (4.1) dans lequel seuls les termes d'interaction sont présents et sans effets propres ni du traitement ni des biomarqueurs. Dans le cas de données de grande dimension, les auteurs proposent également d'appliquer la pénalisation lasso sur ces variables modifiées pour faire de la sélection. Cette approche est nommée **lasso-I** (i.e. interactions) dans la suite du manuscrit.

4.2.3 Réduction de dimension des effets propres (2 approches)

Plutôt que de considérer l'ensemble des effets propres des biomarqueurs (comme le modèle (4.1) illustré en section 4.2.1) ou n'en considérer aucun (comme le modèle (4.2) illustré en section 4.2.2), un compromis peut être de réduire la dimension de la matrice des effets propres en la remplaçant par des combinaisons linéaires. Comme présenté au Chapitre 2, l'utilisation de ces techniques ne permet pas de faire de la sélection de variables mais d'en réduire la dimension en utilisant une projection dans un sous-espace vectoriel. Dans le contexte présent, l'intérêt est de faire de la sélection de variables sur les interactions et non sur les effets propres des biomarqueurs. Il ne serait donc pas problématique de réduire la dimension des effets propres tant que l'on maintient, au travers des combinaisons linéaires, une part de variabilité de la matrice initiale suffisante pour que l'information pronostique contenue puisse expliquer les différences dans les niveaux de risque de base. Le modèle final peut donc s'écrire

$$h(t, \boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{X}) = h_0(t) \exp\left(\alpha \boldsymbol{T} + \sum_{c=1}^{C} \varphi_c \boldsymbol{Z_c} + \sum_{j=1}^{p} \gamma_j \boldsymbol{X_j T}\right), \tag{4.3}$$

avec $\boldsymbol{Z} = (\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_C})$, les C premières combinaisons linéaires des effets propres des biomarqueurs. Le modèle (4.3) contient bien moins de variables que le modèle (4.1) étant

donné que $C \ll p$, et la pénalisation lasso n'est appliquée que sur les termes d'interaction. Dans ce travail, deux techniques de réduction de dimension ont été évaluées. La première, l'analyse en composantes principales ou ACP (Hotelling, 1933), consiste à appliquer une transformation orthogonale sur des variables potentiellement corrélées afin de créer de nouvelles variables linéaires non corrélées que l'on appelle composantes principales. Un inconvénient de cette approche est qu'elle ne tient pas compte de la variable à expliquer (ici la variabilité des temps de survie) dans la création des composantes. Plusieurs approches ont été proposées pour pallier cette lacune comme l'ACP supervisée (Bair et al., 2006) ou encore la régression des moindres carrés partiels (en anglais : *partial least square* ou *PLS*, Martens et Naes, 1989). La première consiste à mettre en place l'analyse en composantes principales sur un sous-ensemble de variables filtrées au préalable comme étant les plus associées avec le critère de jugement. Cette approche a été évaluée dans une étude préliminaire et peu de différence a été observée par rapport à l'ACP non supervisée (résultats non montrés). Ainsi, l'ACP supervisée n'a pas été retenue pour la suite de ce travail car, avec une performance équivalente, elle nécessite de faire un choix supplémentaire par rapport à l'ACP non supervisée (i.e. seuil de préfiltrage). La régression des moindres carrés partiels détermine quant à elle des combinaisons linéaires visant à maximiser la variance des prédicteurs (comme l'ACP) mais aussi à maximiser leur corrélation avec le critère de jugement. Comme l'algorithme *PLS* proposé initialement est basé sur une relation linéaire entre le critère de jugement et les covariables, il n'est pas directement applicable au modèle de Cox. Des extensions ont alors été proposées pour le cas d'un critère de jugement censuré (Bastien et Tenenhaus, 2001 ; Park, Tian et Kohane, 2002 ; Nygård et al., 2006). Les deux approches sont nommées **ACP+lasso** et ***PLS*+lasso** dans la suite du manuscrit.

4.2.4 Le *boosting* (1 approche)

L'approche *boosting* (Schapire, 1990) est une technique de régression basée sur un processus itératif qui consiste à mettre à jour des estimateurs afin d'obtenir un bon modèle final en maximisant la log-vraisemblance partielle. La caractéristique principale de cette approche est que le processus de mise à jour est réalisé lentement et à chaque itération l'amélioration du modèle est faible. Dans le cas de données de grande dimension, la version « composante par composante » du *boosting* est privilégiée. Cette version consiste à partir d'un modèle nul, et seule l'estimation du coefficient de régression permettant la meilleure amélioration du modèle est mise à jour à chaque itération. Contrairement à la sélection de variables pas-à-pas, les autres coefficients du modèle sont fixes. Il existe différents types de *boosting* se différenciant

par la façon dont les coefficients de régression sont mis à jour à chaque itération. Dans ce travail, nous nous sommes intéressés au *gradient boosting* (Bühlmann et Yu, 2003 ; Friedman, 2001) pour lequel la mise à jour possible est calculée en modélisant le gradient de la log-vraisemblance partielle sur chaque prédicteur. Le prédicteur qui est le plus corrélé avec le gradient est utilisé, après multiplication par un paramètre de pénalisation ν, pour mettre à jour son estimation dans le modèle de Cox. Il a été montré que les résultats obtenus par la version « composante par composante » du boosting sont similaires à ceux du lasso (Bühlmann et Yu, 2003). Cependant, en présence de forte corrélation entre les biomarqueurs, le *boosting* a des résultats plus stables que ceux du lasso (Hastie et al., 2007). Dans notre utilisation du *boosting*, la contrainte de hiérarchie n'est pas obligatoirement maintenue. Pour évoquer cette méthode, nous avons gardé la notation anglaise ***gradient boosting***, plus explicite, dans la suite du manuscrit.

4.2.5 Approche univariée en combinaison avec un contrôle du *FDR* (1 approche)

Une approche plus simple conceptuellement peut être d'évaluer la force d'interaction entre le traitement et chaque biomarqueur en utilisant un modèle univarié (Michiels, Potthoff et George, 2011). Cela consiste à effectuer pour chaque biomarqueur un modèle contenant trois composantes – le traitement, le biomarqueur et l'interaction entre ces deux variables – puis à tester la significativité de l'interaction à partir d'un test de Wald. Bien entendu, cela revient à réaliser p tests statistiques. Pour pallier le problème de la multiplicité des tests, plusieurs techniques d'ajustement sont proposées dans la littérature comme le critère de Benjamini et Hochberg (1995) qui permet de contrôler le *FDR*. La valeur *P* ajustée (P^{BH}) du biomarqueur j est égale à $P_j^{\mathrm{BH}} = \min(P_{j+1}^{\mathrm{BH}}, (p \times P_j) / j), j = (p - 1), ..., 1$ avec $P_1 \leq P_2 \leq ... \leq P_p$ les valeurs *P* non ajustées et rangées dans l'ordre croissant, et $P_p^{\mathrm{BH}} = P_p$. Dans cette étude, les biomarqueurs retenus sont ceux pour lesquels P^{BH} est inférieure ou égale à 0,05 pour faire de la sélection. Pour une meilleure capacité de prédiction, un modèle multivarié final contenant le traitement ainsi que l'ensemble des effets propres et interactions des biomarqueurs sélectionnés a été mis en place par la suite pour estimer les différents coefficients. Cette approche est nommée **approche univariée** dans la suite du manuscrit.

4.2.6 Estimation des effets pronostiques par bras (2 approches)

Jusqu'à présent, les approches proposées permettent d'identifier un biomarqueur prédictif dès lors que son interaction avec le traitement est retenue dans le modèle. Par définition, un biomarqueur prédictif correspond à un biomarqueur pour lequel son effet pronostique au sein

de chaque bras de traitement est différent. Ainsi, plutôt que de considérer le modèle (4.1) il est également envisageable d'estimer ces effets pronostiques par bras de traitement, soit dans un modèle pronostique pour chaque bras de traitement, soit dans un seul modèle contenant deux termes d'interaction. La première option a déjà été proposée par Zhao et al. (2013), puis critiquée par Kang, Janes et Huang (2014) qui montraient que le meilleur modèle dans chacun des deux bras de traitement ne donne pas nécessairement le meilleur modèle pour prédire l'effet du traitement. En effet, cette approche peut rater des biomarqueurs qui sont fortement associés à l'effet du traitement mais qui ont des effets propres très modérés, et réciproquement des biomarqueurs peuvent avoir des effets propres très importants mais pour autant des interactions très modérées avec le traitement. La seconde approche, non proposée à notre connaissance, consiste à considérer un seul modèle dans lequel on remplace les effets propres par un second terme d'interaction pour le bras contrôle. Le modèle peut s'écrire

$$h(t, \boldsymbol{T}, \boldsymbol{X}) = h_0(t) \exp\left(\alpha \boldsymbol{T} + \sum_{j=1}^{p} \gamma_{j+} \boldsymbol{X_j} \mathbf{I}(\boldsymbol{T} = +0{,}5) + \sum_{j=1}^{p} \gamma_{j-} \boldsymbol{X_j} \mathbf{I}(\boldsymbol{T} = -0{,}5) \right), \qquad (4.4)$$

avec γ_{j+} et γ_{j-} représentant l'effet pronostique du biomarqueur *j* chez les patients du groupe expérimental et du groupe contrôle respectivement. Dans ce modèle, la pénalisation lasso est appliquée simultanément sur les coefficients $\boldsymbol{\gamma}_+$ et $\boldsymbol{\gamma}_-$ et un biomarqueur est considéré comme prédictif si et seulement si uniquement un des deux effets pronostiques est retenu dans le modèle. Ce choix peut sembler très strict car même dans le cas où les deux effets ne sont pas nuls mais très différents, le biomarqueur ne sera pas retenu comme étant prédictif. Une alternative serait de considérer, *a posteriori*, un test de contraste permettant de tester la différence entre ces deux effets. Cependant, la variabilité des estimateurs n'est pas estimable correctement dans un modèle pénalisé, ni dans un modèle réestimé sans pénalisation (nous reparlerons plus en détails de ce point dans le Chapitre 5). De plus, cela serait en désaccord avec les autres méthodes pour lesquelles aucun test *a posteriori* n'a été réalisé pour tester la significativité des interactions sélectionnées par la pénalisation lasso. Cette approche est nommée **modèle 2-I** (i.e. deux interactions) dans la suite du manuscrit.

Suite à la présentation du modèle (4.4), un questionnement est de savoir s'il n'est pas préférable d'utiliser l'estimation des effets pronostiques dans chacun des bras de traitement pour améliorer le processus de sélection dans le modèle de base (4.1). Un choix possible est alors d'estimer les effets pronostiques $\boldsymbol{\gamma}_+$ et $\boldsymbol{\gamma}_-$ des biomarqueurs à partir du modèle (4.4),

pour lequel ces variables sont soumises à une pénalisation ridge, puis d'utiliser ces estimations pour définir des poids dans l'utilisation de la pénalisation lasso du modèle (4.1). En d'autres termes, il s'agit du même principe que l'approche lasso-pS précédemment présentée (section 4.2.1) mais ici les poids sont estimés à partir du modèle (4.4) plutôt que directement à partir du modèle (4.1). Notre choix de poids et donc de pénalisation du modèle (4.1) a été :

$$p(\lambda, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda \left(\sum_{j=1}^{p} \frac{1}{\left|\tilde{\gamma}_{j+}^{R} + \tilde{\gamma}_{j-}^{R}\right| + \left|\tilde{\gamma}_{j+}^{R} - \tilde{\gamma}_{j-}^{R}\right|} |\beta_j| + \sum_{j=1}^{p} \frac{1}{\left|\tilde{\gamma}_{j+}^{R} - \tilde{\gamma}_{j-}^{R}\right|} |\gamma_j| \right).$$

avec $\tilde{\gamma}_{j+}^{R}$ et $\tilde{\gamma}_{j-}^{R}$ les effets pronostiques du biomarqueur j estimés dans le modèle (4.4) et soumis à la pénalisation ridge, et β_j et γ_j son effet propre et son interaction avec le traitement dans le modèle (4.1). Le poids attribué aux interactions est basé sur la différence des effets pronostiques entre les deux bras de traitement, se traduisant par un effet prédictif. Le poids attribué aux effets propres est basé sur la somme des effets pronostiques des deux bras de traitement à laquelle est ajouté le poids attribué aux interactions. Cet ajout a été proposé pour favoriser la sélection de l'effet propre du biomarqueur lorsque celui-ci interagit fortement avec le traitement et ainsi favoriser la contrainte de hiérarchie dans le modèle. Cette approche est nommée **lasso-pSep** (i.e. poids spécifiques basés sur les effets pronostiques) dans la suite du manuscrit.

4.3 Étude de simulation

Pour évaluer les douze approches discutées dans la section 4.2, une étude de simulation a été mise en place. Dans cette étude, nous étudions principalement la capacité des méthodes à détecter les vrais biomarqueurs prédictifs dans un modèle de Cox à grande dimension. Nous évaluons également leur capacité à prédire l'effet du traitement à partir des signatures développées.

4.3.1 Génération des données

Pour cette étude de simulation, la génération des données est semblable à celle présentée dans le premier travail de thèse (cf. section 3.4.1). Pour la notation, nous appelons p le nombre de biomarqueurs candidats, q_{Po} le nombre de biomarqueurs pronostiques et q_{Pe} le nombre de biomarqueurs prédictifs. Une structure de corrélation autorégressive a été implémentée au

travers de 20 blocs à l'intérieur desquels la corrélation entre deux biomarqueurs est égale à $\rho_{jj'} = 0{,}7^{|j-j'|}$. L'allocation du traitement a été réalisée aléatoirement avec une probabilité de 0,5 pour chacun des deux bras. Enfin, comme précédemment, le temps de survie a été généré à partir d'une distribution de Weibull avec un paramètre de forme a = 1 (i.e. distribution exponentielle correspondant à des risques constants au cours du temps) et des temps de censure indépendants ont été générés de façon à ce que les patients soient inclus uniformément pendant une période de 3 ans, puis suivis pendant une période de 2 ans jusqu'à la clôture de l'étude.

4.3.2 Choix des scénarios

Nous avons considéré une nouvelle fois différents scénarios nuls (i.e. aucun biomarqueur n'interagit avec le traitement) et alternatifs (i.e. au moins un biomarqueur interagit avec le traitement). Un résumé de ces scénarios est présenté en Tableau 4.1. Nous avons généré trois scénarios nuls (scénarios 1–3) : un premier dans lequel il n'y a aucun signal (scénario nul complet), un deuxième avec un fort effet du traitement ($\alpha = \log(0{,}5) \approx -0{,}69$) et un troisième comprenant q_{Po} biomarqueurs pronostiques et non prédictifs ($\beta_j = \log(0{,}5)$). Nous avons proposé également trois scénarios alternatifs (scénarios 4–6). Les scénarios 4 et 5 contiennent uniquement des biomarqueurs prédictifs ($\gamma_j = \log(0{,}5)$), respectivement 1 et q_{Pe}. Enfin, le scénario 6 est plus réaliste car il contient à la fois des biomarqueurs pronostiques (q_{Po}) et des biomarqueurs prédictifs (q_{Pe}). Ces six scénarios ont été générés pour n = 500 patients et p = 500 (scénarios 1–6a) ou p = 1000 (scénarios 1–6b) biomarqueurs avec respectivement q_{Po} = q_{Pe} = 10 ou 20. Dans ces scénarios, le taux de censure était compris entre 10% et 40%. Nous avons volontairement choisi de mettre en place une étude de simulation présentant des situations « optimistes » (i.e. tailles d'effets importantes, taux de censure faible) pour bien comprendre le fonctionnement des méthodes et bien apprécier leurs différences. Afin d'être plus proche d'un exemple réel, nous avons aussi proposé une étude de simulation avec des caractéristiques semblables à notre application (section 4.4). Pour cela, nous avons fortement augmenté le taux de censure (i.e. 60–80%) et réduit la taille d'effet des biomarqueurs actifs (i.e. $\beta_j \sim U(-0{,}5, -0{,}1)$ et $\gamma_j \sim U(-0{,}7, -0{,}1)$). Ces scénarios sont nommés scénarios 1–6c. Enfin, afin d'être plus exhaustif, nous avons également considéré différentes formes de la distribution de Weibull pour la génération des temps de survie (risques décroissants (a = 0,5 < 1) ou croissants (a = 2 > 1) au cours du temps) et différentes structures de corrélation entre les biomarqueurs actifs.

Tableau 4.1 : Scénarios de l'étude de simulation

			Médiane de survie de base (années)		Hazard Ratio		Probabilité de censure moyenne	
			m_0		h(X=1)/h(X=0)			
250 répétitions par scénario			T^-	T^+	T^-	T^+	T^-	T^+
p = 500 biomarqueurs	(**1a**)	Aucun effet	1,0	1,0	1,0	1,0	0,10	0,11
	(**2a**)	Effet du traitement seul	1,0	2,0	1,0	1,0	0,10	0,31
	(**3a**)	10 biomarqueurs pronostiques	1,0	1,0	0,5		0,30	0,30
	(**4a**)	1 biomarqueur prédictif	1,0	1,0	1,0	0,5	0,10	0,15
	(**5a**)	10 biomarqueurs prédictifs	1,0	1,0	1,0	0,5	0,10	0,29
	(**6a**)	10 biomarqueurs prédictifs + 10 biomarqueurs pronostiques	1,0	1,0	1,0 0,5	0,5	0,30	0,35
p = 1000 biomarqueurs	(**1b**)	Aucun effet	1,0	1,0	1,0	1,0	0,11	0,10
	(**2b**)	Effet du traitement seul	1,0	2,0	1,0	1,0	0,11	0,31
	(**3b**)	20 biomarqueurs pronostiques	1,0	1,0	0,5		0,35	0,35
	(**4b**)	1 biomarqueur prédictif	1,0	1,0	1,0	0,5	0,11	0,15
	(**5b**)	20 biomarqueurs prédictifs	1,0	1,0	1,0	0,5	0,11	0,35
	(**6b**)	20 biomarqueurs prédictifs + 20 biomarqueurs pronostiques	1,0	1,0	1,0 0,5	0,5	0,35	0,39
p = 500 biomarqueurs	(**1c**)	Aucun effet	1,0	1,0	1,0	1,0	0,65	0,66
	(**2c**)	Effet du traitement seul	1,0	2,0	1,0	1,0	0,65	0,81
	(**3c**)	10 biomarqueurs pronostiques	1,0	1,0	$\exp(\beta_{Po})$		0,60	0,61
	(**4c**)	1 biomarqueur prédictif	1,0	1,0	1,0	$\exp(\beta_{Pe})$	0,66	0,64
	(**5c**)	10 biomarqueurs prédictifs	1,0	1,0	1,0	$\exp(\beta_{Pe})$	0,66	0,59
	(**6c**)	10 biomarqueurs prédictifs + 10 biomarqueurs pronostiques	1,0	1,0	1,0 $\exp(\beta_{Po})$	$\exp(\beta_{Pe})$	0,61	0,58

Légende. T^- : bras contrôle, T^+ : bras expérimental, X : biomarqueur, $\beta_{Po} \sim U(-0,5, -0,1)$, $\beta_{Pe} \sim U(-0,7, -0,1)$

4.3.3 Critères d'évaluation

Comme lors du premier travail de cette thèse, l'objectif principal est d'évaluer la capacité de sélection des méthodes. Cependant, un vrai positif (voir Tableau 3.1) correspond ici à un biomarqueur prédictif. Les critères tels que le taux de fausses découvertes (*FDR*, pour rappel *FDR* = VP/(VP+FP)) et le taux de faux négatifs (*FNR*, pour rappel *FNR* = FN/(VP+FN)) en termes d'interactions sélectionnées ont donc été évalués pour juger de cette performance. De plus, dans les scénarios présentant des biomarqueurs pronostiques, nous avons rapporté le nombre de faux positifs (faux biomarqueurs prédictifs) qui étaient pronostiques (i.e. faux positifs pronostiques ou FPp) pour évaluer si certaines méthodes étaient amenées à classer plus facilement à tort un biomarqueur comme prédictif selon qu'il ait ou non un rôle pronostique. Pour la quasi-totalité des méthodes, la sélection de variables a été réalisée en estimant le paramètre de pénalisation à partir du critère *cvl* (voir section 4.3.4). Or, ce critère basé principalement sur une erreur de prédiction n'est pas obligatoirement synonyme de bon choix pour la sélection. Comme nous l'avons vu dans le premier travail de thèse, ce critère peut parfois conduire à une forte sélection de faux positifs et l'utilisation d'un critère sélectionnant moins de biomarqueurs peut souvent réduire fortement les faux positifs sans forte augmentation des faux négatifs (section 3.2). Pour évaluer la capacité de sélection des

méthodes, il serait donc intéressant d'avoir un critère qui soit indépendant du paramètre choisi pour la sélection. Une possibilité est de calculer l'aire sous la courbe *ROC* (*Receiver Operating Characteristic*) ou *AUC* afin d'avoir un aperçu plus global de la capacité des méthodes à discriminer les biomarqueurs actifs et inactifs. L'*AUC* permet d'évaluer le pouvoir de discrimination (i.e. classement des biomarqueurs en positifs ou négatifs) parmi les biomarqueurs actifs (sensibilité) et les biomarqueurs inactifs (spécificité) pour différents seuils (dans notre cas, différentes pénalisations). Ainsi, plus la méthode est capable d'éliminer les biomarqueurs inactifs avant les biomarqueurs actifs (i.e. pour des pénalisations plus faibles), plus son *AUC* est élevée, c'est-à-dire proche de 1. A l'inverse, une *AUC* est petite, c'est-à-dire proche de 0,5, lorsque la méthode n'a aucun pouvoir de discrimination dans la sélection des biomarqueurs. Cependant, dans le contexte présent, le critère *AUC* est peu adéquat en raison du nombre très important de biomarqueurs inactifs au regard des biomarqueurs actifs, ce qui conduit à une *AUC* suroptimiste en raison d'une spécificité qui est proche de 1 pour la plupart des valeurs possibles du paramètre de pénalisation. Une alternative est de mesurer l'aire sous la courbe dite *precision-recall* (ou *AUPRC*, Bleakley, Biau et Vert, 2007) qui combine les critères 1-*FNR* et 1-*FDR* (au lieu de la spécificité). Tout comme l'*AUC*, une *AUPRC* est comprise entre 0 et 1 et doit être maximisée. Une absence de discrimination correspond à une *AUPRC* qui tend vers la proportion de paires identiques parmi l'ensemble des paires disponibles.

En plus d'évaluer la capacité des méthodes à effectuer une bonne sélection, il semble important d'évaluer la force du signal des signatures identifiées, i.e. de savoir si les gènes identifiés et leur estimation permettent de suffisamment discriminer le bénéfice du traitement chez des patients futurs. Pour cela, pour chaque jeu de données (d'apprentissage) sur lequel le modèle a été estimé, un autre jeu de données (de validation, $\boldsymbol{X}^{\mathrm{V}}$) avec les mêmes caractéristiques que ce premier a été généré, et un score d'interaction $\hat{\eta}_i$ a été calculé pour chaque patient *i* de ce nouveau jeu de données comme étant

$$\hat{\eta}_i = \sum_{j \in Q_{Pe}} \hat{\gamma}_j \times X_{i,j}^{\mathrm{V}},$$

avec $\hat{\gamma}_j$ les coefficients des interactions retenues dans l'échantillon d'apprentissage, et $X_{i,j}^{\mathrm{V}}$ les données correspondant au patient *i* issu de l'échantillon de validation. Ainsi, plus le score est bas et plus le bénéfice du traitement (en échelle relative) sera fort étant donné que l'effet du traitement est représenté par un coefficient négatif (i.e. réduction du risque). A partir de ce

score, nous avons proposé un critère basé sur une différence de statistiques de concordance pour évaluer la force d'interaction des signatures que nous avons appelé statistique ΔC pour la suite du manuscrit

$$\Delta\mathrm{C}(\tau, \boldsymbol{\eta}) = C_{\mathrm{Uno}}(\tau, \boldsymbol{\eta}, \boldsymbol{T} = +0{,}5) - C_{\mathrm{Uno}}(\tau, \boldsymbol{\eta}, \boldsymbol{T} = -0{,}5).$$

Comme précédemment, nous avons choisi la statistique de Uno, C_{Uno}, comme présentée en (3.3) qui est l'une des mesures les moins biaisées en présence de données censurées. Dans le Chapitre 3, nous avons utilisé ce critère pour évaluer la force pronostique des signatures établies. Ici, nous avons évalué ce critère par bras de traitement pour évaluer la différence de force pronostique entre les deux bras de traitement synonyme de force prédictive. On remarque bien à travers ce critère que nous ne nous sommes pas directement intéressés à la capacité de prédiction du critère de jugement car nous ne nous intéressons pas à la valeur de la statistique C mais seulement au différentiel. Un critère semblable avait déjà été proposé par Schemper (1988) en généralisant le test d'interaction non paramétrique de Patel et Hoel (1973). Ces critères non paramétriques permettent d'être robuste dans les situations ne respectant pas l'hypothèse de proportionnalité des risques du modèle de Cox.

Enfin, pour évaluer le surapprentissage des modèles, les scores d'interaction et la différence de statistiques de concordance ont également été calculés pour les patients ayant contribué à l'estimation du modèle (i.e. échantillon d'apprentissage).

4.3.4 Implémentation

Les méthodes discutées dans ce chapitre ont été implémentées à partir du logiciel `R` en utilisant les packages suivants : `glmnet` (Friedman et al., 2010 ; Friedman et al., 2016) pour les régressions pénalisées ridge et lasso, `grplasso` (Meier, 2015) pour le lasso groupé, `mboost` (Hofner et al., 2014) pour le *gradient boosting*, `corpcor` (Schäfer et al., 2015) pour la construction des composantes principales ou encore `plsRcox` (Bertrand, Maumy-Bertrand et Meyer, 2015) pour l'approche *PLS* dans le cadre d'un modèle de survie. A noter, le package `grplasso` est implémenté uniquement pour les modèles linéaires généralisés. Nous avons transformé la base de données (Annexe B21) de façon à ce que l'utilisation du modèle de Poisson par intervalle (ici, intervalle de 2 mois) corresponde à une approximation du modèle de Cox à risque de base constant par morceaux, se rapprochant bien de l'estimateur de Breslow (Pawitan, 2013 ; Whitehead, 1980). Concernant l'estimation des différents paramètres, nous avons utilisé la technique de validation croisée (5 sous-échantillons) avec le

critère *cvl* de Verweij et van Houwelingen (1993, 1994) pour estimer les paramètres λ pour le lasso et λ_2 pour le ridge, mais aussi le nombre de composantes C pour l'analyse en composantes principales ou pour la régression *PLS* comme déjà réalisé par Bøvelstad et al. (2007), ou encore le nombre d'itérations de l'approche *gradient boosting* (le paramètre de pénalisation étant fixé à $v = 0{,}1$, valeur par défaut). A l'issue des résultats préliminaires, nous avons pu voir que l'utilisation du critère *cvl* pour la régression *PLS* donnait de très mauvais résultats. Nous avons donc décidé de ne garder que la première composante pour cette approche dans notre étude de simulation, pour laquelle les résultats étaient meilleurs.

4.3.5 Résultats

En se basant sur leurs performances au travers de l'étude de simulation, les méthodes peuvent être classées en plusieurs groupes.

Tout d'abord, l'approche **univariée** contrôle bien le *FDR* (ou erreur de type-I) dans les scénarios nuls (Tableau 4.2) mais elle est très conservatrice dans les scénarios alternatifs surtout en présence de biomarqueurs pronostiques (scénario 6). Cela se traduit par un *FDR* bas, un *FNR* élevé (Figure 4.2) et également une faible puissance statistique : 0,59 et 0,43 dans les scénarios 6a et 6b. Par conséquent, la force d'interaction des signatures développées est faible (Figure 4.3). Bien que la faible puissance statistique soit liée au seuil choisi (ici, valeur p ajustée inférieure à 0,05), cela n'explique pas les mauvaises performances de cette méthode. En effet, comme nous l'indique le critère *AUPRC*, indépendant du seuil choisi, cette approche a beaucoup de difficulté à distinguer les biomarqueurs inactifs des biomarqueurs actifs (*AUPRC* souvent inférieure à 0,5, Tableau 4.3). Une analyse de sensibilité présentée en Annexe A10 montre qu'en l'absence de corrélation entre les biomarqueurs, cette approche respecte bien le *FDR* au seuil choisi de 5% dans les scénarios alternatifs.

Le second groupe de méthodes regroupe celles qui ne permettent pas de faire de la sélection de variables sur les effets propres des biomarqueurs : le **lasso-I**, les approches effectuant une réduction de dimension (**ACP+lasso** et ***PLS*+lasso**) et le **ridge+lasso**. Dans les scénarios nuls, leurs *FDR* sont modérés (0,24–0,58) voire élevés (0,49–0,61) avec respectivement peu (scénarios 1–3a et 1–3b) ou beaucoup (scénarios 1–3c) de censure (Tableau 4.2).

Tableau 4.2 : Proportion de modèles sélectionnant au moins un biomarqueur pour l'ensemble des méthodes

		univariée	lasso-I	ACP+lasso	*PLS*+lasso	ridge+lasso	lasso groupé	modèle 2-I	lasso complet	lasso-pS	lasso-pG	lasso-pSep	*gradient boosting*
SCENARIOS NULS	**Scénario 1a**	0,07	0,39	0,38	0,36	0,39	0,48	0,41	0,01	0,14	0,00	0,42	0,68
	Scénario 2a	0,06	0,35	0,43	0,38	0,39	0,56	0,44	0,01	0,12	0,00	0,37	0,66
	Scénario 3a	0,06	0,37	0,24	0,41	0,47	1,00	1,00	0,88	0,20	0,00	0,32	1,00
	Scénario 1b	0,06	0,38	0,35	0,32	0,38	0,52	0,40	0,01	0,12	0,00	0,36	0,68
	Scénario 2b	0,04	0,41	0,43	0,43	0,38	0,52	0,38	0,02	0,16	0,00	0,38	0,69
	Scénario 3b	0,08	0,45	0,27	0,42	0,58	1,00	1,00	0,98	0,32	0,00	0,55	1,00
	Scénario 1c	0,05	0,56	0,57	0,50	0,56	0,56	0,40	0,03	0,25	0,00	0,51	0,73
	Scénario 2c	0,04	0,55	0,56	0,61	0,56	0,46	0,37	0,00	0,13	0,00	0,44	0,65
	Scénario 3c	0,06	0,53	0,49	0,58	0,60	1,00	1,00	0,51	0,63	0,00	0,73	1,00
SCENARIOS ALTERNATIFS	**Scénario 4a**	1,00	1,00	1,00	0,99	1,00	1,00	1,00	1,00	1,00	0,15	1,00	1,00
	Scénario 5a	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,66	1,00	1,00
	Scénario 6a	0,59	0,90	0,80	0,94	0,99	1,00	1,00	1,00	1,00	0,55	1,00	1,00
	Scénario 4b	1,00	1,00	1,00	0,99	1,00	1,00	1,00	0,98	1,00	0,19	1,00	1,00
	Scénario 5b	1,00	1,00	0,98	1,00	1,00	1,00	1,00	0,98	1,00	0,78	1,00	1,00
	Scénario 6b	0,43	0,78	0,64	0,90	0,98	1,00	1,00	1,00	1,00	0,28	1,00	1,00
	Scénario 4c	0,27	0,66	0,70	0,65	0,68	0,72	0,76	0,20	0,48	0,00	0,65	0,76
	Scénario 5c	0,74	0,94	0,93	0,89	0,98	0,99	1,00	0,73	0,96	0,10	0,97	0,99
	Scénario 6c	0,51	0,86	0,83	0,82	0,91	1,00	1,00	0,97	0,99	0,14	0,99	1,00

Légende. Scénario nul : erreur de type-I ou *FDR*. Scénarios alternatifs : puissance. Quantités moyennes basées sur 250 réplications.

Tableau 4.3 : Capacité de sélection des interactions dans les scenarios alternatifs

		univariée	lasso-I	ACP+lasso	*PLS*+lasso	ridge+lasso	lasso groupé	modèle 2-I	lasso complet	lasso-pS	lasso-pG	lasso-pSep	*gradient boosting*
Scénario 4a	Biomarqueurs sélectionnés	4	14	13	14	14	24	18	2	2	0	3	7
	VP / FP(FPp)	1 / 3(0)	1 / 13(0)	1 / 12(0)	1 / 13(0)	1 / 13(0)	1 / 23(0)	1 / 18(0)	1 / 1(0)	1 / 1(0)	0 / 0(0)	1 / 2(0)	1 / 6(0)
	AUPRC	1,00	0,98	0,98	0,95	0,98	0,99		0,99	0,99	0,99	0,99	0,99
Scénario 5a	Biomarqueurs sélectionnés	14	43	42	49	49	100	78	20	20	3	18	37
	VP / FP(FPp)	6 / 8(0)	9 / 34(0)	9 / 33(0)	9 / 40(0)	9 / 40(0)	10 / 90(0)	9 / 69(0)	9 / 11(0)	9 / 11(0)	2 / 1(0)	9 / 9(0)	10 / 27(0)
	AUPRC	0,53	0,63	0,61	0,64	0,68	0,71		0,78	0,78	0,78	0,81	0,68
Scénario 6a	Biomarqueurs sélectionnés	2	25	15	29	37	109	99	23	14	1	13	38
	VP / FP(FPp)	1 / 1(0)	4 / 21(1)	4 / 11(0)	6 / 23(1)	7 / 30(1)	10 / 99(10)	9 / 90(0)	9 / 14(0)	8 / 7(0)	1 / 0(0)	8 / 5(0)	10 / 29(1)
	AUPRC	0,27	0,25	0,37	0,38	0,43	0,21		0,75	0,69	0,71	0,71	0,62
Scénario 4b	Biomarqueurs sélectionnés	4	13	13	14	14	25	18	2	2	0	2	7
	VP / FP(FPp)	1 / 3(0)	1 / 12(0)	1 / 12(0)	1 / 13(0)	1 / 13(0)	1 / 24(0)	1 / 17(0)	1 / 1(0)	1 / 1(0)	0 / 0(0)	1 / 1(0)	1 / 6(0)
	AUPRC	1,00	0,99	0,98	0,94	0,98	1,00		0,98	0,99	0,98	0,99	0,99
Scénario 5b	Biomarqueurs sélectionnés	19	55	50	64	62	127	101	26	31	6	26	46
	VP / FP(FPp)	8 / 11(0)	14 / 41(0)	14 / 36(0)	16 / 48(0)	16 / 46(0)	20 / 107(0)	18 / 83(0)	13 / 13(0)	15 / 16(0)	4 / 2(0)	15 / 12(0)	16 / 31(0)
	AUPRC	0,42	0,45	0,45	0,49	0,51	0,53		0,63	0,63	0,62	0,65	0,51
Scénario 6b	Biomarqueurs sélectionnés	1	20	12	30	39	124	110	28	14	1	13	35
	VP / FP(FPp)	1 / 1(0)	4 / 16(1)	4 / 8(0)	7 / 22(1)	9 / 30(1)	19 / 106(20)	17 / 93(1)	13 / 15(1)	8 / 6(0)	0 / 0(0)	8 / 6(0)	13 / 22(1)
	AUPRC	0,19	0,16	0,26	0,27	0,28	0,17		0,54	0,48	0,47	0,48	0,35
Scénario 4c	Biomarqueurs sélectionnés	1	8	8	6	8	12	14	0	1	0	2	4
	VP / FP(FPp)	0 / 0(0)	0 / 8(0)	0 / 8(0)	0 / 6(0)	0 / 7(0)	1 / 12(0)	1 / 13(0)	0 / 0(0)	0 / 1(0)	0 / 0(0)	0 / 2(0)	0 / 4(0)
	AUPRC	0,36	0,34	0,34	0,25	0,34	0,44		0,33	0,35	0,25	0,37	0,40
Scénario 5c	Biomarqueurs sélectionnés	3	28	27	23	32	51	51	4	9	0	8	21
	VP / FP(FPp)	1 / 1(0)	5 / 23(0)	5 / 22(0)	4 / 19(0)	5 / 27(0)	7 / 44(0)	7 / 44(0)	2 / 2(0)	4 / 6(0)	0 / 0(0)	3 / 5(0)	5 / 16(0)
	AUPRC	0,27	0,29	0,32	0,27	0,33	0,44		0,33	0,35	0,27	0,35	0,35
Scénario 6c	Biomarqueurs sélectionnés	2	22	17	19	27	73	75	6	8	0	7	27
	VP / FP(FPp)	1 / 1(0)	3 / 19(0)	3 / 14(0)	3 / 16(0)	4 / 22(1)	7 / 66(8)	7 / 68(3)	3 / 4(0)	3 / 5(0)	0 / 0(0)	3 / 4(0)	5 / 22(0)
	AUPRC	0,21	0,22	0,26	0,24	0,29	0,19		0,32	0,33	0,26	0,34	0,31

Légende. VP : vrai positif, FP : faux positif, FPp : faux positif pronostique, *AUPRC* : *area under the precision-recall curve*. Quantités moyennes basées sur 250 réplications.

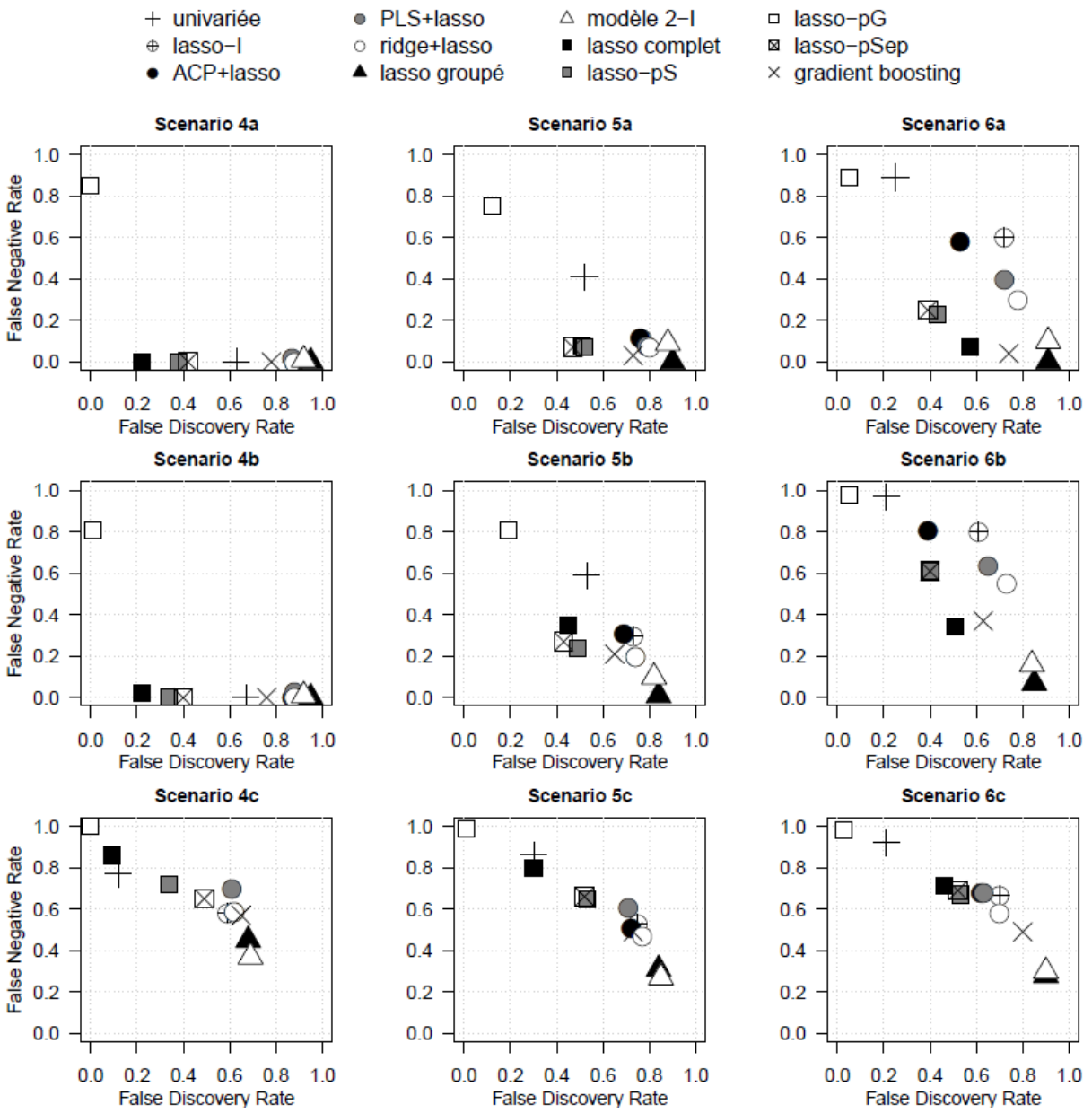
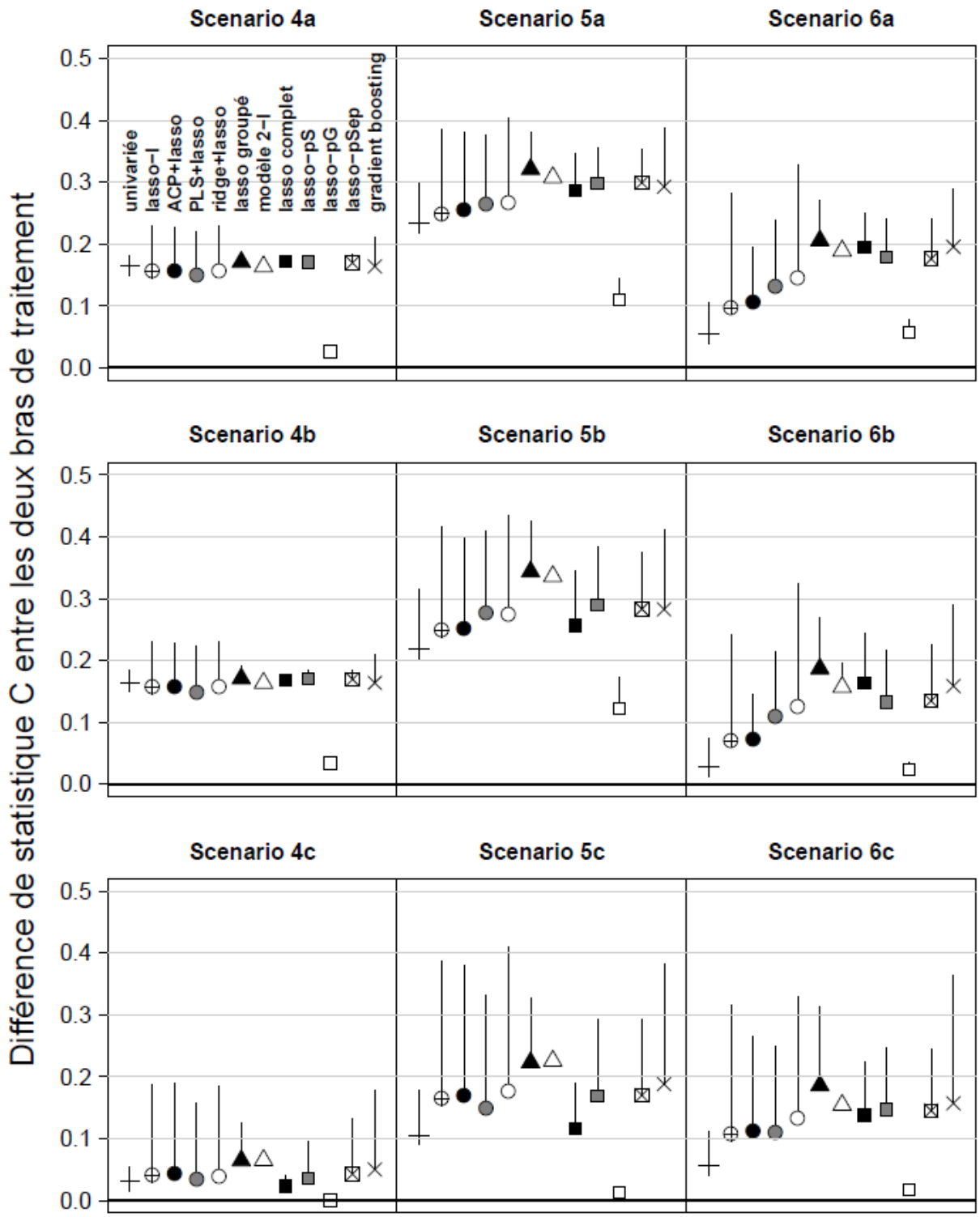


Figure 4.2 : Taux de faux négatifs (*FNR*) en fonction du taux de fausses découvertes (*FDR*) dans les scénarios alternatifs



Légende. Lignes verticales : réduction de la statistique ΔC entre le jeu d'apprentissage et le jeu de validation, symboles : statistique ΔC dans le jeu de validation. Quantités moyennes basées sur 250 réplications.

Figure 4.3 : Différence de statistique C entre les deux bras de traitement dans les scénarios alternatifs

Dans les scénarios alternatifs (Figure 4.2, Tableau 4.3) sans biomarqueurs pronostiques (scénarios 4–5a et 4–5b), les méthodes identifient la plupart des biomarqueurs prédictifs (*FNR* faible : 0,00–0,12 pour p = 500, 0,00–0,31 pour p = 1000) avec un nombre important de faux

positifs (*FDR* élevé : 0,76–0,88 pour p = 500, 0,69–0,88 pour p = 1000). Les signatures développées ont une bonne force d'interaction (Figure 4.3). En revanche, en présence de biomarqueurs pronostiques, ces méthodes fonctionnent moins bien car elles deviennent généralement beaucoup plus conservatrices (Tableau 4.3). En effet, sans biomarqueurs pronostiques (scénarios 5a et 5b) les méthodes sélectionnent entre 42 et 49 biomarqueurs pour p = 500 et entre 50 et 64 pour p = 1000, alors qu'en présence de biomarqueurs pronostiques (scénarios 6a et 6b), elles en sélectionnent seulement entre 15 et 37 pour p = 500 et entre 12 et 39 pour p = 1000. La diminution du nombre de biomarqueurs sélectionnés conduit également à la diminution du nombre de biomarqueurs prédictifs identifiés (*FNR* compris entre 0,30 et 0,60 pour p = 500 et entre 0,55 et 0,81 pour p = 1000). Dans les scénarios avec beaucoup de censure et des effets plus faibles des biomarqueurs, les méthodes sélectionnent globalement moins de biomarqueurs dans le scénario 5c et encore moins dans le scénario 6c. En présence d'effets propres (scénarios 6a et 6b), les méthodes qui tiennent compte du critère de jugement pour réduire la dimension des effets propres (*PLS*+lasso et ridge+lasso) fonctionnent mieux que celles qui n'en tiennent pas compte (lasso-I et ACP+lasso) pour l'identification des biomarqueurs prédictifs (*FNR* plus faible : 0,30–0,40 vs. 0,58–0,60 pour p = 500 et 0,55–0,64 vs. 0,80–0,81 pour p = 1000) et plus généralement pour la discrimination des biomarqueurs (*AUPRC* plus élevée : 0,38–0,43 vs. 0,25–0,37 pour p = 500 et 0,27–0,28 vs. 0,16–0,26 pour p = 1000). La solution extrême de ne pas considérer les effets propres (lasso-I) est la moins bonne solution dans ces scénarios. De plus, les mauvais résultats du lasso-I sont amplifiés lorsque les biomarqueurs prédictifs sont fortement corrélés avec les biomarqueurs pronostiques (Annexe A11), ce qui n'est pas le cas de l'ACP+lasso ou le *PLS*+lasso. Enfin, concernant la force d'interaction des signatures développées, il est intéressant de voir que ces quatre méthodes semblent être fortement impactées par le surapprentissage. En effet, la statistique ΔC estimée sur le jeu d'apprentissage est nettement supérieure à celle estimée sur le jeu de validation. Le lasso-I et le ridge+lasso sont les méthodes les plus impactées.

Un troisième panel de méthodes regroupe le **lasso groupé** et le **modèle 2-I**. Ces deux méthodes identifient globalement bien les biomarqueurs prédictifs mais réagissent assez mal en présence de biomarqueurs pronostiques notamment dans les scénarios nuls. En effet, dans les scénarios nuls 1 et 2 sans biomarqueurs pronostiques, le *FDR* est égal à 0,38–0,44 et 0,48–0,56 respectivement pour le modèle 2-I et le lasso groupé, alors qu'il est égal à 1 en présence de biomarqueurs pronostiques (scénarios 3, Tableau 4.2). Dans les scénarios alternatifs, les deux méthodes identifient bien les biomarqueurs prédictifs (scénarios 4–6a et 4–6b : *FNR* <

0,2, Figure 4.2) mais avec une quantité extrêmement importante de faux positifs (parfois supérieure à 100, Tableau 4.3). Malgré cela, l'*AUPRC* relativement élevée nous montre que le lasso groupé pourrait éliminer de nombreux faux positifs sans fortement augmenter le nombre de faux négatifs pour une augmentation modérée du paramètre de pénalisation λ. En raison de l'identification de la majorité des vrais biomarqueurs prédictifs, les signatures développées par le lasso groupé et par le modèle 2-I ont de bonnes forces d'interaction (Figure 4.3). De plus, le modèle 2-I est très faiblement impacté par le surapprentissage (peu de différence entre la statistique ΔC estimée sur l'échantillon d'apprentissage et de validation). Enfin, un point très important est que le lasso groupé est très fortement impacté par la présence de biomarqueurs pronostiques. En effet, le lasso groupé considère la grande majorité des biomarqueurs pronostiques comme étant des biomarqueurs prédictifs (scénario 6a : 10 FPp sur 10, scénario 6b : 20 FPp sur 20 et scénario 6c : 8 FPp sur 10, Tableau 4.3).

Les méthodes qui ne respectent pas obligatoirement la contrainte de hiérarchie forment un quatrième groupe : le **lasso complet** et les trois lassos adaptatifs (**lasso-pS**, **lasso-pG** et **lasso-pSep**). Dans les scénarios nuls 1 et 2, le lasso complet et les lassos adaptatifs (lasso-pS et lasso-pG) sélectionnent bien le modèle nul dans la majorité des cas, se traduisant par des *FDR* faibles (respectivement 0,01–0,02, 0,12–0,16 et 0,00, Tableau 4.2). Cependant, le premier est fortement affecté par la présence de biomarqueurs pronostiques (scénarios 3a et 3b, *FDR* : 0,88–0,98). Le lasso-pSep a un *FDR* modéré (0,36–0,42), puis large en présence de biomarqueurs pronostiques (0,55, $p = 1000$). Dans les scénarios alternatifs, le lasso complet et les deux lassos adaptatifs avec des poids spécifiques (lasso-pS et lasso-pSep) identifient la plupart des effets prédictifs (*FNR* faible : 0,00–0,25, Figure 4.2) avec un faible nombre de faux positifs (*FDR* relativement bas : 0,22–0,57, Figure 4.2) pour $p = 500$. Par conséquent, les signatures développées par ces méthodes ont une grande force d'interaction (Figure 4.3). Ces résultats sont légèrement moins bons lorsque le nombre de biomarqueurs candidats p augmente. En effet, pour $p = 1000$, le lasso-pS et le lasso-pSep sélectionnent trop peu de biomarqueurs et ratent de nombreux biomarqueurs prédictifs (en moyenne 14 et 13 biomarqueurs sélectionnés respectivement, $FNR = 0,61$). A noter également que ces deux méthodes ont de très mauvaises performances lorsque les biomarqueurs prédictifs sont fortement corrélés avec les biomarqueurs pronostiques (Annexe A11), avec aucune interaction sélectionnée dans 100% des cas. Le lasso-pG sélectionne le modèle nul dans la majorité des cas ($FDR \approx 0$, $FNR \approx 1$, faible puissance statistique et faible force d'interaction pour les signatures). Cependant, lorsque l'on s'intéresse à un critère plus global tel que l'*AUPRC*, le

lasso-pG affiche des résultats semblables aux trois autres méthodes (Tableau 4.3). Cela signifie donc que le paramètre de pénalisation choisi par cette méthode est trop important, la rendant trop stricte. Un choix plus réduit de ce paramètre permettrait à la méthode d'avoir des performances similaires aux trois autres méthodes. Enfin, bien que non obligatoirement maintenue par ces méthodes, la contrainte de hiérarchie est respectée pour plus d'une interaction sur deux dans les scénarios alternatifs (proportions d'interactions respectant cette contrainte : 0,54–0,68 pour le lasso complet, 0,52–0,62 pour le lasso-pS, 0,57–0,74 pour le lasso-pSep et 0,69–0,97 pour le lasso-pG). Comme souhaité, le poids utilisé pour le lasso-pSep permet de respecter plus souvent la contrainte de hiérarchie par rapport au lasso-pS de par sa nature.

Le ***gradient boosting*** ne se comporte comme aucun des groupes présentés ci-dessus. Dans les scénarios nuls, il affiche des résultats semblables au lasso groupé et au modèle 2-I en ayant un *FDR* élevé pour les scénarios 1a–1b et 2a–2b sans biomarqueurs pronostiques (0,66–0,69) puis égal à 1 en présence de biomarqueurs pronostiques (Tableau 4.2). Dans les scénarios alternatifs, le *gradient boosting* est semblable aux approches ne respectant pas obligatoirement la contrainte de hiérarchie : bonne sélection des biomarqueurs prédictifs avec relativement peu de faux positifs (Tableau 4.3, Figure 4.2). Cependant, le *gradient boosting* est moins conservateur que ces autres méthodes lorsque p augmente. En matière de force d'interaction des signatures développées, le *gradient boosting* a de bonnes performances bien qu'il soit fortement impacté par le surapprentissage, d'où l'importance d'avoir des données externes pour correctement évaluer la méthode.

Enfin, l'analyse de sensibilité générant les données à partir d'une distribution de Weibull montre que la variation du risque (décroissant ou croissant) au cours du temps n'impacte pas les performances relatives des méthodes (Annexe A12).

4.4 Application

Les douze approches ont été appliquées à des données réelles. Il s'agit d'une base de données de 614 patientes ayant un cancer du sein et pour lesquelles des données d'expression de gènes et des données cliniques telles que le statut nodulaire ou le stade de la tumeur sont disponibles (Desmedt et al., 2011 ; Hatzis et al., 2011). Ces patientes ont été traitées par chimiothérapie adjuvante à base d'anthracycline avec (n = 507) ou sans (n = 107) l'ajout de taxanes et ont respectivement une survie sans rechute à distance à 3 ans égale à 78% [IC95% : 74%–82%] et

79% [IC95% : 71%–87%]. L'objectif de cette application est d'identifier des biomarqueurs prédictifs de l'effet des taxanes.

Avant d'effectuer l'analyse, une étape de normalisation des données et de filtrage des variables a été effectuée. Comme précédemment, la normalisation des données a été réalisée par la technique *fRMA* pour rendre les puces comparables (Mccall et al., 2010) mais aussi en utilisant une technique pour normaliser les différentes plateformes (technique *XPN*, Shabalin et al., 2008). Les gènes filtrés ont été ceux ayant un écart interquartile inférieur à 1. Au final, $p = 1689$ gènes standardisés sur 22277 ont été retenus pour l'analyse.

Les 614 patientes ont été divisées aléatoirement en deux groupes : un groupe d'apprentissage pour construire le modèle ($n = 315$) et un groupe de validation pour évaluer les signatures développées par ces méthodes ($n = 299$). Les résultats (Tableau 4.4) montrent que le nombre de biomarqueurs prédictifs sélectionnés est grandement variable entre les méthodes : de 0 à 39.

Tableau 4.4 : Nombre de biomarqueurs prédictifs sélectionnés et force d'interaction des signatures dans l'application du cancer du sein

	Nombre de biomarqueurs prédictifs	statistique ΔC
univariée	4	0,10
lasso-I	21	0,09
ACP+lasso	13	0,12
***PLS*+lasso**	20	0,01
ridge+lasso	39	0,04
lasso groupé	4	0,06
modèle 2-I	34	0,12
lasso complet	0	0
lasso-pS	1	0,06
lasso-pG	0	0
lasso-pSep	2	0,14
gradient boosting	8	0,18

Les méthodes ne permettant pas de faire de la sélection de variables sur les effets propres des biomarqueurs (**lasso-I**, **ACP+lasso**, ***PLS*+lasso** et **ridge+lasso**) ont des résultats assez proches (14 gènes sélectionnés par au moins trois des quatre méthodes). A l'inverse, la majorité des gènes sélectionnés par le **lasso groupé** et le **modèle 2-I** sont très différents de ceux identifiés par les autres méthodes (3 sur 4 et 10 sur 34 respectivement). L'ensemble des gènes identifiés par l'**approche univariée** le sont également pour le ***gradient boosting***. Enfin, les méthodes ne respectant pas nécessairement la contrainte de hiérarchie semblent très conservatrices dans cette application. En effet, le **lasso complet** et le **lasso-pG** n'identifient aucun biomarqueur

prédictif, le **lasso-pSep** en identifie deux et le **lasso-pS** en identifie un seul : le gène IFIH1. A noter que ce gène est identifié par toutes les méthodes sélectionnant au moins un biomarqueur prédictif à l'exception du lasso groupé et de l'approche univariée. L'expression de ce gène est déjà connue dans la littérature comme étant associée à la récidive chez des patientes ayant un cancer du sein et non répondeurs aux taxanes (Magbanua et al., 2015). Ce gène est inclus dans deux brevets ayant pour objectif de prédire le bénéfice des taxanes (Gehrmann et Von Törne, 2009 ; Wang et al., 2013). Des études fonctionnelles suggèrent également que l'expression du gène IFIH1 est associée à la résistance aux taxanes chez des patients atteints d'un cancer de la prostate (Marín-Aguilera et al., 2011). Dans notre application, ce gène a une force d'interaction modérée avec une statistique ΔC égale à 0,06 dans l'échantillon de validation. En sélectionnant plus de biomarqueurs, les autres méthodes augmentent légèrement cette force d'interaction (Tableau 4.4). Le *gradient boosting* a la signature présentant la plus grande force d'interaction (8 gènes, statistique ΔC = 0,18).

4.5 Conclusion

Dans ce deuxième axe du travail de thèse, nous avons proposé des méthodes visant à identifier des biomarqueurs prédictifs de l'effet du traitement en vue d'établir une signature. Ces signatures ont pour objectif de sélectionner les patients susceptibles de bénéficier du traitement et ne pas traiter ceux pour qui le traitement n'aurait pas d'effet ou un effet délétère (Buyse et Michiels, 2013 ; Hingorani et al., 2013). Bien que l'identification de biomarqueurs prédictifs soit nécessaire (Michiels et al., 2011), aucune recommandation n'est encore établie quant à la technique à utiliser pour les sélectionner dans un contexte de données de grande dimension. Comme dit précédemment, la sélection de biomarqueurs peut avoir plusieurs objectifs : sélectionner les biomarqueurs qui ont un rôle biologique pour comprendre des mécanismes sous-jacents et identifier des gènes cibles, ou encore sélectionner les patients qui sont le plus susceptibles de bénéficier de la thérapie. Dans le premier cas, il est important que les biomarqueurs sélectionnés soient effectivement prédictifs, quitte à ne pas tous les identifier ; il faut donc contrôler le *FDR*. Cependant, si l'intérêt principal est de sélectionner une sous-population de patients à traiter, il est important que le plus grand nombre de biomarqueurs prédictifs soit sélectionné quitte à inclure des biomarqueurs n'apportant que du bruit, afin que le plus d'information possible soit disponible pour classer les patients ; dans ce cas, il est important de contrôler le *FNR*. Dans mon deuxième axe de thèse, l'intérêt s'est porté sur le premier objectif en proposant douze approches pouvant être utilisées pour

identifier des biomarqueurs prédictifs tout en limitant la sélection de faux positifs. Naturellement, il n'est pas possible d'établir une recommandation générale sur la meilleure approche à utiliser car les méthodes discutées ne sont qu'un sous-ensemble des approches possibles. Ainsi, nous avons présenté les avantages et inconvénients de chaque approche. Nous avons également proposé un nouveau critère d'évaluation, lié à la sélection correcte des biomarqueurs, qui consiste à regarder la force d'interaction des signatures développées par les méthodes.

Sur la base des résultats de l'étude de simulation, plusieurs groupes de méthodes ont été identifiés. Tout d'abord, l'approche simple visant à identifier les interactions à partir de p modèles univariés doit être écartée lorsque les biomarqueurs sont corrélés entre eux, même avec une procédure de contrôle du *FDR*. Si l'on souhaite choisir une méthode démarrant du modèle nul, le *gradient boosting* peut être une meilleure option bien qu'il ne contrôle pas du tout le *FDR* dans les scénarios nuls. Les méthodes ne permettant pas de faire de sélection sur les effets propres dans le modèle complet (i.e. non inclus dans le modèle : lasso-I, réduction de dimension : ACP+lasso ou *PLS*+lasso, ou soumis à la pénalisation ridge ne réalisant pas de sélection : ridge+lasso) ont des profils semblables et fonctionnent bien à l'exception des situations contenant également des biomarqueurs pronostiques. Les mauvais résultats dans ces situations sont amplifiés pour les méthodes ne tenant pas compte du critère de jugement dans la transformation de la matrice des effets propres (lasso-I ou ACP+lasso). De plus, bien que cela ne soit pas l'objectif principal de cette étude, ces méthodes ne permettent pas d'identifier des biomarqueurs pronostiques ce qui est souvent souhaité par les médecins pour qu'ils puissent facilement et de manière fiable les reproduire sur d'autres plateformes (e.g. RT-PCR, technique qui associe une transcription inverse de l'ARN pour synthétiser l'ADN, suivie d'une méthode d'amplification génétique *in vitro* visant à multiplier spécifiquement le segment d'ADN d'intérêt pour repérer un gène particulier dans un génome entier). A l'inverse, les méthodes permettant de faire de la sélection sur les effets propres fonctionnent globalement bien indépendamment du fait de tenir compte ou non de la contrainte de hiérarchie. En effet, le non-respect de cette contrainte (lasso complet et lassos adaptatifs) ne semble pas impacter la performance des méthodes en matière de sélection, bien qu'elles aient tendance à devenir trop conservatrices avec l'augmentation du nombre de biomarqueurs p. La performance du lasso groupé est très impactée par la présence d'effets propres du fait que les groupes considérés ne permettent pas de séparer l'effet propre et l'interaction de chaque biomarqueur. Pour pallier cela, différents tests *a posteriori* peuvent être mis en place tels que

tester la significativité des interactions sélectionnées (Lockhart et al., 2014) ou encore tester la significativité pratique des interactions à partir d'un seuil préalablement établi. Cependant, dans tous les cas, des choix arbitraires doivent être effectués. Une autre idée serait d'implémenter une version adaptative du lasso groupé qui consisterait à ajouter des poids à chaque groupe sur la base d'une estimation préalable des effets des interactions. Nous n'avons pas considéré cette approche dans ce travail. Le modèle estimant les effets pronostiques de chaque bras de traitement fonctionne également bien pour identifier les biomarqueurs prédictifs. Bien entendu, il serait utile d'appliquer un test de contraste pour évaluer la différence entre les deux effets pronostiques, mais comme pour le lasso groupé cela nécessite de faire un test *a posteriori* qui requiert des choix supplémentaires. De plus, cela n'est pas très fiable dans ce modèle car les coefficients de régression sont pénalisés. Enfin, dans cette étude nous avons souhaité rappeler deux points : (i) l'importance d'avoir un jeu de données externe pour évaluer correctement la force d'interaction des signatures du fait que la plupart des modèles sont impactés par le surapprentissage, tout particulièrement le lasso-I et les approches en deux étapes (e.g. traitement fixé en *offset* ou encore réduction de dimension de la matrice des effets propres) ; (ii) la nécessité d'avoir un nombre important d'évènements pour identifier des interactions quelle que soit la méthode utilisée.

Pour la quasi-totalité des méthodes discutées, la liste des biomarqueurs est définie par l'estimation d'un paramètre de pénalisation par le critère *cvl* (Verweij et van Houwelingen, 1993, 1994). Cependant, nous avons observé que cette technique peut être sous-optimale pour certaines approches : e.g. le paramètre de pénalisation est trop important pour le lasso-pG le rendant trop conservateur et il est trop faible pour le lasso groupé conduisant à la sélection d'un grand nombre de faux positifs. Pour ce dernier cas, il est envisageable d'utiliser l'extension de la pénalisation lasso que nous avons proposée dans ma première partie de thèse pour réduire le nombre de faux positifs dans un contexte pronostique (Chapitre 3, Ternès et al., 2016a).

Un autre point important à discuter concerne la standardisation des interactions au regard des effets propres. Dans ce travail, le traitement est codé en $\pm 0{,}5$ signifiant que la variabilité des interactions $\boldsymbol{X_j T}$ (variance = 0,25) est moins importante que celle des effets propres $\boldsymbol{X_j}$ (variance = 1). Dans une analyse de sensibilité, nous avons recodé le traitement en ± 1 pour obtenir la même variabilité entre les effets propres des biomarqueurs et leur interaction avec le traitement. Les résultats montrent de moins bonnes capacités de sélection des interactions pour la plupart des méthodes avec ce codage différent (résultats non montrés).

L'illustration des méthodes sur des données réelles chez des patientes atteintes d'un cancer du sein montre une forte disparité quant au nombre de biomarqueurs sélectionnés. Comme pour toute application, les résultats sont uniquement illustratifs et non comparatifs étant donné qu'il n'est pas possible de savoir quels gènes sont réellement prédictifs de l'effet du traitement. De plus, dans cette illustration les données ne sont pas collectées dans un essai clinique randomisé. Cependant, il est intéressant de noter que le gène identifié par la majorité des méthodes (IFIH1) est déjà reporté dans la littérature comme étant prédictif de la réponse aux taxanes, et que certains groupes de méthodes mis en évidence dans l'étude de simulation sont également présents dans cette application. Par exemple, on retrouve une forte similitude des méthodes ne permettant pas de faire de la sélection de variables sur les effets propres et une forte similitude des méthodes ne respectant pas la contrainte de hiérarchie, avec des listes de biomarqueurs très conservatrices lorsque p est grand. Il est aussi important de noter que pour certaines méthodes la liste des biomarqueurs sélectionnés peut varier en fonction de l'assignation des observations aux différents sous-groupes de la validation croisée (résultats non montrés). Enfin, nous avons tenu compte des variables cliniques (statut nodulaire et stade de la tumeur) en estimant leur effet et en les fixant en *offset* dans les modèles de sélection des biomarqueurs. Cependant, l'intégration des données cliniques et génomiques est une question importante qui dépasse ce travail et nous en discuterons plus en détails dans la conclusion de ce manuscrit (Chapitre 6).

En conclusion de ce travail, nous avons proposé et évalué un grand panel d'approches possibles pour identifier des biomarqueurs prédictifs. Cette étude permet de mettre en exergue les avantages et inconvénients de chacune des approches considérées. Nous proposons également un nouveau critère pour évaluer ces signatures prédictives.

Chapitre 5 Prédiction individuelle de l'effet du traitement

Jusqu'à présent, le travail de thèse s'est focalisé sur la bonne sélection de biomarqueurs pronostiques ou prédictifs jouant un vrai rôle biologique afin d'établir des signatures. Dans cette troisième partie de thèse, je me suis intéressé à différentes stratégies permettant d'obtenir une prédiction individuelle de la probabilité de survie pour un patient futur à l'aide d'une signature génomique. Il est important d'un point de vue pratique d'implémenter un outil de prédiction à partir des données des patients utilisées pour le développement de la signature (échantillon d'apprentissage) qui soit représentatif d'une population externe n'ayant pas été utilisée pour la mise en place de cet outil (échantillon de validation). D'un point de vue statistique, nous avons alors proposé une méthodologie permettant (i) d'éviter un éventuel surapprentissage du modèle dans l'étude utilisée pour le développement de la signature et (ii) de calculer une mesure d'incertitude autour de cette prédiction, telle que l'intervalle de confiance, à partir d'un modèle de régression pénalisé.

5.1 État des connaissances

Pour effectuer une bonne prédiction avec une bonne mesure d'incertitude, il est nécessaire d'identifier un modèle représentatif de la réalité biologique pour lequel les coefficients de régression et leur variabilité sont correctement estimés. Lorsque l'on estime les coefficients de régression à partir d'un modèle pénalisé, les prédictions individuelles de la probabilité de

survie et leur intervalle de confiance ne sont pas simples à mettre en place. En effet, les modèles de régression pénalisés introduisent un biais dans l'estimation des coefficients. Par conséquent, l'estimation de leur variabilité a un intérêt limité (Goeman, Meijer et Chaturvedi, 2016). De plus, Knight et Fu (2000) ont montré que le lasso a des résultats peu concluants en matière de sélection de variables et les coefficients de régression estimés n'ont pas une distribution asymptotique habituelle, synonyme de non-respect de la propriété *oracle*. Des solutions alternatives ont alors été proposées pour respecter cette propriété *oracle*, telles que le lasso adaptatif, le scad ou encore l'*elastic net* adaptatif (Fan et Li, 2002 ; Wu, 2012 ; Zou, 2006), permettant une estimation juste de la variabilité des coefficients de régression lorsque ceux-ci sont non nuls. Très récemment, certains auteurs tels que Sinnott et Cai (2016) ou Lin et Halabi (2016) ont montré que les techniques existantes sont peu efficaces pour estimer la variabilité des coefficients, notamment lorsque leur estimation est nulle. En effet, dans cette situation l'approche analytique tend à sous-estimer la variabilité des coefficients et les approches de rééchantillonage comme le bootstrap tendent à la surestimer. Dans ces deux articles les auteurs proposent des solutions alternatives. Sinnott et Cai (2016) suggèrent une approche en deux étapes considérée comme un compromis entre la propriété *oracle* et le rééchantillonage. Elle consiste à générer des échantillons bootstrapés et à (i) implémenter le modèle pénalisé sur l'échantillon original et sur les échantillons bootstrapés pour définir les coefficients non nuls et (ii) réestimer les modèles sur les mêmes jeux de données avec seulement les variables retenues à l'issue de l'étape 1. Les auteurs ont également étendu leur approche pour prédire la fonction de survie et son intervalle de confiance pour un patient futur. Lin et Halabi (2016) suggèrent une approche basée sur des perturbations qui consiste à ajouter un poids aléatoire à chaque contribution à la vraisemblance en s'inspirant des travaux précédents de Chatterjee et Lahiri (2011) et Minnier, Tian et Cai (2012). Cependant, dans ces deux articles récents, les auteurs n'ont évalué ces approches que dans une situation de données de petite dimension. En effet, Sinnott et Cai (2016) ont effectué une étude de simulation contenant seulement p = 10, 20 ou 30 biomarqueurs dont 5 ayant un effet pronostique. En plus de scénarios avec p = 10 ou 20 biomarqueurs, Lin et Halabi (2016) ont rapidement étudié une situation avec p = 1000 biomarqueurs montrant des résultats peu concluants en matière de sélection. A noter également que ces deux articles se focalisent uniquement sur le cas de biomarqueurs pronostiques. Matsui et al. (2012) avaient précédemment proposé une approche pour développer et valider une signature génomique dans un essai clinique randomisé. Contrairement aux autres papiers discutés, ces auteurs se sont intéressés à la fois à la composante pronostique permettant d'estimer la survie spécifique

de chaque patient et à la composante prédictive permettant d'estimer les interactions entre le traitement et les biomarqueurs. Dans leur travail, les deux composantes sont estimées indépendamment l'une de l'autre et la sélection de variables est réalisée au travers d'une approche univariée similaire à celle présentée en section 4.2.5. Dans le modèle final, les auteurs ont utilisé des polynômes fractionnaires pour modéliser les effets des variables, comme suggéré par Royston et Sauerbrei (2004). Un autre point intéressant discuté par Matsui et al. (2012) concerne la validation de la signature. En l'absence d'un jeu de validation, les auteurs suggèrent d'utiliser la technique de la validation croisée pour imiter une validation externe des résultats. Malheureusement, ce développement a été appliqué uniquement sur un jeu de données d'exemple et aucune étude de simulation n'a été réalisée pour évaluer ses performances. L'utilisation d'une technique de rééchantillonage pour valider les résultats est une recommandation déjà faite par Michiels et al. (2005) ou encore Simon et al. (2011) qui ont montré différents exemples pour lesquels l'absence de validation des résultats conduit à des conclusions trop optimistes. Enfin, pour visualiser les interactions traitement-biomarqueur, différentes approches graphiques ont été proposées, comme par exemple : Bonetti et Gelber, 2015 ; Huang, Pepe et Feng, 2007 ; Lazar et al., 2010 ; Pepe et al., 2008 ; Yang et al., 2015.

5.2 Approches étudiées

Dans ce travail et à la lecture des articles présentés ci-dessus, nous avons souhaité évaluer l'impact de différentes méthodes et paramètres sur la prédiction individuelle de l'effet du traitement, notamment dans un contexte de données de grande dimension. Les principaux paramètres considérés étaient le choix du modèle de sélection, la méthode d'estimation ponctuelle de la prédiction de la probabilité de survie et l'estimation des bornes de confiance de ces prédictions.

5.2.1 Choix du modèle de sélection

Pour établir une signature pronostique et une signature prédictive, nous proposons d'utiliser le modèle (4.1) contenant les effets propres du traitement et des biomarqueurs ainsi que leur interaction comme présenté au Chapitre 4. Afin d'effectuer une sélection de variables et au vu de sa dimension ($2p + 1 \gg n$), nous avons soumis le modèle à deux pénalisations possibles $p(\lambda, \boldsymbol{\beta}, \boldsymbol{\gamma})$: soit le lasso (plus simple du point de vue computationnel, mais ne garantissant pas la propriété *oracle*), soit le lasso adaptatif (garantissant la propriété *oracle*,

mais plus demandeur en ressources de calcul) comme présenté en section 4.2.1. Pour rappel, le modèle pénalisé s'écrit :

$$h_{\mathrm{P}}(t, \boldsymbol{T}, \boldsymbol{X}) = h_0(t) \exp\left(\alpha \boldsymbol{T} + \sum_{j=1}^{p} \beta_j \boldsymbol{X_j} + \sum_{j=1}^{p} \gamma_j \boldsymbol{X_j T} \right),$$

avec $\{\hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\} = \mathrm{argmax}_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma}}\{l(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}) - p(\lambda, \boldsymbol{\beta}, \boldsymbol{\gamma})\}$. Pour chaque pénalisation, le paramètre λ est estimé à partir d'une validation croisée en K_1 sous-échantillons (K_1 = 5 dans ce travail). A titre de comparaison, nous avons également mis en place un modèle réestimant sans pénalisation les coefficients des biomarqueurs retenus dans le modèle pénalisé. Ce modèle non pénalisé s'écrit :

$$h_{\mathrm{nP}}(t, \boldsymbol{T}, \boldsymbol{X}) = h_0(t) \exp\left(\alpha \boldsymbol{T} + \sum_{j \in Q_{Po}} \beta_j \boldsymbol{X_j} + \sum_{j \in Q_{Pe}} \gamma_j \boldsymbol{X_j T} \right)$$

avec $\{\hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\} = \mathrm{argmax}_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma}}\{l(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma})\}$ et respectivement Q_{Po} et Q_{Pe} les indices des biomarqueurs pronostiques et prédictifs sélectionnés dans le modèle pénalisé.

5.2.2 Estimation ponctuelle de la probabilité de survie

A partir du modèle choisi $h(t, \boldsymbol{T}, \boldsymbol{X})$ (i.e. $h_{\mathrm{P}}(t, \boldsymbol{T}, \boldsymbol{X})$ ou $h_{\mathrm{nP}}(t, \boldsymbol{T}, \boldsymbol{X})$), il est possible d'estimer la probabilité de survie d'un nouvel individu *k* à un instant *t* que l'on note

$$\hat{S}_k(t) = \hat{S}(t, T_k, \boldsymbol{X_k}) = \exp\left(-\hat{H}_k(t)\right)$$

avec $\hat{H}_k(t)$ le risque cumulé de l'individu *k* à l'instant *t* noté

$$\hat{H}_k(t) = \hat{H}_0(t) \times \exp\left(\hat{\alpha} T_k + \sum_{j=1}^{p} \hat{\beta}_j X_{kj} + \sum_{j=1}^{p} \hat{\gamma}_j X_{kj} T_k \right),$$

et $\hat{H}_0(t)$ l'estimation du risque de base cumulé jusqu'au temps *t*. On estime classiquement cette quantité à partir de l'estimateur non paramétrique de Breslow (1972), une version pondérée de celui de Nelson et Aalen (Nelson, 1969), que l'on note :

$$\hat{H}_0(t) = \sum_{i\,:\,t_i \leq t} \frac{d(t_i)}{Y(t_i) \exp\left(\hat{\alpha} T_i + \sum_{j=1}^{p} \hat{\beta}_j X_{ij} + \sum_{j=1}^{p} \hat{\gamma}_j X_{ij} T_i\right)}$$

avec $Y(t_i)$ le nombre de sujets encore à risque juste avant t_i et $d(t_i)$ le nombre d'évènements à l'instant t_i. Malgré le fait que la validation externe reste fondamentale, une première validation interne est souvent effectuée sur le jeu de données utilisé pour le développement de la signature. Pour estimer la probabilité de survie de chaque patient de l'échantillon d'apprentissage, nous avons évalué deux stratégies. La première consiste à utiliser l'ensemble des patients pour estimer le modèle $h(t, \boldsymbol{T}, \boldsymbol{X})$ à partir d'une validation croisée en K_1 sous-échantillons pour estimer le paramètre λ, et à utiliser ce modèle pour estimer la probabilité de survie de ces mêmes patients. On appelle cette stratégie *simple validation croisée* ou 1CV. La deuxième stratégie cherche à imiter les résultats que l'on obtiendrait sur un nouveau jeu de données (i.e. validation externe). Cette stratégie utilise le principe de validation croisée comme suggéré par Simon et al. (2011) puis illustré par Matsui et al. (2012). Cela consiste à diviser l'échantillon en K_2 sous-échantillons (K_2 = 5 dans ce travail) et à estimer la probabilité de survie des patients d'un sous-groupe à partir du modèle $h(t, \boldsymbol{T}, \boldsymbol{X})$ estimé avec les données des autres sous-groupes au travers d'une validation croisée de K_1 sous-échantillons pour estimer λ. Ce processus est répété de façon à ce que l'on puisse estimer la probabilité de survie de l'ensemble des patients. On appelle cette stratégie *double validation croisée* ou 2CV.

5.2.3 Estimation des bornes de confiance

Nous avons également évalué deux approches pour construire les bornes de l'intervalle de confiance de ces prédictions : une approche semi-paramétrique et une approche non paramétrique.

L'approche semi-paramétrique présentée par Therneau et Grambsch (2000) consiste à estimer la variance du risque cumulé $\widehat{H}_k(t)$ à partir de l'estimateur de Breslow (Breslow, 1974) dans le modèle semi-paramétrique de Cox. L'intervalle de confiance à $1 - \theta$ de $\hat{S}_k(t)$ est donc :

$$\mathrm{IC}_{1-\theta}\left(\hat{S}_k(t)\right) = \exp\left(-\widehat{H}_k(t) \pm z_{1-\frac{\theta}{2}}\sqrt{\widehat{\mathrm{var}}\left(\widehat{H}_k(t)\right)}\right).$$

L'implémentation de cette approche est directement réalisable avec la fonction `survfit()` du package `survival` de `R` (Therneau et Lumley, 2016) en faisant appel à un modèle de Cox estimé via la fonction `coxph()`. Dans notre cas, le modèle pénalisé estimé à partir du package `glmnet` ne fournit des informations ni sur la variabilité des coefficients ni sur le risque de base du modèle. Une astuce consiste donc à utiliser la fonction `coxph()` en

initialisant les coefficients aux valeurs estimées dans le modèle pénalisé et à fixer à zéro le nombre d'itérations de l'algorithme de maximisation de la vraisemblance partielle, permettant ainsi d'estimer les erreurs types aux valeurs fixées des paramètres à partir de la matrice hessienne et d'obtenir l'intervalle de confiance semi-paramétrique de la survie prédite.

L'approche non paramétrique consiste quant à elle à générer *B* échantillons tirés au sort avec remise à partir de l'échantillon initial (i.e. *bootstrap*) et à estimer le modèle $h(t, \boldsymbol{T}, \boldsymbol{X})$ (voir section 5.2.1) au sein de chaque échantillon. On a donc *B* modèles $h_1(t, \boldsymbol{T}, \boldsymbol{X})$, ..., $h_B(t, \boldsymbol{T}, \boldsymbol{X})$ permettant d'estimer *B* probabilités de survie d'un individu *k* au temps *t* (section 5.2.2) notées $\hat{S}_{k(\text{boot})}(t) = \{\hat{S}_{k(1)}(t), \ldots, \hat{S}_{k(B)}(t)\}$. Un intervalle de confiance à $1 - \theta$ de $\hat{S}_k(t)$ peut être construit à partir des percentiles de la distribution de $\hat{S}_{k(\text{boot})}(t)$ tel que :

$$\text{IC}_{1-\theta}\left(\hat{S}_k(t)\right) = \left[\text{q}_{\frac{\theta}{2}}\left(\hat{S}_{k(\text{boot})}(t)\right) ; \text{q}_{1-\frac{\theta}{2}}\left(\hat{S}_{k(\text{boot})}(t)\right)\right].$$

5.2.4 Représentation graphique

Graphiquement nous avons souhaité représenter la probabilité de survie des patients à un instant *t* en fonction de leur score prédictif comme illustré en Figure 5.1. Pour rappel, le score d'interaction d'un patient *k* noté $\hat{\eta}_k$ s'écrit :

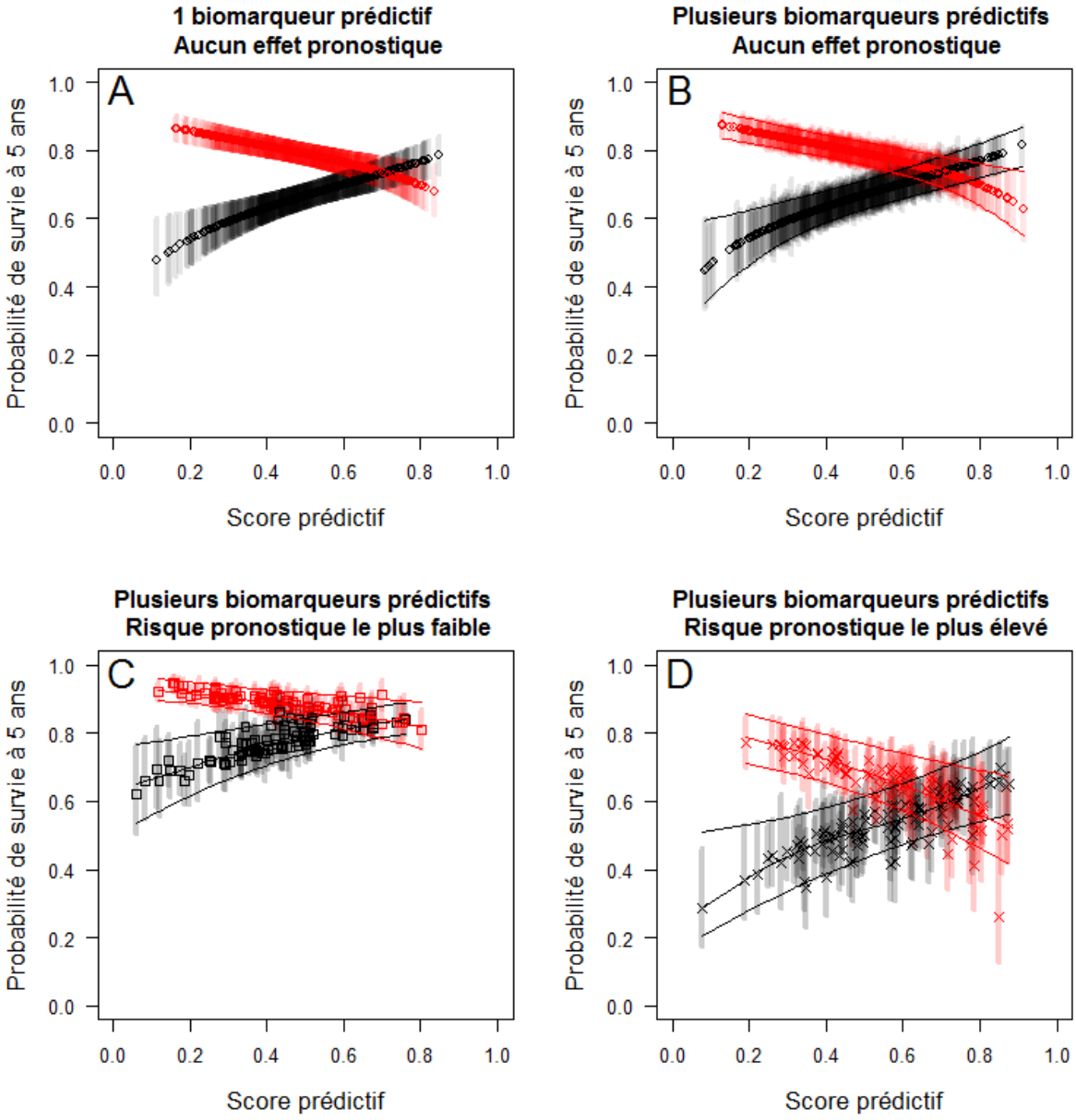
$$\hat{\eta}_k = \sum_{j=1}^{p} \hat{\gamma}_j X_{kj}.$$

Pour une facilité de lecture, le score $\widehat{\boldsymbol{\eta}}$ est normalisé entre 0 et 1 en utilisant la fonction de répartition d'une loi normale centrée réduite $\Phi : \mathbb{R} \longrightarrow (0, 1)$.

Pour faciliter la présentation des méthodes, nous considérons dans un premier temps le cas simple où aucun biomarqueur pronostique n'est sélectionné dans le modèle ($\hat{\beta}_j = 0, \forall j$).

Dans le cas où un seul biomarqueur prédictif est identifié, la relation entre le score d'interaction $\widehat{\boldsymbol{\eta}}$ et la probabilité de survie $\hat{S}(t)$ est strictement monotone, ainsi que les intervalles de confiance (Figure 5.1a). En revanche, si plusieurs biomarqueurs prédictifs sont sélectionnés, il est possible que pour deux patients ayant des scores d'interaction $\widehat{\boldsymbol{\eta}}$ (et donc des probabilités de survie prédites) semblables ou même identiques, les intervalles de confiance des probabilités de survie prédites $\hat{S}(t)$ soient différents (Figure 5.1b) en raison des différentes combinaisons linéaires possibles $\widehat{\boldsymbol{\gamma}}' \boldsymbol{X}_k$ pour obtenir un $\widehat{\boldsymbol{\eta}}$ identique à partir des

mêmes coefficients $\hat{\boldsymbol{\gamma}}$. Pour pallier cela, nous avons proposé de lisser la relation entre le score $\hat{\boldsymbol{\eta}}$ et chacune des deux bornes des intervalles de confiance par des *splines* dans l'optique de la rendre univoque et possiblement monotone (Figure 5.1b). Nous avons utilisé le package `cobs` sous `R` pour estimer des splines $spl(t)$ en y imposant une contrainte $0 \leq spl(t) \leq 1$.



Légende. Probabilité de survie prédite en fonction du score prédictif $\hat{\eta}$ dans le cas de : aucun biomarqueur identifié comme pronostique et un seul (A) ou plusieurs (B) comme prédictifs ; plusieurs pronostiques et plusieurs prédictifs dans le groupe à meilleur (C) et moins bon (D) pronostic. Noir : bras contrôle, rouge : bras expérimental. Bandes verticales : intervalles de confiance à 95%. Courbes : prédiction et limites des intervalles lissées par des splines. Jeu de données simulé provenant du scénario 6 de l'étude de simulation présentée en section 5.3.1.

Figure 5.1 : Illustration graphique de la probabilité de survie en fonction du score prédictif

Le nombre de nœuds des splines a été estimé par le critère d'*AIC* (méthode par défaut du package). Dans ce travail, le critère choisi pour estimer le nombre de nœuds (*AIC*, *BIC* ou

critère d'information de Schwartz : *SIC*) n'a pas d'impact sur l'estimation des splines (résultats non montrés). Un problème restant concerne la prise en compte des variables pronostiques. En effet, hormis le fait que deux patients peuvent avoir des intervalles de confiance de la probabilité de survie $\hat{S}(t)$ différents bien qu'ayant des scores d'interaction $\widehat{\boldsymbol{\eta}}$ identiques, il est aussi possible que la prédiction même de la survie $\hat{S}(t)$ soit différente pour la même valeur de $\widehat{\boldsymbol{\eta}}$ à cause d'un score pronostique $\widehat{\boldsymbol{\varphi}}$ différent entre les patients (Figure 5.1c et Figure 5.1d). Pour un patient donné, ce dernier est égal à :

$$\hat{\varphi}_k = \sum_{j=1}^{p} \hat{\beta}_j X_{kj}.$$

Comme précédemment, le score $\widehat{\boldsymbol{\varphi}}$ est normalisé entre 0 et 1 en utilisant la fonction de répartition d'une loi normale centrée réduite (i.e. fonction `pnorm()` sous `R`) pour une facilité de lecture. En outre, à l'intérieur de chacun de ces groupes une hétérogénéité du niveau pronostique $\widehat{\boldsymbol{\varphi}}$ persiste et une *spline* est encore utilisée pour lisser les estimations ponctuelles en plus des bornes des intervalles de confiance. Cela permet d'avoir une seule prédiction (par bras) de la probabilité de survie pour chaque valeur du score prédictif. A titre d'exemple, considérons le modèle avec $h_0(t) = 1, \alpha = 0, \varphi_i = X_{i1}$, et $\eta_i = X_{i1} + X_{i2}$. Deux patients A et B ayant $X_{A1} = X_{A2} = 1$ et $X_{B1} = 2, X_{B2} = 0$ ont les deux $\eta_A = \eta_B = 2$, alors que $\hat{S}_A(t) = \exp(-t\,exp(3)) \neq \hat{S}_B(t) = \exp(-t\,exp(4))$. Ainsi, nous avons proposé de distinguer différents sous-groupes de patients en fonction de leur score $\widehat{\boldsymbol{\varphi}}$. Après normalisation du score pour que celui-ci soit compris entre 0 et 1, nous avons choisi de séparer les patients en quatre groupes selon la catégorisation proposée par Cox (1957). Celle-ci consiste à utiliser les quantiles du score afin d'obtenir la distribution suivante : 16,4%, 33,6%, 33,6% et 16,4%. Les Figure 5.1c et Figure 5.1d correspondent respectivement aux groupes ayant le meilleur et le moins bon pronostic selon le score défini. Dans ce travail, nous avons donc évalué deux approches : l'approche par *spline* avec catégorisation en fonction du risque pronostique et l'approche par *point*. A noter, pour cette dernière approche, la *double validation croisée* peut être utilisée uniquement pour la validation interne des résultats et non pour effectuer une prédiction pour des patients futurs. En effet, un inconvénient possible de la *double validation croisée* est qu'elle repose sur K_2 modèles et donc K_2 risques de base plutôt qu'un seul.

5.3 Étude de simulation

Pour évaluer les approches discutées dans la section 5.2, une étude de simulation a été mise en place. Dans cette étude, nous avons étudié principalement leur capacité de prédiction de la probabilité de survie à un instant *t* donné pour un nouveau patient, ainsi que l'estimation de l'incertitude autour de ces prédictions.

5.3.1 Choix des scénarios

Pour cette étude de simulation, des scénarios semblables à ceux présentés dans la section 4.3.2 ont été générés. Il s'agit d'une étude comprenant six scénarios – trois scénarios nuls et trois scénarios alternatifs (Tableau 5.1) – dans laquelle les caractéristiques des jeux de données sont proches de celles de l'application (section 5.4).

Tableau 5.1 : Scénarios de l'étude de simulation

	Tailles d'effets			**Taux de censure**	
Scénarios	α	β_j	γ_j	T^-	T^+
(**1**) Aucun effet	0	0	0	0,73	0,73
(**2**) Effet du traitement seul	$-0{,}8$	0	0	0,63	0,81
(**3**) 20 bm. pronostiques	0	$\sim U(-0{,}20, -0{,}05)$	0	0,70	0,70
(**4**) 15 bm. prédictifs	0	0	$\sim U(-0{,}40, -0{,}10)$	0,70	0,71
(**5**) Effet du traitement + (4)	$-0{,}8$	0	$\sim U(-0{,}40, -0{,}10)$	0,61	0,79
(**6**) 20 bm. pronostiques + (5)	$-0{,}8$	$\sim U(-0{,}20, -0{,}05)$	$\sim U(-0{,}40, -0{,}10)$	0,60	0,76

Légende. α : effet du traitement, β_j : effet des biomarqueurs pronostiques, γ_j : effet des biomarqueurs prédictifs, T^+ : bras expérimental, T^- : bras contrôle, bm : biomarqueurs.

Le premier scénario nul ne contient aucun signal (scénario nul complet), le deuxième contient un fort effet du traitement ($\alpha = -0{,}8$) et le troisième contient 20 biomarqueurs pronostiques (avec $\beta_j \sim U(-0{,}20, -0{,}05)$). Concernant les trois scénarios alternatifs (i.e. effets des biomarqueurs interagissant avec le traitement), le scénario 4 contient 15 biomarqueurs prédictifs (avec $\gamma_j \sim U(-0{,}40, -0{,}10)$), le scénario 5 ajoute un fort effet du traitement au scénario 4 ($\alpha = -0{,}8$) et le scénario 6 ajoute 20 biomarqueurs pronostiques (avec $\beta_j \sim U(-0{,}20, -0{,}05)$) au scénario 5. Ces six scénarios ont été générés pour n = 1500 patients et p = 500 biomarqueurs. Les temps de survie ont été générés selon une distribution exponentielle, et la médiane de survie m_0 a été choisie telle que la probabilité de survie moyenne à l'horizon de prédiction τ, notée $S(\tau)$, soit proche de 80%. Dans ce travail, l'horizon de prédiction τ a été fixé à 5 ans. Pour ce critère, une analyse de sensibilité a été

réalisée pour $S(\tau) \approx 50\%$. Des temps de censure ont été générés selon une distribution uniforme $U(0\,,u)$ avec u choisi tel que le taux de censure soit compris entre 60% et 80%. Enfin, une structure de corrélation autorégressive a été générée par blocs de 25 biomarqueurs avec $\rho_{jj'} = 0{,}8^{|j-j'|}$. Pour chaque scénario, 250 jeux de données d'apprentissage et 250 jeux de données de validation ayant les mêmes caractéristiques que les jeux d'apprentissage ont été simulés.

5.3.2 Critères d'évaluation

Dans ce travail, l'objectif principal est d'évaluer la capacité de prédiction des approches. Tous les critères d'évaluation présentés ci-dessous ont été mesurés à l'horizon de prédiction τ égal à 5 ans. Pour mesurer la capacité de prédiction globale des modèles, nous avons utilisé le score de Brier (Brier, 1950). Ce score, dépendant du temps, correspond à un score quadratique mesurant les écarts entre probabilités attendues et prédites. On note ce score :

$$\text{Brier}(\tau) = \frac{1}{n}\sum_{k=1}^{n}\left[\frac{\left(\hat{S}_k(t)\right)^2 \mathbf{I}(t_k \leq \tau, \delta_k = 1)}{\hat{S}_k^C(t)} + \frac{\left(1-\hat{S}_k(t)\right)^2 \mathbf{I}(t_k > \tau)}{\hat{S}^C(\tau)}\right].$$

A noter, le score de Brier tient compte de la censure au travers de $\hat{S}^C(\cdot)$ la distribution de la censure estimée par Kaplan-Meier. Afin de ne pas garder d'horizon fixe, une solution est de calculer l'intégrale de ce score pour des temps $t \in [0, \tau]$ en utilisant une fonction de poids $W(t)$ pour laquelle un choix classiquement utilisé est $\left(1-\hat{S}(t)\right) / \left(1-\hat{S}(\tau)\right)$ (Graf et al., 1999). Ce score intégré, noté *iBrier*, doit être minimisé.

Pour évaluer la force pronostique et prédictive des signatures identifiées, nous avons également mesuré la concordance entre les scores prédits et les temps de survie en utilisant la statistique de concordance de Uno, comme présenté dans la section 3.4.3. Pour le cas pronostique, nous avons classiquement mesuré $\text{C}(\tau)$ la concordance entre le prédicteur linéaire de $h(t, \boldsymbol{T}, \boldsymbol{X})$ noté $\boldsymbol{\pi}$ et les temps de survie. Pour le cas prédictif, nous avons mesuré la différence de statistiques de concordance entre les deux bras de traitement comme proposé et présenté dans la section 4.3.3. On note ce critère

$$\Delta\text{C}(\tau, \boldsymbol{\eta}) = C_{\text{Uno}}(\tau, \boldsymbol{\eta}, \boldsymbol{T} = +0{,}5) - C_{\text{Uno}}(\tau, \boldsymbol{\eta}, \boldsymbol{T} = -0{,}5)$$

avec $\boldsymbol{\eta}$ le score d'interaction.

Pour évaluer les estimations des prédictions de survie individuelles, nous avons comparé pour chaque patient *k*, sa survie prédite $\hat{S}_k(\tau)$ avec sa survie théorique $S_k(\tau)$ estimée à partir du modèle de simulation. Ainsi, la qualité d'estimation de $\hat{S}_k(\tau)$ a été évaluée au travers du biais moyen correspondant à la différence moyenne entre ces deux probabilités de survie

$$\text{Biais moyen} = \frac{1}{n}\sum_{k=1}^{n}\left(\hat{S}_k(\tau) - S_k(\tau)\right),$$

et la précision d'estimation $\hat{S}_k(\tau)$ au travers de son erreur type estimée comme la moyenne des écarts au carré entre la probabilité de survie estimée et sa valeur attendue

$$\text{Erreur type} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(\hat{S}_k(\tau) - E\left(\hat{S}_k(\tau)\right)\right)^2}$$

$$= \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(\hat{S}_k(\tau) - [S_k(\tau) + \text{Biais Moyen}]\right)^2}.$$

Enfin, pour évaluer la qualité des intervalles de confiance, un troisième critère consiste à estimer le taux de couverture empirique

$$\text{Couverture} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{I}\left(\mathrm{q}_{\frac{\theta}{2}}\left(\hat{S}_k(\tau)\right) \leq S_k(\tau) \leq \mathrm{q}_{1-\frac{\theta}{2}}\left(\hat{S}_k(\tau)\right)\right)$$

avec $\mathrm{q}_{\frac{\theta}{2}}\left(\hat{S}_k(\tau)\right)$ et $\mathrm{q}_{1-\frac{\theta}{2}}\left(\hat{S}_k(\tau)\right)$ les bornes de l'intervalle de confiance de $\hat{S}_k(\tau)$ pour un niveau de confiance de $1-\theta$.

5.3.3 Résultats

Les résultats de l'étude de simulation sont présentés dans le Tableau 5.2 et le Tableau 5.3. Ces résultats montrent la capacité de prédiction des modèles de régression soumis à la pénalisation lasso adaptatif sans réestimation des coefficients de régression. Les résultats obtenus avec la pénalisation lasso standard (non-respect de la propriété *oracle*) et la pénalisation lasso adaptatif (respect de la propriété *oracle*) sont relativement proches.

Tableau 5.2 : Précision des modèles sélectionnés à partir du lasso adaptatif (score de Brier, statistiques de concordance)

	Brier score intégré (iBrier)				**Statistique de concordance (C)**				**Δ statistique de concordance (ΔC)**			
	Apprentissage		Validation	Vrai	Apprentissage		Validation	Vrai	Apprentissage		Validation	Vrai
	1cv	2cv	1cv	modèle†	1cv	2cv	1cv	modèle†	1cv	2cv	1cv	modèle†
Coefficients de régression pénalisés												
(**1**) Aucun effet	0,094	0,099	0,099	0,098	0,634	0,497	0,499	0,500	0,071	0,001	0,002	0,000
(**2**) Effet du traitement seul	0,096	0,102	0,101	0,100	0,666	0,586	0,589	0,558	0,065	-0,001	-0,002	0,000
(**3**) 20 biomarqueurs pronostiques	0,097	0,105	0,105	0,102	0,716	0,630	0,640	0,665	0,062	-0,005	-0,002	0,000
(**4**) 15 biomarqueurs prédictifs	0,094	0,106	0,105	0,101	0,726	0,570	0,584	0,641	0,333	0,209	0,229	0,283
(**5**) Effet du traitement + (4)	0,094	0,107	0,106	0,102	0,740	0,621	0,630	0,675	0,332	0,203	0,224	0,284
(**6**) 20 biomarqueurs pronostiques + (5)	0,096	0,111	0,109	0,104	0,768	0,669	0,680	0,718	0,297	0,184	0,207	0,266

Légende. † modèle contenant uniquement les variables ayant un vrai effet. Quantités moyennes basées sur 250 réplications.

Tableau 5.3 : Biais et variabilité de l'estimation ponctuelle de la probabilité de survie à 5 ans et taux de couverture des intervalles de confiance pour les différentes approches à partir de coefficients pénalisés estimés par la pénalisation lasso adaptatif

	Estimation ponctuelle de la probabilité de survie à 5 ans						**Intervalle de confiance à 95% de la prédiction**				
	Biais moyen			Erreur standard			Taux de couverture empirique				
	Point	Spline		Point	Spline		Point		Spline		
	Anly.1cv	Anly.1cv	Anly.2cv	Anly.1cv	Anly.1cv	Anly.2cv	Anly.1cv	Boot	Anly.1cv	Anly.2cv	Boot
Coefficients de régression pénalisés											
(**1**) Aucun effet	0,00	0,00	0,00	0,05	0,05	0,04	0,93	0,96	0,94	0,99	1,00
(**2**) Effet du traitement seul	0,00	0,00	0,00	0,06	0,05	0,04	0,92	0,96	0,93	0,99	1,00
(**3**) 20 biomarqueurs pronostiques	0,00	0,00	0,00	0,08	0,09	0,09	0,91	0,97	0,89	0,92	0,98
(**4**) 15 biomarqueurs prédictifs	0,00	0,00	0,01	0,11	0,13	0,12	0,88	0,96	0,83	0,87	0,93
(**5**) Effet du traitement + (4)	0,00	0,00	0,01	0,11	0,14	0,13	0,88	0,96	0,83	0,87	0,93
(**6**) 20 biomarqueurs pronostiques + (5)	0,00	0,00	0,00	0,13	0,16	0,15	0,88	0,96	0,83	0,86	0,92

Légende. 1cv and 2cv: simple et double validation croisée, Anly: approche analytique, Boot: approche non paramétrique par bootstrap. Quantités moyennes basées sur 250 réplications.

Une exception est à noter pour la construction des intervalles de confiance à partir de l'approche analytique pour laquelle la pénalisation lasso standard affiche de mauvais résultats pour certains scénarios. De plus, il a déjà été montré que la pénalisation lasso standard affichait de moins bons résultats que le lasso adaptatif en matière de sélection d'interactions traitement-biomarqueurs (Chapitre 4), nous avons donc choisi de présenter les résultats de celle-ci en annexes (Annexe A15 et Annexe A16).

Les modèles identifiés par la pénalisation lasso adaptatif montrent de bonnes capacités de prédiction avec des performances relativement proches du vrai modèle (i.e. modèle non pénalisé contenant uniquement les véritables biomarqueurs actifs, Tableau 5.2). En effet, lorsqu'ils sont estimés sur un jeu de données de validation, le score de Brier intégré est relativement bas (variant de 0,099 à 0,109) et la statistique de concordance de Uno (variant de 0,584 à 0,680 pour les scénarios 2–6) ainsi que la différence de statistiques de concordance (variant de 0,207 à 0,229 pour les scénarios 4–6) sont relativement élevées. Lorsque l'on estime ces critères sur le jeu de données d'apprentissage (Anly.1cv) utilisé pour la construction du modèle, les résultats observés sont nettement meilleurs que ceux observés sur un jeu de données de validation traduisant le surapprentissage des modèles. En effet, pour les statistiques de concordance, on observe des variations allant de +0,076 à +0,142 pour le critère C et de +0,064 à +0,108 pour le critère ΔC. Dans le cas où aucun jeu de données externe n'est disponible, la double validation croisée (Anly.2cv) permet de réduire le suroptimisme des modèles en se rapprochant des résultats que l'on obtiendrait sur un jeu de données externe (Tableau 5.2).

Lorsque l'on se focalise sur la capacité de prédiction de la probabilité de survie (Tableau 5.3), la pénalisation lasso adaptatif est également performante en utilisant la technique de simple ou double validation croisée. En effet, le biais moyen est presque nul dans tous les scénarios. Ce constat est moins vrai lorsque la survie attendue à l'horizon τ est proche de 50% (Annexe A14) montrant un léger biais négatif variant de -0,02 à -0,01. En matière de précision, l'erreur standard de l'estimation de la survie ponctuelle est faible en l'absence d'effet de biomarqueurs (variant de 0,04 à 0,06 pour les scénarios 1 et 2) et augmente avec la présence de biomarqueurs pronostiques (0,08–0,09 dans le scénario 3) ou de modificateurs d'effet du traitement (0,11–0,14 pour les scénarios 4 et 5). La variabilité est la plus grande lorsque les deux types de biomarqueurs sont présents simultanément (0,13–0,16 pour le scénario 6). La catégorisation des scores pronostiques et l'utilisation de splines augmentent légèrement cette

variabilité, tout particulièrement dans les scénarios alternatifs. De plus, le choix de réestimer les paramètres de régression afin qu'ils soient non pénalisés augmente la variabilité des estimations de probabilité de survie (erreur standard variant de 0,17 à 0,20 pour le scénario 6, Annexe A13).

Concernant les intervalles de confiance à 95% des prédictions de probabilité de survie, l'approche non paramétrique par rééchantillonnage (bootstrap) affiche de meilleurs résultats que l'approche analytique en ce qui concerne le taux de couverture empirique des intervalles (Tableau 5.3). En effet, en utilisant l'approche par point et non par spline, l'utilisation du bootstrap tend à produire des intervalles de confiance proches et légèrement au-dessus du niveau nominal de 95% (couverture empirique variant de 0,96 à 0,97). En revanche, la couverture des intervalles de confiance produit par l'approche analytique est souvent inférieure au niveau nominal de 95% (variant de 0,91 à 0,93 pour les scénarios nuls et 0,88 pour les scénarios alternatifs). Ces derniers résultats sont encore moins bons lorsque les coefficients de régression sont réestimés dans une seconde étape (couverture empirique variant de 0,65 à 0,71), alors que les résultats restent inchangés pour l'approche par bootstrap (résultats non montrés). Enfin, la catégorisation des scores pronostiques par groupe de risque et l'utilisation de splines réduisent le niveau empirique de couverture des intervalles de confiance pour l'approche analytique et par bootstrap (réduction de 0,88 à 0,83 et de 0,96 à 0,92–0,93 respectivement pour les scénarios alternatifs 4–6) avec une réduction moindre pour l'approche non paramétrique. L'utilisation de splines permet d'utiliser la technique de double validation croisée pour effectuer des prédictions sur un jeu de données externe. Les résultats de cette double validation croisée (Anly.2cv) montrent que cette dernière permet d'améliorer légèrement la qualité des intervalles de confiance avec une augmentation absolue de +0,03 à +0,06 du niveau empirique des intervalles par rapport à l'utilisation d'une simple validation croisée (Anly.1cv). Il est cependant à noter que ce gain n'a pas été observé avec l'utilisation de la pénalisation lasso simple. En effet, une réduction de la couverture empirique des intervalles de confiance est même observée (Annexe A16).

Dans tous les cas, peu de différences en matière de biais, variabilité et taux de couverture ont été observées entre les données d'apprentissage et les données de validation (résultats non montrés), ce qui est synonyme d'un très faible surapprentissage des modèles dans cette étude de simulation.

5.4 Application

Pour illustrer cette approche, nous l'avons appliquée à des données issues d'un essai randomisé contrôlé financé par l'institut américain de la santé (*NIH*) évaluant l'effet du trastuzumab chez 1574 patientes atteintes d'un cancer du sein (Pogue-Geile et al., 2013). Dans cet essai, les patientes ont été randomisées en deux groupes : un bras chimiothérapie seule (n = 795) et un bras chimiothérapie plus trastuzumab (n = 779). La survie sans récidive à distance à 5 ans est égale à 75% [IC95% : 73%–77%]. Dans la population d'étude, il a été montré en moyenne un bénéfice de l'ajout du trastuzumab en matière de survie sans récidive (HR = 0,46, IC95% : 0,38–0,56).

Cependant, il est possible que cet effet ne soit pas le même chez toutes les patientes. Ainsi, nous avons proposé de calculer un score prédictif (ou score d'interaction) pour distinguer les patientes chez lesquelles l'effet du trastuzumab est important, faible voire inexistant ou délétère. De plus, particulièrement dans le cancer du sein, de nombreux outils pronostiques basés sur l'information génomique ont été développés. Nous avons donc proposé de calculer également un score pronostique par patiente afin de prédire sa courbe de survie et aider dans la décision du traitement.

La signature développée (Tableau 5.4) a été implémentée à partir d'un modèle pénalisé avec pénalisation lasso adaptatif. Cette signature est composée de 29 variables pronostiques (4 variables cliniques et 25 variables génomiques) et de 36 variables prédictives (uniquement des variables génomiques). L'effet pronostique d'une seule des variables prédictives a été sélectionné par la régression pénalisée (gène MED13L). Quelques biomarqueurs ayant déjà été identifiés dans la littérature ont été retrouvés comme pronostiques tels que les gènes SOX4 (Song et al., 2015) et CSNK1D (Abba et al., 2007). A noter également la présence de gènes immunitaires tels que CD9 et CCL21 identifiés comme étant prédictifs. De récents articles ont mis en évidence l'impact des voies immunitaires sur l'effet du trastuzumab dans le cancer du sein (André et al., 2013 ; Loi et al., 2014).

Les coefficients de régression des variables sélectionnées n'ont pas été réestimés, il s'agit donc de coefficients pénalisés. La Figure 5.2 représente la probabilité de survie sans récidive à distance à 5 ans en fonction du score prédictif pour différentes classes de risque pronostique. L'estimation ponctuelle de la probabilité de survie a été faite directement à partir des coefficients du Tableau 5.4 (i.e. sans double validation croisée) et les intervalles de confiance ont été générés par bootstrap à partir de B = 200 jeux de données rééchantillonnés.

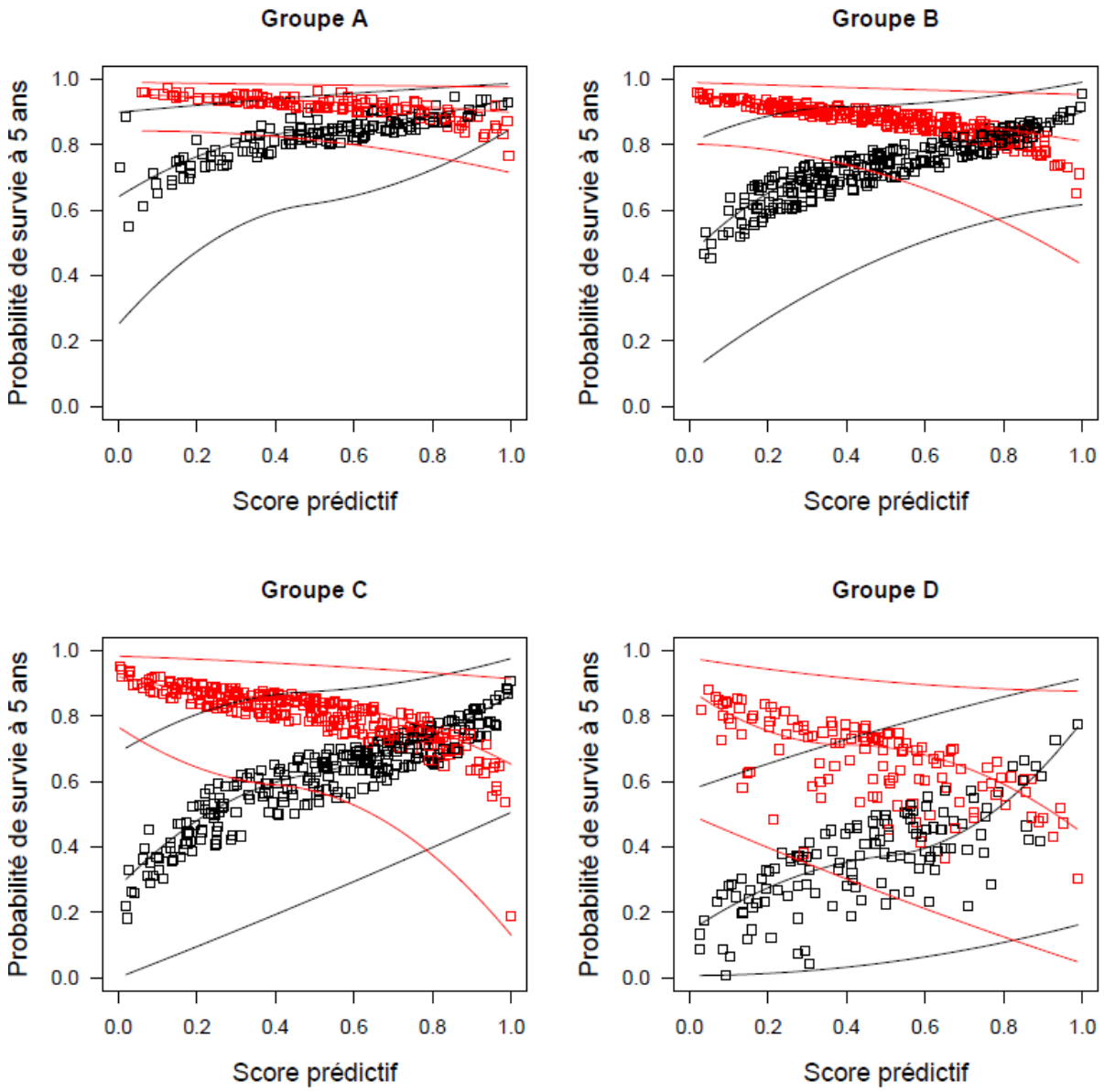
Tableau 5.4 : Signature clinico-génomique développée à partir du lasso adaptatif pour la prédiction de l'effet du trastuzumab dans le cancer du sein

Composante pronostique	
Variables cliniques (p = 4)	Treatment (-0,874*), ER status (-0,182*), Tumor size (0,143*), Nodal status (0,412*)
Variables génomiques (p = 25)	CFLP1 (-0,053), CSNK1D (-0,002), CXXC5 (-0,018), DNAJC4 (-0,018), ELAVL4 (-0,006), FABP5 (0,045), GHR (-0,022), IST1H2AA (-0,037), IGH, (-0,010), IGJ (-0,108), ILF2 (0,016), KIF2C (0,029), MAD2L2 (-0,013), MED13L (-0,033), METTL3 (-0,113), RPS2 (-0,024), SOX4 (0,029), SPP1 (0,070), ST6GALNAC4 (-0,004), SULT1A2 (-0,046), TCEB2 (0,010), TRABD (-0,064), TUBB2C (0,034), XYLT1 (0,060), ZNF609 (-0,009)
Composante prédictive	
Variables génomiques (p=36)	ACBD6 (0,020), ADNP (0,004), ANGPTL4 (0,109), ATAD3A (-0,157), C16orf14 (0,237), C1orf93 (-0,153), CACNG7 (-0,039), CCL21 (-0,105), CD9 (-0,231), CIAPIN1 (-0,096), CLIC1 (0,130), DKFZP434A0131 (0,240), FAM148A (-0,126), FNDC4 (0,071), FURIN (0,082), GPRIN1 (-0,086), GUSBL2 (-0,040), HERC2P4 (-0,010), HNRPAB (-0,006), KRTAP2,4 (0,299), LOC442260 (-0,084), MED13L (0,116), MIA (-0,081), MMD (-0,160), ORMDL3 (-0,015), RPLP0 (0,025), SCNN1D (-0,006), SIAH2 (0,112), SLC25A31 (-0,177), SSBP2 (-0,215), THOP1 (-0,336), THRAP1 (-0,064), TMEM45B (-0,027), TMSB10 (0,118), UGCG (0,076), UNC119 (0,036)

Légende. * variable non pénalisée.

Pour cette application, les quatre groupes de risque pronostique ont également été générés à partir de la distribution proposée par Cox (1957). Ainsi, les quatre groupes allant du meilleur au moins bon pronostic correspondent à : groupe A ($\widehat{\boldsymbol{\varphi}} < 0{,}61$), groupe B ($0{,}61 \leq \widehat{\boldsymbol{\varphi}} < 0{,}77$), groupe C ($0{,}77 \leq \widehat{\boldsymbol{\varphi}} < 0{,}91$) et groupe D ($0{,}91 \geq \widehat{\boldsymbol{\varphi}}$).

D'un point de vue global, la signature a une forte capacité de discrimination des patientes quant à leur probabilité de survie. En effet, dans le groupe de risque le plus bas, la survie des patientes oscille autour de 90% et 80% respectivement dans le bras expérimental et le bras contrôle, alors qu'elle se situe autour de 60% et 30% dans le groupe de risque le plus élevé. Cela se traduit par une bonne statistique de concordance : C = 0,77 pour une simple validation croisée et 0,68 pour une double validation croisée. La composante prédictive permet quant à elle de légèrement discriminer les patientes selon le bénéfice du trastuzumab : plus le score prédictif est bas plus le bénéfice du trastuzumab est important, quel que soit le groupe pronostique. Cela se traduit par une différence de statistiques de concordance modérée : ΔC = +0,27 pour une simple validation croisée et +0,07 pour une double validation croisée.



Légende. rouge : bras chimiothérapie plus trastuzumab, noir : bras chimiothérapie seule, ligne : lissage pour spline. Intervalle de confiance construit à partir des percentiles basés sur 200 rééchantillonnages.

Figure 5.2 : Probabilité de survie à 5 ans en fonction du score prédictif de l'effet du trastuzumab dans le cancer du sein

On remarque que pour ces différents groupes de patientes, le trastuzumab a toujours un effet soit positif soit nul mais jamais réellement délétère. Enfin, il est important de considérer la taille des intervalles de confiance qui est un élément très informatif dans la prise de décision thérapeutique. Ceux-ci sont souvent très grands (e.g. entre 10% et 80%) et se chevauchent largement entre les deux bras.

5.5 Conclusion

Dans ce troisième axe de thèse, nous avons proposé des méthodes visant à prédire la probabilité de survie d'un nouveau patient à un instant *t* avec une mesure d'incertitude à partir d'un modèle de régression pénalisé. Nous nous sommes placés dans le cadre d'un essai contrôlé randomisé et nous avons souhaité implémenter simultanément une signature pronostique et une signature prédictive pour prédire la probabilité individuelle de survie selon le niveau pronostique et le traitement choisi. Dans cette étude, plusieurs stratégies ont été évaluées. Pour le modèle de régression, les pénalisations lasso (non-respect de la propriété *oracle*) et lasso adaptatif (respect de la propriété *oracle*) ont été implémentées, ainsi qu'une approche en deux étapes visant à réestimer les coefficients du modèle pénalisé afin qu'ils ne soient plus pénalisés. Pour l'estimation ponctuelle des prédictions, les approches avec et sans validation croisée ont été évaluées pour tenter d'imiter les résultats que l'on obtiendrait sur un jeu de données indépendant. Enfin, pour l'estimation des bornes de confiance, une approche paramétrique (analytique) et une approche non paramétrique (bootstrap) ont été évaluées. L'évaluation de ces stratégies a été réalisée au travers d'une étude de simulation pour laquelle les paramètres tels que la taille de l'échantillon, le nombre d'événements, le nombre de biomarqueurs et le nombre et les effets des biomarqueurs actifs ont été choisis à partir de l'application présentée dans le cancer du sein.

Les principaux résultats suggèrent d'utiliser le modèle pénalisé pour prédire la probabilité de survie d'un patient à un instant donné plutôt que de réestimer le modèle afin d'avoir des coefficients non pénalisés. En effet, bien que les prédictions issues du modèle pénalisé puissent être légèrement biaisées, notamment lorsque les probabilités de survie attendues sont proches de 50%, la variabilité de ces prédictions est moins importante et la probabilité de couverture des intervalles de confiance est plus proche de la valeur nominale. Nous avons observé une légère amélioration des performances du modèle avec l'utilisation d'une pénalisation garantissant la propriété *oracle* telle que le lasso adaptatif par rapport à une pénalisation (le lasso standard) dépourvue de cette propriété, notamment pour quelques scénarios en ce qui concerne la construction des intervalles de confiance avec la stratégie analytique. Par ailleurs, dans notre étude, l'utilisation de la technique de validation croisée pour réduire le surapprentissage du modèle ne semble pas avoir d'impact en matière de prédiction de probabilités de survie. Il serait intéressant d'évaluer cela pour un échantillon d'apprentissage contenant moins d'observations (d'événements) qui serait potentiellement

plus favorable à un surapprentissage du modèle. L'utilisation de splines et la catégorisation du score pronostique augmentent légèrement la variabilité des prédictions et réduisent légèrement le taux de couverture des différentes approches. Bien que la catégorisation d'une variable continue réduise l'information qui y est contenue (Royston, Altman et Sauerbrei, 2006), dans notre situation et avec l'utilisation de splines cela permet d'avoir une estimation qui soit globalement plus représentative d'un sous-groupe de patients. Un lissage par splines n'aurait pas été possible sans regrouper les patients. Bien entendu, le nombre et le choix des groupes restent arbitraires et différentes approches peuvent être utilisées. Enfin, en matière de couverture des intervalles de confiance, l'approche par bootstrap est celle qui affiche les meilleurs résultats en étant la plus proche du niveau nominal prédéfini bien que parfois trop conservatrice. A l'inverse, l'approche analytique est parfois bien en dessous du niveau nominal prédéfini. De plus, l'approche non paramétrique ne réduit pas de façon importante le taux de couverture si le souhait est d'utiliser des splines. En revanche, contrairement à l'approche analytique, nous n'avons pas implémenté la double validation croisée pour l'approche par bootstrap en raison du temps de calcul extrêmement long.

Concernant les modèles de régression, il serait intéressant d'évaluer d'autres pénalisations telles que l'elastic net adaptatif (Wu, 2012) qui a été implémenté par Sinnott et Cai (2016) et Lin et Halabi (2016). Par ailleurs, dans l'optique d'obtenir une signature de biomarqueurs plus réduite, il serait également intéressant d'évaluer l'apport de notre extension pour le lasso, le lasso-*pcvl* (Ternès et al., 2016a), présentée dans le Chapitre 3. Une signature contenant moins de biomarqueurs peut éventuellement permettre d'avoir des prédictions plus précises mais potentiellement plus biaisées si de véritables biomarqueurs actifs sont écartés de la signature. Toutefois, dans l'étude de simulation présentée, l'utilisation de la *pcvl* sur le lasso standard semble avoir un intérêt limité (Annexe A17 et Annexe A18). En effet, bien que cette dernière réduise confortablement le nombre de biomarqueurs sélectionnés, le lasso-*pcvl* n'améliore pas les résultats en matière d'estimations ponctuelles des probabilités de survie ou de construction des intervalles de confiance par rapport au lasso-*cvl* standard. L'utilisation de coefficients non pénalisés estimés dans une seconde étape semble être plus favorable pour le lasso-*pcvl*. En effet, bien que les estimations ponctuelles soient légèrement plus variables cela améliore la qualité des intervalles de confiance notamment en utilisant l'approche non paramétrique par bootstrap. A noter, la pénalisation *pcvl* a initialement été proposée pour un cadre pronostique et n'a pas encore été évaluée pour le cadre prédictif mélangeant effets propres des biomarqueurs et interactions traitement-biomarqueurs.

Enfin, nous avons appliqué notre méthodologie à des données réelles provenant d'un essai contrôlé randomisé. Bien entendu, il n'a alors pas été possible d'évaluer le biais, la variabilité des prédictions ou le taux de couverture des intervalles de confiance. Néanmoins, un résultat important de cette application est que l'incertitude autour des prédictions individuelles est très grande. Ce résultat avait déjà été souligné dans une autre application par Sinnott et Cai (2016) dans un contexte uniquement pronostique avec parfois des intervalles de confiance allant de 0% à 100%.

Chapitre 6 Conclusion générale

L'objectif de ce travail de thèse a été de développer des méthodologies statistiques adaptées à des données de grande dimension pour l'identification de biomarqueurs et la prédiction individuelle, pour de futurs patients, du bénéfice apporté par le traitement et de la probabilité de survie selon le choix thérapeutique. Dans un premier temps, nous nous sommes intéressés à la sélection de biomarqueurs, qui permettent de comprendre les mécanismes biologiques sous-jacents au cancer et de cibler l'effet des traitements, ce qui est particulièrement important dans le cadre de la médecine stratifiée. Nous souhaitions privilégier une signature avec peu de biomarqueurs pouvant être mesurés avec une bonne capacité prédictive (e.g. ELISA) plutôt que des technologies moins précises (e.g. spectrométrie de masse). Par ailleurs, il a été montré dans la littérature que le nombre de biomarqueurs faussement identifiés comme étant positifs est de plus en plus important au fur et à mesure que l'information disponible est de plus en plus grande à des coûts de plus en plus bas. Ainsi, notre volonté a été de proposer des méthodologies permettant de limiter le nombre de faux positifs tout en conservant le plus de biomarqueurs réellement actifs (i.e. limiter le nombre de faux négatifs) et la plus grande capacité de prédiction. Nous avons travaillé sur deux types de biomarqueurs : les biomarqueurs prédictifs de la survie (appelés pronostiques) et les biomarqueurs prédictifs de l'effet du traitement (appelés biomarqueurs prédictifs ou modificateurs de l'effet du traitement). Bien que les biomarqueurs prédictifs soient nécessaires pour prédire l'effet relatif d'un traitement pour un patient futur, la meilleure recommandation pour effectuer ces

prédictions reste l'estimation de la réduction absolue de risque (Rothwell, 2005) pour laquelle il faut évaluer le pronostic des patients (Windeler, 2000) et donc identifier des biomarqueurs pronostiques.

Pour la sélection de biomarqueurs pronostiques et prédictifs, nous nous sommes essentiellement focalisés sur la pénalisation lasso. Ce choix s'explique car cette pénalisation est relativement populaire pour faire de la sélection de variables en présence de données de grande dimension et est relativement simple à implémenter. Dans le cas pronostique, nous avons proposé une extension de cette pénalisation, appelée lasso-*pcvl* (*penalized cross-validated log-likelihood*), visant à sélectionner moins de biomarqueurs en éliminant majoritairement des faux positifs. La principale limite de notre extension est qu'elle est fondée sur des résultats empiriques et non théoriques. Néanmoins, nous l'avons comparée dans une étude de simulation à onze autres approches de différentes natures afin d'être le plus exhaustif possible. Les résultats de cette étude ont montré que notre extension peut réduire très considérablement le nombre de faux positifs sans réelle augmentation du nombre de faux négatifs et n'est jamais trop conservatrice dans les différents scénarios considérés. Tout comme le lasso standard, une autre limite de notre extension est qu'elle ne garantit pas la propriété *oracle*. Cette propriété, initialement décrite par Fan et Li (2001), est utilisée pour définir une approche permettant dans un échantillon de taille infinie de (i) correctement sélectionner les variables ayant un effet et (ii) estimer leur coefficient sans biais comme s'il était estimé dans le modèle non pénalisé contenant uniquement les variables réellement actives. Cependant, les articles originaux des approches garantissant la propriété *oracle* tels celui de Fan et Li (2001) pour le SCAD, Zou et Hastie (2005) pour l'elastic net, Zou (2006) pour le lasso adaptatif ou encore Wu (2012) pour l'*elastic net* adaptatif présentent uniquement des études de simulation pour un faible nombre de variables (i.e. une dizaine de prédicteurs). En pratique, pour des échantillons de taille finie et avec un nombre de variables important, notre étude de simulation a montré que les méthodes garantissant la propriété *oracle* telles que le lasso adaptatif ou le *stability selection* ne sont pas plus performantes que le lasso standard ou notre extension en matière de sélection. De plus, ces variantes du lasso ont de mauvais résultats dans les scénarios nuls avec des erreurs de type-I extrêmement élevées dans certains cas et nécessitent le choix de paramètres arbitraires. A notre connaissance, l'évaluation de ces méthodes n'avait jamais été réalisée dans ces scénarios nuls. Ainsi, nous proposons d'appliquer notre extension, simple à implémenter et applicable à différents types de

régression, pour estimer le paramètre de pénalisation λ de la régression lasso lorsque l'on cherche à minimiser le nombre de faux positifs tout en limitant les faux négatifs.

Dans le cas prédictif, nous avons présenté différentes paramétrisations de modèles visant à correctement sélectionner les biomarqueurs interagissant avec l'effet du traitement et à estimer cette interaction. Pour la majorité des approches évaluées, la technique de sélection de variables a également été réalisée au travers d'une pénalisation lasso. A ce jour, les articles s'intéressant à la sélection de biomarqueurs prédictifs sont bien moins nombreux que ceux s'intéressant aux biomarqueurs pronostiques. Nous avons proposé et comparé douze approches pour cette sélection. Bien que le nombre d'approches (i.e. modèles et paramétrisations) possibles reste infini, ce travail présente un premier aperçu des avantages et inconvénients d'un large éventail d'approches : la pénalisation des effets propres et des interactions ; la réduction de dimension des effets propres ; la considération des effets pronostiques par bras ; ou d'autres approches qui ne sont pas basées sur la pénalisation lasso pour faire de la sélection telles que le boosting ou l'approche univariée. Pour la grande majorité des approches discutées, nous avons rappelé qu'il est nécessaire d'avoir un nombre d'évènements suffisamment important pour identifier des interactions. Il s'agit d'une des raisons pour laquelle nous n'avons pas implémenté notre extension *pcvl* car cela aurait conduit à une sélection encore plus limitée des interactions. Cependant, certaines approches telles que le lasso groupé ou encore le modèle à deux interactions sélectionnaient énormément de faux positifs et il aurait été intéressant d'appliquer l'extension *pcvl* sur ces méthodes.

En plus de la bonne sélection de biomarqueurs, nous nous sommes intéressés à l'élaboration d'une prédiction individuelle de la probabilité de survie pour de nouveaux patients correspondant au but ultime de la médecine stratifiée. Jusqu'à présent, de nombreux outils de prédiction fournissaient l'estimation ponctuelle des prédictions, cependant une mesure d'incertitude autour de celles-ci était rarement ajoutée pour juger de leur variabilité. Dans notre travail, nous avons évalué différentes approches proposant d'ajouter une mesure d'incertitude aux prédictions. Nous avons également proposé d'utiliser la technique de validation croisée pour imiter les résultats que l'on obtiendrait sur un nouveau jeu de données (i.e. validation externe). Comme attendu, les résultats de l'étude de simulation montrent que les prédictions de la probabilité de survie sont légèrement biaisées lorsqu'elles sont estimées à partir d'un modèle pénalisé. Lorsque l'on réestime le modèle pour obtenir des coefficients non pénalisés, les prédictions sont généralement moins biaisées mais sont plus variables. En matière de couverture des intervalles de confiance, l'approche par bootstrap semble plus

adaptée que l'approche analytique. Enfin, d'une manière générale, l'incertitude autour de ces prédictions individuelles est très importante avec des intervalles de confiance très larges.

Dans l'ensemble de ce travail de thèse, nous avons mis en place différentes études de simulation permettant d'évaluer les approches proposées. L'intérêt des études de simulation est de pouvoir évaluer les performances des approches sur des jeux de données où la réalité est connue (i.e. activité et effet des biomarqueurs, probabilité de survie attendue, etc.), ce qui n'est pas le cas dans de réelles applications. Bien qu'il n'existe pas d'étude de simulation permettant de couvrir l'ensemble des cas de figure possibles, il est important que l'étude de simulation ait des caractéristiques semblables à des cas réels. Cela n'est pas toujours le cas dans la littérature, comme par exemple dans des études proposant des effets des biomarqueurs $\boldsymbol{\beta}$ égaux à 1,5 correspondant à un *HR* proche de 4,5 (Benner et al., 2010). Cela est pertinent pour comprendre les performances des méthodes dans des situations optimales, mais la généralisation des résultats reste alors délicate dans ce cas. Dans les études de simulation que nous avons proposées, nous avons souvent privilégié la proximité avec ce que l'on observe dans de véritables applications en oncologie.

Un point important à souligner est la question de l'intégration à la fois des données cliniques et génomiques dans les modèles de sélection et de prédiction. Pour la sélection de variables, nous nous sommes principalement focalisés sur la sélection des variables génomiques et non des variables cliniques. En oncologie, les variables cliniques pertinentes (i.e. statut nodulaire, taille de la tumeur ou encore le grade) sont souvent connues et il n'est pas toujours nécessaire de faire de sélection sur ces variables. En revanche, pour la prédiction, la question de l'intégration des données cliniques et génomiques est primordiale. En effet, la manière dont les deux types de variables sont combinés peut impacter la qualité de prédiction des modèles. Certains auteurs comme Bøvelstad, Nygård et Borgan (2009) ou De Bin, Sauerbrei et Boulesteix (2014) ont travaillé sur ce sujet avec différents jeux de données réels dans le cadre de données de survie. Les premiers auteurs ont évalué différentes techniques pour établir une prédiction à partir des données génomiques telles qu'un modèle univarié, une réduction de dimension (ACP ou encore *PLS*), une pénalisation ridge ou une pénalisation lasso. En ce qui concerne les variables cliniques, celles-ci sont considérées dans les différents modèles sans sélection ni pénalisation. L'étude proposée par De Bin et al. (2014) évalue simultanément différentes techniques de sélection de variables (les approches univariées et multivariées avec ou sans ajustement sur les variables cliniques, le lasso ou encore le *boosting*) et différentes approches permettant de combiner les variables cliniques et génomiques. Les méthodes

étudiées par les auteurs comprenaient une approche dite naïve, une approche fixant les variables cliniques en *offset*, une approche considérant les variables cliniques non pénalisées et deux approches réduisant les informations sous forme de score : un score uniquement pour les données génomiques ou un score pour les données cliniques et un autre pour les données génomiques. Le principal inconvénient de ces deux études est que l'évaluation des approches proposées n'est réalisée que sur des applications réelles et non sur des études de simulation. Il est vrai que l'inclusion de variables cliniques dans les études de simulation n'est pas chose simple pour la compréhension des scénarios et nécessite des choix supplémentaires arbitraires. C'est pourquoi nous n'avons pas simulé de variables cliniques dans nos études de simulation. A notre connaissance, aucun travail ne propose d'études simulant des données contenant à la fois des variables cliniques et génomiques.

Dans ce travail, l'intérêt s'est porté sur le développement d'une signature génomique, cela reste néanmoins une première étape dans son processus de validation. Des groupes de travail tels que l'EGAPP (*Evaluation of Genomic Applications in Practice and Prevention*) ont proposé des critères pour la validation de résultats (Teutsch et al., 2009) et ceux-ci ont été récemment transposés pour la validation de signatures génomiques (Michiels, Ternès et Rotolo, 2016). Un des critères que nous n'avons pas réellement discuté dans ce manuscrit concerne la valeur ajoutée de la signature génomique. En effet, l'intérêt d'une signature génomique sera limité si elle apporte peu de valeur ajoutée à une signature clinique déjà existante, simple et peu coûteuse à mettre en place. Il s'agit d'un point longuement discuté par Boulesteix et Sauerbrei (2011) qui est aussi observé par Bøvelstad et al. (2009) et De Bin et al. (2014) dans différentes applications pour lesquelles la signature génomique apporte dans certains cas très peu d'information supplémentaire comparée à la signature clinique. D'une manière très générale, l'utilisation d'un échantillon indépendant au développement d'une signature reste primordiale pour sa validation (Michiels et al., 2005 ; Michiels, Koscielny et Hill, 2007 ; Simon, Paik et Hayes, 2009).

Enfin, l'ensemble de ce travail de thèse a été réalisé à partir du logiciel `R` et les codes implémentant les différentes approches discutées sont facilement utilisables pour d'autres applications. De plus, ils sont publiquement disponibles, ayant été soumis pour publication en même temps que les manuscrits. Ainsi, une perspective de ce travail est de développer et mettre à disposition des utilisateurs un package `R` implémentant l'ensemble des codes avec la soumission d'un article scientifique expliquant leur utilisation.

Références

Abba, M. C., Sun, H., Hawkins, K. A., Drake, J. A., Hu, Y., Nunez, M. I., … Sahin, A. (2007). Breast cancer molecular signatures as determined by SAGE: correlation with lymph node status. *Molecular Cancer Research*, *5*(9), 881–890.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Amado, R. G., Wolf, M., Peeters, M., Van Cutsem, E., Siena, S., Freeman, D. J., … Chang, D. D. (2008). Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of Clinical Oncology*, *26*(10), 1626–34.

Andre, F., Dieci, M. V., Dubsky, P., Sotiriou, C., Curigliano, G., Denkert, C., & Loi, S. (2013). Molecular pathways: involvement of immune pathways in the therapeutic response and outcome in breast cancer. *Clinical cancer research*, *19*(1), 28-33.

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning* (pp. 33–40). ACM.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, *101*(473), 119–137.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W., Ledoux, P., … Edgar, R. (2005). NCBI GEO : mining millions of expression profiles — database and tools. *Nucleic Acids Research*, *33*, 562–566.

Bastien, P., & Tenenhaus, M. (2001). PLS generalized linear regression. Application to the analysis of life time data. In *Proceedings of the PLS'01 International Symposium, Anacapri (Italy)* (pp. 131–140).

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*, *57*(1), 289–300.

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., & Mansmann, U. (2010). High-dimensional Cox models: the choice of penalty as part of the model building process. *Biometrical Journal*, *52*(1), 50–69.

Bertrand, F., Maumy-Bertrand, M., & Meyer, N. (2015). Partial Least Squares Regression for Cox Models and Related Techniques. *R-Package Version 1.7.2*.

Bien, J., Taylor, J., & Tibshirani, R. (2013). a Lasso for Hierarchical Interactions. *The Annals of Statistics*, *41*(3), 1111–1141.

Bleakley, K., Biau, G., & Vert, J. (2007). Supervised reconstruction of biological networks with local models. *Bioinformatics*, *23*, i57–i65.

Bonetti, M., & Gelber, R. D. (2015). A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. *Statistics in Medicine*, *19*, 2595–2609.

Boulesteix, A.-L., & Sauerbrei, W. (2011). Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics*, *12*(3), 215–229.

Bøvelstad, H. M., Nygård, S., & Borgan, Ø. (2009). Survival prediction from clinico-genomic models-a comparative study. *BMC Bioinformatics*, *10*, 413.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., & Lingjaerde, O. C. (2007). Predicting survival from microarray data--a comparative study. *Bioinformatics*, *23*(16), 2080–2087.

Breiman, L., & Spector, P. (1992). Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review*, *60*(3), 291–319.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89–99.

Breslow, N. E. (1972). Contribution to the discussion of the paper by DR Cox. *Journal of the Royal Statistical Society: Series B*, *34*(2), 216–217.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Bühlmann, P., & Yu, B. (2003). Boosting with the L 2 loss: regression and classification. *Journal of the American Statistical Association*, *98*(462), 324–339.

Buyse, M., & Michiels, S. (2013). Omics-based clinical trial designs. *Current Opinion in Oncology*, *25*(3), 289–95.

Chatterjee, A., & Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, *106*(494), 608–625.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Cheng, C.-J., Lin, Y.-C., Tsai, M.-T., Chen, C.-S., Hsieh, M.-C., Chen, C.-L., & Yang, R.-B. (2009). SCUBE2 suppresses breast tumor cell proliferation and confers a favorable prognosis in invasive breast cancer. *Cancer Research*, *69*(8), 3634–3641.

Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association*, *52*(280), 543-547.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, *34*(2), 187–220.

Culhane, A. C., Schröder, M. S., Sultana, R., Picard, S. C., Martinelli, E. N., Kelly, C., … Quackenbush, J. (2012). GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Research*, *40*, D1060–1066.

Davis, S., & Meltzer, P. S. (2007). GEOquery : a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, *23*(14), 1846–1847.

De Bin, R., Sauerbrei, W., & Boulesteix, A. (2014). Investigating the prediction ability of

survival models based on both clinical and omics data : two case studies. *Statistics in Medicine*, *33*(30), 5310–5329.

Desmedt, C., Di Leo, A., de Azambuja, E., Larsimont, D., Haibe-Kains, B., Selleslags, J., … Sotiriou, C. (2011). Multifactorial approach to predicting resistance to anthracyclines. *Journal of Clinical Oncology*, *29*(12), 1578–86.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, *78*(382), 316–331.

Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., & Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, *30*(1), 74–99.

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*(3), 928–961.

Farcomeni, A., & La, R. (2007). A review of modern multiple hypothesis testing , with particular attention to the false. *Statistical Methods in Medical Research*, 1–42.

Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, *22*(4), 1947–1975.

Frank, Ll. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109–135.

Friedman, J. H. (2001). Greedy function approximation: a gradient bosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Friedman, J., Hastie, T., Simon, N., & Tibshirani, R. (2016). Lasso and Elastic-Net Regularized Generalized Linear Models. *R-Package Version 2.0-5*.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1).

Gehrmann, M., & Von Törne, C. (2009, September 24). Prediction of Breast Cancer Response to Taxane-Based Chemotherapy. Google Patents. Retrieved from http://www.google.com/patents/US20090239223

Genovese, C., & Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B*, *64*(3), 499–517.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., & Dudoit, S. (2006). *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Science & Business Media.

Goeman, J., Meijer, R., & Chaturvedi, N. (2016). penalized: L1 (Lasso and Fused Lasso) and L2 (Ridge) Penalized Estimation in GLMs and in the Cox Model. *R-Package Version 0.9-47*.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, *18*(17-18),

2529-2545.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B*, *41*, 190–195.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543–2546.

Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., & Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, *3*(2), 143–152.

Harrell, F. E, Lee, K. L. & Mark, D. B. (1996). Tutorial in Biostatistics: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequaty, and measuring and reducing errors. *Statistics in Medicine*, 23, 2109–2123.

Hastie, T., Taylor, J., Tibshirani, R., & Walther, G. (2007). Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, *1*, 1–29. http://doi.org/10.1214/07-EJS004

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning* (pp. 485–585). Springer.

Hatzis, C., Pusztai, L., Valero, V., Booser, D. J., Esserman, L., Lluch, A., … Symmans, W. F. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA : The Journal of the American Medical Association*, *305*(18), 1873–81.

Haute Autorité de Santé. (2014). Test compagnon associé à une thérapie ciblée : définitions et méthode d'évaluation - Guide méthodologique. *Http://www.has-sante.fr/portail/jcms/c_1735034/fr/*.

Hingorani, A. D., van der Windt, D. A., Riley, R. D., Abrams, K., Moons, K. G. M., Steyerberg, E. W., … Hemingway, H. (2013). Prognosis research strategy (PROGRESS) 4: stratified medicine research. *British Medical Journal*, *345*, e5793.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *42*(1), 80–86.

Hofner, B., Mayr, A., Robinzonov, N., & Schmid, M. (2014). Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computational Statistics*, *29*(1-2), 3–35.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417.

Huang, J., Ma, S., & Zhang, C. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, *18*, 1603–1618.

Huang, Y., Pepe, M. S., & Feng, Z. (2007). Evaluating the Predictiveness of a Continuous Marker. *Biometrics*, *63*, 1181–1188.

Jansen, M. P. H. M., Sas, L., Sieuwerts, A. M., Van Cauwenberghe, C., Ramirez-Ardila, D., Look, M., … Van Laere, S. (2015). Decreased expression of ABAT and STC2 hallmarks ER-positive inflammatory breast cancer and endocrine therapy resistance in advanced

disease. *Molecular Oncology*, *9*(6), 1218–33.

Kalisch, M., & Bühlmann, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, *8*, 613–636.

Kang, C., Janes, H., & Huang, Y. (2014). Combining Biomarkers to Optimize Patient Treatment Recommendations. *Biometrics*, *70*, 695–720.

Knight, K., & Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *ICML* (Vol. 97, pp. 179–186). Nashville, USA.

Lazar, A. A., Cole, B. F., Bonetti, M., & Gelber, R. D. (2010). Evaluation of Treatment-Effect Heterogeneity Using Biomarkers Measured on a Continuous Scale : Subpopulation Treatment Effect Pattern Plot. *Journal of Clinical Oncology*, *28*, 4539–4544.

Lin, C., & Halabi, S. (2016). A Simple Method for Deriving the Confidence Regions for the Penalized Cox's Model via the Minimand Perturbation. *Communications in Statistics - Theory and Methods*. In press.

Lin, C.-Y., & Halabi, S. (2013). On model specification and selection of the Cox proportional hazards model. *Statistics in Medicine*, *32*(26), 4609–4623.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, *42*(2), 413–468.

Loi, S., Michiels, S., Salgado, R., Sirtaine, N., Jose, V., Fumagalli, D., ... & Piccart, M. J. (2014). Tumor infiltrating lymphocytes are prognostic in triple negative breast cancer and predictive for trastuzumab benefit in early breast cancer: results from the FinHER trial. *Annals of oncology*, *25*(8), 1544-1550.

Maathuis, M. H., Kalisch, M., & Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, *37*(6A), 3133–3164.

MacArthur, D. (2012). Face up to false positives. *Nature*, *487*(7408), 427–428.

Magbanua, M. J. M., Wolf, D. M., Yau, C., Davis, S. E., Crothers, J., Au, A., … van 't Veer, L. J. (2015). Serial expression analysis of breast tumors during neoadjuvant chemotherapy reveals changes in cell cycle and immune pathways associated with recurrence and response. *Breast Cancer Research*, *17*, 73.

Marín-Aguilera, M., Codony-Servat, J., Kalko, S. G., Fernández, P. L., Bermudo, R., Buxo, E., … Mellado, B. (2011). Identification of docetaxel resistance genes in castration-resistant prostate cancer. *Molecular Cancer Therapeutics*, *11*(2), 329–339.

Martens, H., & Naes, T. (1989). Assessment, validation and choice of calibration method. *Multivariate Calibration*, 237–266.

Matsui, S., Simon, R., Qu, P., Matsui, S., Simon, R., Qu, P., … Crowley, J. (2012). Developing and Validating Continuous Genomic Signatures in Randomized Clinical Trials for Predictive Medicine. *Clinical Cancer Research*, *18*, 6065–6073.

Mccall, M. N., Bolstad, B. M., & Irizarry, R. A. (2010). Frozen robust multiarray analysis (fRMA). *Biostatistics*, *11*(2), 242–253.

Mehra, R., Varambally, S., Ding, L., Shen, R., Sabel, M. S., Ghosh, D., … Kleer, C. G. (2005). Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer Research*, *65*(24), 11259–11264.

Meier, L. (2015). Fitting user specified models with Group Lasso penalty. *R-Package Version 0.4-5*.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis*, *52*(1), 374–393.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, *34*(3), 1436–1462.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B*, *72*(4), 417–473.

Michiels, S., Koscielny, S., & Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, *365*, 488–492.

Michiels, S., Koscielny, S., & Hill, C. (2007). Interpretation of microarray data in cancer. *British Journal of Cancer*, *96*(8), 1155–8.

Michiels, S., Potthoff, R. F., & George, S. L. (2011). Multiple testing of treatment-effect-modifying biomarkers in a randomized clinical trial with a survival endpoint. *Statistics in Medicine*, *30*(13), 1502–1518.

Michiels, S., Ternès, N., & Rotolo, F. (2016). Statistical Controversies in Clinical Research: Prognostic gene signatures are not (yet) useful in clinical practice. *Annals of Oncology*. In press.

Minnier, J., Tian, L., & Cai, T. (2012). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*.

Mok, T. S., Wu, Y., Thongprasert, S., Yang, C., Saijo, N., Sunpaweravong, P., … Fukuoka, M. (2009). Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma. *New England Journal of Medicine*, *361*(10), 947–957.

Müller, S., & Welsh, A. H. (2010). On Model Selection Curves. *International Statistical Review*, *78*(2), 240–256.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, *1*(1), 27–52.

NIH Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology Therapeutics*, *69*, 89–95.

Nygård, S., Borgan, Ø., Lingjærde, O. C., & Størvold, H. L. (2006). Partial least squares Cox regression on genomic data handling additional covariates. *Preprint Series. Statistical Research Report Http://urn. Nb. no/URN: NBN: No-23420*.

Pang, H., Tong, T., & Zhao, H. (2009). Shrinkage-based diagonal discriminant analysis and

its applications in high-dimensional data. *Biometrics*, *65*(4), 1021–1029.

Park, M. Y., & Hastie, T. (2015). glmpath: L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model. *R-Package Version 0.97*.

Park, P. J., Tian, L., & Kohane, I. S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, *18*, S120–S127.

Patel, K. M., & Hoel, D. G. (1973). A nonparametric test for interaction in factorial experiments. *Journal of the American Statistical Association*, *68*(343), 615–620.

Pawitan, Y. (2013). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., & Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics (Oxford, England)*, *21*(13), 3017–24.

Pearl, J. (2000). *Causality: models, reasoning and inference* (Vol. 29). Cambridge Univ Press.

Pencina, M. J., & Agostino, R. B. D. (2004). Overall C as a measure of discrimination in survival analysis : model speciÿc population value and conÿdence interval estimation. *Statistics in Medicine*, *23*, 2109–2123.

Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M., & Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*, *167*(3), 362–8.

Perez, E. a, Thompson, E. A., Ballman, K. V, Anderson, S. K., Asmann, Y. W., Kalari, K. R., … Reinholz, M. M. (2015). Genomic analysis reveals that immune function genes are strongly linked to clinical outcome in the North Central Cancer Treatment Group n9831 Adjuvant Trastuzumab Trial. *Journal of Clinical Oncology*, *33*(7), 701–8.

Plant, D., Wilson, A. G., & Barton, A. (2014). Genetic and epigenetic predictors of responsiveness to treatment in RA. *Nat Rev Rheumatol*, *10*(6), 329–337. Retrieved from

Pogue-Geile, K. L., Kim, C., Jeong, J.-H., Tanaka, N., Bandos, H., Gavin, P. G., … Paik, S. (2013). Predicting degree of benefit from adjuvant trastuzumab in NSABP trial B-31. *Journal of the National Cancer Institute*, *105*(23), 1782–1788.

Poste, G. (2011). Bring on the biomarkers. *Nature*, *469*, 156–257.

Roberts, S., & Nowak, G. (2014). Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis*, *70*, 198–211.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*, *365*(9454), 176–86.

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression : a bad idea. *Statistics in Medicine*, 127–141.

Royston, P., & Sauerbrei, W. (2004). A new measure of prognostic separation in survival data. *Statistics in Medicine*, *748*, 723–748.

Sabourin, J. A., Valdar, W., & Nobel, A. B. (2015). A Permutation Approach for Selecting the Penalty Parameter in Penalized Model Selection. *Biometrics*, *71*, 1185–1194.

Schäfer, J., Opgen-rhein, R., Zuber, V., Ahdesmäki, A., Duarte Silva, A., & Strimmer, K. (2015). Efficient Estimation of Covariance and (Partial) Correlation. *R-Package Version 1.6.8*.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227.

Schemper, M. (1988). Non-parametric analysis of treatment-covariate interaction in the presence of censoring. *Statistics in Medicine*, *7*, 1257–1266.

Schmid, M., & Potapov, S. (2012). A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in Medicine*, *31*(23), 2588–2609.

Schneider, D., Bianchini, G., Horgan, D., Michiels, S., Witjes, W., Hills, R., … Lawler, M. (2015). Establishing the Evidence Bar for Molecular Diagnostics in Personalised Cancer Care. *Public Health Genomics*, *18*(6), 349–358.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shabalin, A. a, Tjelmeland, H., Fan, C., Perou, C. M., & Nobel, A. B. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, *24*(9), 1154–1160.

Simon, R. M., Paik, S., & Hayes, D. F. (2009). Use of Archived Specimens in Evaluation of Prognostic and Predictive Biomarkers. *Journal of National Cancer Institute*, *101*, 1446–1452.

Simon, R. M., Subramanian, J., Li, M.-C., & Menezes, S. (2011). Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings in Bioinformatics*, *12*(3), 203–214.

Sinnott, J. A., & Cai, T. (2016). Inference for survival prediction under the regularized Cox model. *Biostatistics*, 1–16.

Song, G.-D., Sun, Y., Shen, H., & Li, W. (2015). SOX4 overexpression is a novel biomarker of malignant status and poor prognosis in breast cancer patients. *Tumor Biology*, *36*(6), 4167–4173.

Stekhoven, D. J., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M. H., & Bühlmann, P. (2012). Causal stability ranking. *Bioinformatics*, *28*(21), 2819–23.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., … Kattan, M. W. (2010). Assessing the Performance of Prediction Models A Framework for Traditional and Novel Measures. *Epidemiology*, *21*, 128–138.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, *36*(2), 111–147.

Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-Theory and Methods*, *7*(1), 13–26.

Ternès, N., Arnedos, M., Koscielny, S., Michiels, S., & Lanoy, E. (2014). Statistical methods applied to omics data : predicting response to neoadjuvant therapy in breast cancer. *Current Opinion in Oncology*, *26*(6), 576–583.

Ternès, N., Rotolo, F., & Michiels, S. (2016a). Empirical extensions of the lasso penalty to reduce the false discovery rate in high- dimensional Cox regression models. *Statistics in Medicine*, *35*(15), 2561–2573.

Ternès, N., Rotolo, F., Heinze, G., & Michiels, S. (2016b). Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. *Biometrical Journal*. Accepted.

Teutsch, S. M., Bradley, L. A., Palomaki, G. E., Haddow, J. E., Piper, M., Calonge, N., … Berg, A. O. (2009). The evaluation of genomic applications in practice and prevention (EGAPP) initiative: methods of the EGAPP working group. *Genetics in Medicine*, *11*(1), 3–14.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.

Therneau, T. M., & Lumley, T. (2016). survival: Survival Analysis. *R-Package Version 2.39-4*.

Tian, L., Alizadeh, A. a, Gentles, A. J., & Tibshirani, R. (2014). A Simple Method for Estimating Interactions between a Treatment and a Large Number of Covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532.

Tiang, J. M., Butcher, N. J., & Minchin, R. F. (2015). Effects of human arylamine N-acetyltransferase I knockdown in triple-negative breast cancer cell lines. *Cancer Medicine*, *4*(4), 565–74.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*(1), 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, *16*, 385–395.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso : a retrospective. *Journal of the Royal Statistical Society, Series B*, *73*, 273–282.

Tibshirani, R., & Saunders, M. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, *67*(1), 91–108.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, *98*(9), 5116–5121.

Ulloa-Montoya, F., Louahed, J., Dizier, B., Gruselle, O., Spiessens, B., Lehmann, F. F., … Brichard, V. G. (2013). Predictive gene signature in MAGE-A3 antigen-specific cancer immunotherapy. *Journal of Clinical Oncology*, *31*(19), 2388–95.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, *30*(10), 1105–1117.

van ’t Veer, L. J., Dai, H., Vijver, M. J. Van De, Kooy, K. Van Der, Marton, M. J., Witteveen, A. T., … Linsley, P. S. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530–536.

van de Geer, S., Bühlmann, P., & Zhou, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics*, *5*, 688–749.

Verweij, P. J. M., & van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine*, *13*, 2427–2436.

Verweij, P. J., & van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, *12*(24), 2305–2314.

Wang, E., Li, J., O’connor-Mccourt, M., & Purisima, E. (2013, June 6). Paclitaxel response markers for cancer. Google Patents. Retrieved from http://www.google.com/patents/CA2857191A1?cl=en

Weiss, S. M. (1991). Small sample error rate estimation for k-NN classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(3), 285–289.

Whitehead, J. (1980). Fitting Cox’s regression model to survival data using GLIM. *Applied Statistics*, *29*(3), 268–275.

Windeler, J. (2000). Prognosis - what does the clinician associate with this notion ? *Statistics in Medicine*, *19*, 425–430.

Wu, Y. (2012). Elastic net for Cox’s proportional hazards model. *Statistica Sinica*, *22*, 27.

Yang, H., Tang, R., Hale, M., & Huang, J. (2015). A visualization method measuring the performance of biomarkers for guiding treatment decisions. *Pharmaceutical Statistics*. http://doi.org/10.1002/pst.1728

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with. *Journal of the Royal Statistical Society: Series B*, *68*(1), 49–67.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*(1), 19–35.

Yuan, Y. (2016). Lots Of Lasso. *R-Package Version 1.20.0*.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942.

Zhang, H. H., & Lu, W. (2007). Adaptive lasso for Cox’s proportional hazards model. *Biometrika*, *94*(3), 691–703.

Zhao, L., Tian, L., Cai, T., Claggett, B., & Wei, L. J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, *108*(502), 527–539.

Zhao, P., & Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American*

Statistical Association, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320. http://doi.org/10.1111/j.1467-9868.2005.00503.x

Table des annexes

A Résultats complémentaires

Annexe A1 : Chapitre 3. Impact du choix du nombre de folds dans le processus de validation croisée sur le taux de fausses découvertes et le taux de faux négatifs (*FDR* / *FNR*)

			lasso-*cvl*		
q	***p***	***n***	**3 *folds***	**5 *folds***	**10 *folds***
		100	0,45 / -	0,38 / -	0,28 / -
	10	*500*	0,43 / -	0,32 / -	0,29 / -
		1000	0,42 / -	0,36 / -	0,29 / -
		100	0,54 / -	0,42 / -	0,30 / -
0	*100*	*500*	0,53 / -	0,45 / -	0,34 / -
		1000	0,49 / -	0,42 / -	0,35 / -
		100	0,78 / -	0,47 / -	0,40 / -
	1000	*500*	0,58 / -	0,48 / -	0,34 / -
		1000	0,62 / -	0,49 / -	0,36 / -
		100	0,48 / 0,40	0,41 / 0,47	0,32 / 0,51
	10	*500*	0,66 / 0,01	0,59 / 0,02	0,52 / 0,02
		1000	0,66 / 0,00	0,60 / 0,00	0,52 / 0,00
		100	0,67 / 0,63	0,52 / 0,68	0,47 / 0,71
1	*100*	*500*	0,81 / 0,06	0,77 / 0,06	0,70 / 0,04
		1000	0,87 / 0,00	0,82 / 0,00	0,74 / 0,00
		100	0,75 / 0,82	0,51 / 0,84	0,42 / 0,87
	1000	*500*	0,88 / 0,07	0,80 / 0,12	0,69 / 0,14
		1000	0,96 / 0,00	0,92 / 0,00	0,85 / 0,00
		100	0,44 / 0,33	0,38 / 0,39	0,35 / 0,38
2	*10*	*500*	0,62 / 0,00	0,56 / 0,00	0,52 / 0,00
		1000	0,59 / 0,00	0,54 / 0,00	0,50 / 0,00
		100	0,58 / 0,57	0,53 / 0,60	0,51 / 0,61
10	*100*	*500*	0,70 / 0,01	0,68 / 0,01	0,67 / 0,00
		1000	0,72 / 0,00	0,70 / 0,00	0,68 / 0,00
		100	0,67 / 0,90	0,49 / 0,92	0,46 / 0,92
20	*1000*	*500*	0,80 / 0,03	0,78 / 0,04	0,76 / 0,04
		1000	0,86 / 0,00	0,83 / 0,00	0,81 / 0,00

Légende. *q* : nombre de biomarqueurs actifs, *p* : nombre de biomarqueurs, *n* : taille d'échantillon. Quantités moyennes basées sur 250 réplications.

Annexe A2 : Chapitre 3. Impact du choix de la procédure d'estimation des poids pour le lasso adaptatif sur le taux de fausse découverte et le taux de faux négatifs (*FDR* / *FNR*)

			lasso-*cvl*	**lasso adaptatif**		
q	***p***	***n***		**univarié**	**ridge**	**lasso-*cvl***
		100	0,38 / -	0,52 / -	0,53 / -	0,38 / -
	10	*500*	0,32 / -	0,53 / -	0,52 / -	0,32 / -
		1000	0,36 / -	0,50 / -	0,51 / -	0,36 / -
		100	0,42 / -	0,84 / -	0,81 / -	0,42 / -
0	*100*	*500*	0,45 / -	0,82 / -	0,81 / -	0,45 / -
		1000	0,42 / -	0,82 / -	0,81 / -	0,42 / -
		100	0,47 / -	0,99 / -	0,97 / -	0,47 / -
	1000	*500*	0,48 / -	0,99 / -	0,99 / -	0,48 / -
		1000	0,49 / -	0,99 / -	0,98 / -	0,49 / -
		100	0,41 / 0,47	0,45 / 0,37	0,45 / 0,38	0,33 / 0,50
	10	*500*	0,59 / 0,02	0,42 / 0,01	0,43 / 0,01	0,35 / 0,02
		1000	0,60 / 0,00	0,37 / 0,00	0,36 / 0,00	0,30 / 0,00
		100	0,52 / 0,68	0,87 / 0,47	0,84 / 0,48	0,52 / 0,69
1	*100*	*500*	0,77 / 0,06	0,80 / 0,01	0,81 / 0,01	0,68 / 0,06
		1000	0,82 / 0,00	0,76 / 0,00	0,79 / 0,00	0,67 / 0,00
		100	0,51 / 0,84	0,97 / 0,63	0,94 / 0,68	0,51 / 0,84
	1000	*500*	0,80 / 0,12	0,97 / 0,01	0,97 / 0,01	0,79 / 0,12
		1000	0,92 / 0,00	0,98 / 0,00	0,98 / 0,00	0,88 / 0,00
		100	0,38 / 0,39	0,39 / 0,33	0,38 / 0,32	0,30 / 0,47
2	*10*	*500*	0,56 / 0,00	0,37 / 0,00	0,38 / 0,00	0,32 / 0,01
		1000	0,54 / 0,00	0,31 / 0,00	0,30 / 0,00	0,23 / 0,00
		100	0,53 / 0,60	0,61 / 0,51	0,64 / 0,50	0,48 / 0,66
10	*100*	*500*	0,68 / 0,01	0,55 / 0,01	0,58 / 0,01	0,53 / 0,01
		1000	0,70 / 0,00	0,51 / 0,00	0,55 / 0,00	0,53 / 0,00
		100	0,49 / 0,92	0,84 / 0,80	0,85 / 0,80	0,47 / 0,93
20	*1000*	*500*	0,78 / 0,04	0,75 / 0,05	0,77 / 0,05	0,70 / 0,06
		1000	0,83 / 0,00	0,75 / 0,00	0,78 / 0,00	0,77 / 0,00

Légende. *q* : nombre de biomarqueurs actifs, *p* : nombre de biomarqueurs, *n* : taille d'échantillon.
Quantités moyennes basées sur 250 réplications.

Annexe A3 : Chapitre 3. Impact des paramètres du *stability selection* sur le taux de fausses découvertes et le taux de faux négatifs (*FDR* / *FNR*)

			lasso sans randomisation ($\zeta = 1$)				lasso avec randomisation ($\zeta = 0{,}5$)			
			$\alpha_{FWER} = 0{,}05p$		$\alpha_{FWER} = 0{,}01p$		$\alpha_{FWER} = 0{,}05p$		$\alpha_{FWER} = 0{,}01p$	
q	***p***	***n***	$\pi_{thr} = 0{,}6$	$\pi_{thr} = 0{,}3$	$\pi_{thr} = 0{,}6$	$\pi_{thr} = 0{,}3$	$\pi_{thr} = 0{,}6$	$\pi_{thr} = 0{,}3$	$\pi_{thr} = 0{,}6$	$\pi_{thr} = 0{,}3$
		100	0,26 / -	0,17 / -	0,07 / -	0,07 / -	0,19 / -	0,19 / -	0,06 / -	0,07 / -
	10	*500*	0,29 / -	0,30 / -	0,10 / -	0,08 / -	0,24 / -	0,22 / -	0,06 / -	0,06 / -
		1000	0,23 / -	0,21 / -	0,08 / -	0,07 / -	0,25 / -	0,24 / -	0,08 / -	0,07 / -
		100	0,88 / -	0,91 / -	0,40 / -	0,39 / -	0,83 / -	0,84 / -	0,31 / -	0,34 / -
0	*100*	*500*	0,98 / -	0,97 / -	0,52 / -	0,54 / -	0,95 / -	0,94 / -	0,39 / -	0,43 / -
		1000	0,97 / -	0,97 / -	0,56 / -	0,58 / -	0,97 / -	0,94 / -	0,47 / -	0,44 / -
		100	0,34 / -	0,37 / -	0,36 / -	0,33 / -	0,18 / -	0,16 / -	0,16 / -	0,18 / -
	1000	*500*	1,00 / -	1,00 / -	0,99 / -	1,00 / -	1,00 / -	1,00 / -	0,89 / -	0,92 / -
		1000	1,00 / -	1,00 / -	1,00 / -	1,00 / -	1,00 / -	1,00 / -	0,95 / -	0,96 / -
		100	0,15 / 0,74	0,12 / 0,71	0,03 / 0,92	0,05 / 0,91	0,12 / 0,74	0,12 / 0,76	0,03 / 0,86	0,03 / 0,88
	10	*500*	0,05 / 0,16	0,07 / 0,14	0,00 / 0,58	0,01 / 0,57	0,03 / 0,24	0,05 / 0,19	0,01 / 0,68	0,01 / 0,61
		1000	0,03 / 0,08	0,04 / 0,09	0,00 / 0,43	0,00 / 0,50	0,03 / 0,09	0,04 / 0,11	0,00 / 0,56	0,00 / 0,49
		100	0,83 / 0,69	0,83 / 0,70	0,44 / 0,78	0,44 / 0,80	0,81 / 0,71	0,79 / 0,71	0,37 / 0,82	0,32 / 0,83
1	*100*	*500*	0,70 / 0,03	0,72 / 0,02	0,26 / 0,07	0,24 / 0,08	0,67 / 0,03	0,68 / 0,04	0,20 / 0,11	0,20 / 0,11
		1000	0,70 / 0,00	0,71 / 0,00	0,28 / 0,00	0,26 / 0,00	0,67 / 0,00	0,67 / 0,00	0,19 / 0,00	0,22 / 0,00
		100	0,35 / 0,96	0,36 / 0,95	0,36 / 0,95	0,36 / 0,94	0,17 / 0,97	0,18 / 0,96	0,21 / 0,96	0,17 / 0,97
	1000	*500*	0,93 / 0,12	0,93 / 0,09	0,80 / 0,08	0,80 / 0,08	0,92 / 0,12	0,92 / 0,13	0,68 / 0,09	0,67 / 0,14
		1000	0,96 / 0,00	0,96 / 0,00	0,84 / 0,00	0,83 / 0,00	0,95 / 0,00	0,95 / 0,00	0,76 / 0,00	0,73 / 0,00
		100	0,10 / 0,74	0,08 / 0,74	0,02 / 0,92	0,02 / 0,92	0,05 / 0,76	0,07 / 0,77	0,03 / 0,92	0,02 / 0,94
2	*10*	*500*	0,01 / 0,40	0,01 / 0,39	0,00 / 0,80	0,00 / 0,76	0,01 / 0,43	0,01 / 0,46	0,00 / 0,82	0,00 / 0,81
		1000	0,00 / 0,28	0,00 / 0,28	0,00 / 0,74	0,00 / 0,72	0,00 / 0,35	0,00 / 0,32	0,00 / 0,82	0,00 / 0,79
		100	0,41 / 0,76	0,42 / 0,76	0,24 / 0,89	0,21 / 0,89	0,39 / 0,78	0,38 / 0,78	0,17 / 0,91	0,16 / 0,91
10	*100*	*500*	0,18 / 0,10	0,17 / 0,09	0,05 / 0,45	0,06 / 0,45	0,16 / 0,11	0,16 / 0,11	0,05 / 0,53	0,04 / 0,51
		1000	0,15 / 0,00	0,15 / 0,00	0,03 / 0,26	0,03 / 0,25	0,13 / 0,00	0,14 / 0,00	0,02 / 0,34	0,02 / 0,34
		100	0,26 / 0,99	0,29 / 0,99	0,26 / 0,99	0,27 / 0,99	0,15 / 0,99	0,14 / 0,99	0,17 / 0,99	0,11 / 0,99
20	*1000*	*500*	0,45 / 0,26	0,45 / 0,26	0,23 / 0,31	0,23 / 0,31	0,41 / 0,28	0,41 / 0,28	0,18 / 0,38	0,17 / 0,39
		1000	0,52 / 0,01	0,52 / 0,01	0,21 / 0,01	0,21 / 0,01	0,46 / 0,01	0,48 / 0,01	0,16 / 0,02	0,16 / 0,03

Légende. q : nombre de biomarqueurs actifs, p : nombre de biomarqueurs, n : taille d'échantillon. Quantités moyennes basées sur 250 réplications.

Annexe A4 : Chapitre 3. Statistique de concordance de Uno dans les scénarios alternatifs (q > 0) à partir des coefficients de régression pénalisés

q	***p***	***n***	**lasso-*cvl***	**lasso-*1se***	**lasso-*pcvl***	**lasso-*AIC***	**lasso-*RIC***	**lasso adaptatif**
		100	0,511	0,501	0,510	0,503	0,510	0,510
	10	*500*	0,524	0,506	0,525	0,523	0,525	0,524
		1000	0,524	0,508	0,525	0,524	0,525	0,524
		100	0,506	0,500	0,505	0,501	0,503	0,506
1	*100*	*500*	0,521	0,503	0,523	0,517	0,523	0,517
		1000	0,523	0,508	0,524	0,524	0,524	0,520
		100	0,503	0,500	0,503	0,500	0,504	0,503
	1000	*500*	0,516	0,502	0,518	0,509	0,519	0,510
		1000	0,522	0,506	0,524	0,523	0,524	0,515
		100	0,521	0,504	0,521	0,506	0,519	0,521
2	*10*	*500*	0,534	0,518	0,535	0,534	0,535	0,535
		1000	0,535	0,529	0,535	0,535	0,535	0,535
		100	0,542	0,510	0,539	0,506	0,524	0,540
10	*100*	*500*	0,576	0,575	0,576	0,576	0,551	0,575
		1000	0,580	0,581	0,581	0,581	0,581	0,580
		100	0,518	0,501	0,517	0,500	0,511	0,517
20	*1000*	*500*	0,588	0,583	0,587	0,519	0,527	0,585
		1000	0,600	0,602	0,602	0,602	0,546	0,595

Légende. q : nombre de biomarqueurs actifs, p : nombre de biomarqueurs, n : taille d'échantillon. Quantités moyennes basées sur 250 réplications.

Annexe A5 : Chapitre 3. Multiplicateur θ_λ pour l'ensemble des extensions pénalisées dans les différents scénarios

q	***p***	***n***	**lasso-*cvl***	**lasso-*pcvl*[1]**	**lasso-*AIC***	**lasso-*RIC*[1]**	**lasso-*HQIC***	**lasso-*BIC***	***lasso-eBIC***
		100	0	0,15	2	≤ 0,77	3,05	4,61	9,21
	10	*500*	0	0,14	2	≤ 0,66	3,65	6,21	10,8
		1000	0	0,17	2	≤ 0,75	3,87	6,91	11,5
		100	0	0,07	2	≤ 1,14	3,05	4,61	13,8
0	*100*	*500*	0	0,08	2	≤ 1,28	3,65	6,21	15,4
		1000	0	0,09	2	≤ 1,15	3,87	6,91	16,1
		100	0	0,05	2	≤ 1,37	3,05	4,61	18,4
	1000	*500*	0	0,04	2	≤ 1,75	3,65	6,21	20,0
		1000	0	0,04	2	≤ 1,93	3,87	6,91	20,7
		100	0	0,51	2	≤ 1,39	3,05	4,61	9,21
	10	*500*	0	3,96	2	≤ 2,20	3,65	6,21	10,8
		1000	0	8,68	2	≤ 2,22	3,87	6,91	11,5
		100	0	0,18	2	≤ 1,80	3,05	4,61	13,8
1	*100*	*500*	0	1,35	2	≤ 3,62	3,65	6,21	15,4
		1000	0	3,42	2	≤ 3,93	3,87	6,91	16,1
		100	0	0,08	2	≤ 1,83	3,05	4,61	18,4
	1000	*500*	0	0,51	2	≤ 4,86	3,65	6,21	20,0
		1000	0	1,39	2	≤ 5,52	3,87	6,91	20,7
		100	0	0,89	2	≤ 1,82	3,05	4,61	9,21
2	*10*	*500*	0	4,79	2	≤ 3,19	3,65	6,21	10,8
		1000	0	10,2	2	≤ 3,12	3,87	6,91	11,5
		100	0	0,41	2	≤ 4,21	3,05	4,61	13,8
10	*100*	*500*	0	2,75	2	≤ 6,93	3,65	6,21	15,4
		1000	0	5,99	2	≤ 7,05	3,87	6,91	16,1
		100	0	0,10	2	≤ 2,83	3,05	4,61	18,4
20	*1000*	*500*	0	0,84	2	≤ 8,98	3,65	6,21	20,0
		1000	0	2,17	2	≤ 9,56	3,87	6,91	20,7

Légende. q : nombre de biomarqueurs actifs, p : nombre de biomarqueurs, n : taille d'échantillon. [1]Quantités moyennes basées sur 250 réplications.

Annexe A6 : Chapitre 3. Taux de fausses découvertes et taux de faux négatifs (*FDR*/*FNR*) pour les cinq méthodes non discutées dans le manuscrit

q	***p***	***n***	**lasso-*HQIC***	**lasso-*BIC***	**lasso-*eBIC***	**lasso-*AICC***	***percentile lasso***
		100	0,00 / -	0,00 / -	0,00 / -	0,33 / -	0,07 / -
	10	*500*	0,00 / -	0,00 / -	0,00 / -	0,31 / -	0,07 / -
		1000	0,01 / -	0,00 / -	0,00 / -	0,35 / -	0,09 / -
		100	0,00 / -	0,00 / -	0,00 / -	0,33 / -	0,08 / -
0	*100*	*500*	0,00 / -	0,00 / -	0,00 / -	0,41 / -	0,08 / -
		1000	0,00 / -	0,00 / -	0,00 / -	0,40 / -	0,09 / -
		100	0,02 / -	0,02 / -	0,02 / -	0,34 / -	0,10 / -
	1000	*500*	0,00 / -	0,00 / -	0,00 / -	0,41 / -	0,13 / -
		1000	0,00 / -	0,00 / -	0,00 / -	0,46 / -	0,14 / -
		100	0,01 / 0,94	0,00 / 0,98	0,00 / 0,99	0,38 / 0,50	0,16 / 0,74
	10	*500*	0,05 / 0,24	0,01 / 0,46	0,00 / 0,64	0,58 / 0,02	0,41 / 0,04
		1000	0,06 / 0,01	0,02 / 0,05	0,02 / 0,10	0,59 / 0,00	0,43 / 0,00
		100	0,00 / 0,99	0,00 / 1,00	0,00 / 1,00	0,44 / 0,71	0,18 / 0,85
1	*100*	*500*	0,03 / 0,52	0,00 / 0,71	0,00 / 0,89	0,76 / 0,06	0,58 / 0,14
		1000	0,08 / 0,05	0,02 / 0,14	0,00 / 0,38	0,82 / 0,00	0,69 / 0,00
		100	0,01 / 0,99	0,01 / 1,00	0,01 / 1,00	0,42 / 0,86	0,15 / 0,94
	1000	*500*	0,01 / 0,70	0,00 / 0,81	0,00 / 0,98	0,78 / 0,12	0,67 / 0,20
		1000	0,06 / 0,13	0,01 / 0,26	0,00 / 0,61	0,91 / 0,00	0,85 / 0,00
		100	0,02 / 0,91	0,00 / 0,96	0,00 / 0,98	0,36 / 0,42	0,19 / 0,62
2	*10*	*500*	0,10 / 0,13	0,03 / 0,35	0,01 / 0,58	0,56 / 0,00	0,45 / 0,00
		1000	0,13 / 0,00	0,07 / 0,01	0,04 / 0,04	0,54 / 0,00	0,44 / 0,00
		100	0,01 / 0,99	0,00 / 1,00	0,00 / 1,00	0,49 / 0,62	0,34 / 0,75
10	*100*	*500*	0,24 / 0,27	0,03 / 0,82	0,00 / 1,00	0,68 / 0,01	0,65 / 0,01
		1000	0,33 / 0,00	0,25 / 0,00	0,11 / 0,33	0,70 / 0,00	0,66 / 0,00
		100	0,00 / 1,00	0,00 / 1,00	0,00 / 1,00	0,43 / 0,92	0,19 / 0,98
20	*1000*	*500*	0,00 / 1,00	0,00 / 1,00	0,00 / 1,00	0,78 / 0,04	0,75 / 0,05
		1000	0,36 / 0,05	0,04 / 0,81	0,00 / 1,00	0,83 / 0,00	0,80 / 0,00

Légende. *q* : nombre de biomarqueurs actifs, *p* : nombre de biomarqueurs, *n* : taille d'échantillon. Quantités moyennes basées sur 250 réplications.

Annexe A7 : Chapitre 3. Impact de la corrélation entre les biomarqueurs en matière de taux de fausses découvertes et taux de faux négatifs (*FDR* / *FNR*)

q	***p***	***n***	**lasso-*cvl***	**lasso-*1se***	**lasso-*pcvl***	**lasso-*AIC***	**lasso-*RIC***	**lasso adaptatif**	***stability selection***
		100	+0,01 / -	–0,01 / -	+0,01 / -	+0,00 / -	+0,01 / -	+0,01 / -	–0,06 / -
	10	*500*	–0,04 / -	+0,00 / -	–0,04 / -	+0,00 / -	–0,04 / -	–0,04 / -	–0,05 / -
		1000	–0,08 / -	+0,00 / -	–0,08 / -	+0,00 / -	–0,08 / -	–0,08 / -	–0,14 / -
		100	+0,04 / -	+0,00 / -	+0,04 / -	+0,01 / -	+0,04 / -	+0,04 / -	+0,01 / -
0	*100*	*500*	+0,02 / -	+0,00 / -	+0,02 / -	+0,00 / -	+0,02 / -	+0,02 / -	+0,00 / -
		1000	–0,02 / -	+0,00 / -	–0,02 / -	+0,00 / -	–0,02 / -	–0,02 / -	+0,00 / -
		100	+0,08 / -	+0,00 / -	+0,08 / -	+0,02 / -	+0,08 / -	+0,08 / -	–0,05 / -
	1000	*500*	–0,02 / -	+0,00 / -	–0,02 / -	+0,00 / -	–0,02 / -	–0,02 / -	+0,00 / -
		1000	–0,05 / -	+0,00 / -	–0,05 / -	+0,00 / -	–0,05 / -	–0,05 / -	+0,00 / -
		100	+0,07 / –0,02	+0,01 / –0,02	+0,07 / +0,00	+0,01 / +0,03	+0,05 / –0,01	+0,05 / +0,02	+0,02 / +0,01
	10	*500*	+0,05 / –0,01	+0,00 / +0,02	+0,03 / –0,02	–0,01 / +0,00	+0,08 / +0,00	+0,05 / –0,02	+0,02 / +0,03
		1000	–0,06 / –0,01	+0,00 / +0,00	–0,06 / –0,03	+0,00 / +0,00	–0,04 / +0,00	–0,06 / –0,02	–0,15 / –0,02
		100	–0,04 / +0,00	–0,01 / +0,05	–0,01 / +0,00	+0,00 / –0,01	+0,01 / +0,00	–0,08 / +0,00	+0,02 / –0,04
1	*100*	*500*	–0,03 / +0,03	+0,00 / +0,00	+0,01 / +0,03	+0,01 / +0,04	+0,06 / +0,06	–0,07 / +0,04	–0,02 / +0,02
		1000	–0,03 / +0,00	+0,01 / –0,02	+0,00 / +0,00	+0,02 / +0,01	–0,02 / +0,00	–0,02 / +0,00	–0,01 / +0,00
		100	–0,03 / +0,00	+0,00 / +0,03	+0,03 / +0,00	+0,00 / +0,00	+0,00 / +0,00	–0,03 / +0,00	+0,01 / +0,02
	1000	*500*	–0,02 / +0,00	+0,00 / +0,03	+0,04 / +0,00	+0,04 / +0,00	+0,04 / +0,00	–0,04 / +0,00	+0,00 / +0,00
		1000	+0,03 / +0,00	+0,00 / +0,02	+0,00 / +0,00	–0,01 / +0,02	+0,01 / +0,00	+0,01 / +0,00	+0,00 / +0,00
		100	+0,00 / +0,04	+0,00 / –0,03	+0,04 / +0,02	+0,01 / –0,02	+0,01 / –0,01	+0,01 / +0,05	–0,02 / +0,01
2	*10*	*500*	+0,09 / –0,07	+0,02 / –0,05	+0,08 / –0,04	+0,02 / –0,02	+0,04 / –0,01	+0,08 / –0,05	+0,01 / –0,02
		1000	–0,04 / –0,01	+0,00 / +0,00	–0,05 / –0,01	+0,00 / +0,00	–0,02 / +0,00	–0,05 / –0,01	–0,04 / +0,00
		100	–0,02 / +0,00	+0,01 / +0,09	+0,04 / +0,01	+0,03 / +0,00	+0,03 / +0,01	–0,01 / +0,00	+0,01 / +0,04
10	*100*	*500*	–0,02 / +0,00	+0,08 / +0,01	+0,03 / +0,02	+0,10 / –0,02	+0,08 / –0,15	–0,04 / +0,01	+0,08 / +0,03
		1000	+0,07 / +0,00	+0,10 / –0,12	+0,07 / +0,00	+0,03 / –0,04	+0,05 / +0,00	+0,06 / +0,00	+0,00 / +0,00
		100	–0,02 / +0,00	+0,00 / +0,04	+0,02 / +0,00	+0,01 / +0,00	+0,04 / +0,00	–0,01 / +0,00	+0,00 / +0,05
20	*1000*	*500*	–0,03 / +0,00	+0,04 / +0,00	+0,05 / +0,00	+0,06 / +0,00	+0,05 / +0,00	–0,08 / +0,00	+0,06 / +0,00
		1000	+0,07 / –0,01	+0,04 / –0,07	+0,06 / +0,00	–0,06 / –0,09	+0,03 / –0,02	+0,06 / –0,01	+0,00 / +0,01

Légende. *q* : nombre de biomarqueurs actifs, *p* : nombre de biomarqueurs, *n* : taille d'échantillon. Quantités moyennes basées sur 250 réplications. Différence de *FDR* et *FNR* entre les scénarios avec corrélation et les scénarios sans corrélation.

Annexe A8 : Chapitre 3. Impact du taux de censure en matière de taux de fausses découvertes et taux de faux négatifs (*FDR* / *FNR*)

q	***p***	***n***	**lasso-*cvl***	**lasso-*1se***	**lasso-*pcvl***	**lasso-*AIC***	**lasso-*RIC***	**lasso adaptatif**	***stability selection***
		100	+0,02 / -	–0,17 / -	+0,02 / -	+0,01 / -	+0,02 / -	+0,02 / -	+0,00 / -
	10	*500*	–0,03 / -	–0,17 / -	–0,03 / -	+0,02 / -	–0,03 / -	–0,03 / -	+0,03 / -
		1000	+0,03 / -	–0,12 / -	+0,03 / -	+0,02 / -	+0,03 / -	+0,03 / -	–0,02 / -
		100	–0,03 / -	–0,24 / -	–0,03 / -	+0,00 / -	–0,03 / -	–0,03 / -	–0,04 / -
0	*100*	*500*	+0,04 / -	–0,19 / -	+0,04 / -	+0,00 / -	+0,04 / -	+0,04 / -	+0,02 / -
		1000	–0,03 / -	–0,23 / -	–0,03 / -	+0,00 / -	–0,03 / -	–0,03 / -	–0,02 / -
		100	–0,13 / -	–0,21 / -	–0,13 / -	+0,00 / -	–0,13 / -	–0,13 / -	–0,07 / -
	1000	*500*	–0,02 / -	–0,24 / -	–0,02 / -	+0,00 / -	–0,02 / -	–0,02 / -	+0,00 / -
		1000	+0,00 / -	–0,26 / -	+0,00 / -	+0,00 / -	+0,00 / -	+0,00 / -	+0,00 / -
		100	+0,05 / –0,04	–0,13 / +0,24	+0,04 / –0,04	+0,01 / –0,02	–0,01 / –0,08	+0,01 / –0,03	+0,03 / +0,02
	10	*500*	–0,02 / +0,02	–0,16 / +0,76	+0,00 / +0,02	–0,01 / +0,05	+0,00 / +0,01	–0,04 / +0,02	–0,01 / +0,00
		1000	–0,02 / +0,00	–0,11 / +0,67	+0,03 / +0,00	–0,01 / +0,00	+0,00 / +0,00	+0,00 / +0,00	+0,00 / –0,02
		100	–0,01 / –0,02	–0,26 / +0,16	–0,03 / –0,02	–0,01 / +0,00	–0,02 / –0,03	–0,01 / –0,04	+0,04 / +0,00
1	*100*	*500*	–0,01 / +0,03	–0,36 / +0,80	+0,03 / +0,03	–0,05 / +0,10	–0,01 / +0,04	–0,01 / +0,03	+0,00 / +0,02
		1000	–0,01 / +0,00	–0,24 / +0,69	+0,03 / +0,00	+0,02 / +0,01	+0,03 / +0,00	–0,01 / +0,00	–0,02 / +0,00
		100	–0,06 / +0,01	–0,27 / +0,07	–0,06 / +0,02	+0,00 / +0,00	–0,04 / +0,02	–0,06 / –0,01	–0,06 / +0,00
	1000	*500*	–0,03 / +0,04	–0,53 / +0,74	+0,01 / +0,03	+0,00 / +0,07	–0,01 / +0,04	–0,06 / +0,06	+0,01 / +0,04
		1000	+0,00 / +0,00	–0,45 / +0,77	+0,05 / +0,00	+0,03 / +0,01	+0,01 / +0,00	–0,01 / +0,00	+0,00 / +0,00
		100	–0,02 / +0,02	–0,12 / +0,28	+0,00 / +0,01	+0,01 / –0,02	–0,02 / –0,01	–0,01 / +0,02	–0,03 / +0,00
2	*10*	*500*	+0,01 / +0,00	–0,14 / +0,50	+0,01 / +0,01	+0,01 / +0,01	+0,00 / +0,01	+0,02 / +0,00	–0,01 / +0,03
		1000	–0,01 / +0,00	–0,08 / +0,18	+0,02 / +0,00	+0,02 / +0,00	+0,03 / +0,00	+0,02 / +0,00	+0,00 / +0,01
		100	+0,02 / +0,03	–0,22 / +0,16	–0,01 / +0,03	+0,02 / –0,01	–0,01 / +0,00	–0,01 / +0,03	+0,01 / +0,03
10	*100*	*500*	+0,01 / +0,00	–0,11 / +0,04	+0,02 / +0,01	+0,02 / +0,02	+0,01 / –0,01	+0,02 / +0,01	+0,01 / +0,05
		1000	+0,00 / +0,00	–0,09 / +0,00	+0,01 / +0,00	+0,03 / +0,00	+0,02 / +0,00	+0,01 / +0,00	+0,01 / +0,00
		100	+0,01 / –0,01	–0,20 / +0,03	+0,03 / –0,01	–0,01 / +0,00	+0,02 / +0,00	+0,01 / –0,01	+0,03 / +0,01
20	*1000*	*500*	+0,01 / +0,01	–0,12 / +0,11	+0,01 / +0,03	+0,03 / –0,06	+0,01 / +0,00	+0,01 / +0,02	+0,01 / +0,06
		1000	+0,01 / +0,00	–0,08 / +0,00	+0,03 / +0,00	+0,04 / +0,00	+0,01 / –0,01	+0,01 / +0,00	+0,01 / +0,00

Légende. *q* : nombre de biomarqueurs actifs, *p* : nombre de biomarqueurs, *n* : taille d'échantillon. Quantités moyennes basées sur 250 réplications. Différence de *FDR* et *FNR* entre les scénarios avec censure et les scénarios sans censure.

Annexe A9 : Chapitre 3. Taux de fausses découvertes et taux de faux négatifs (*FDR* / *FNR*) dans les scénarios avec des effets variables dans le scénario alternatif 2 ($q > 1$)

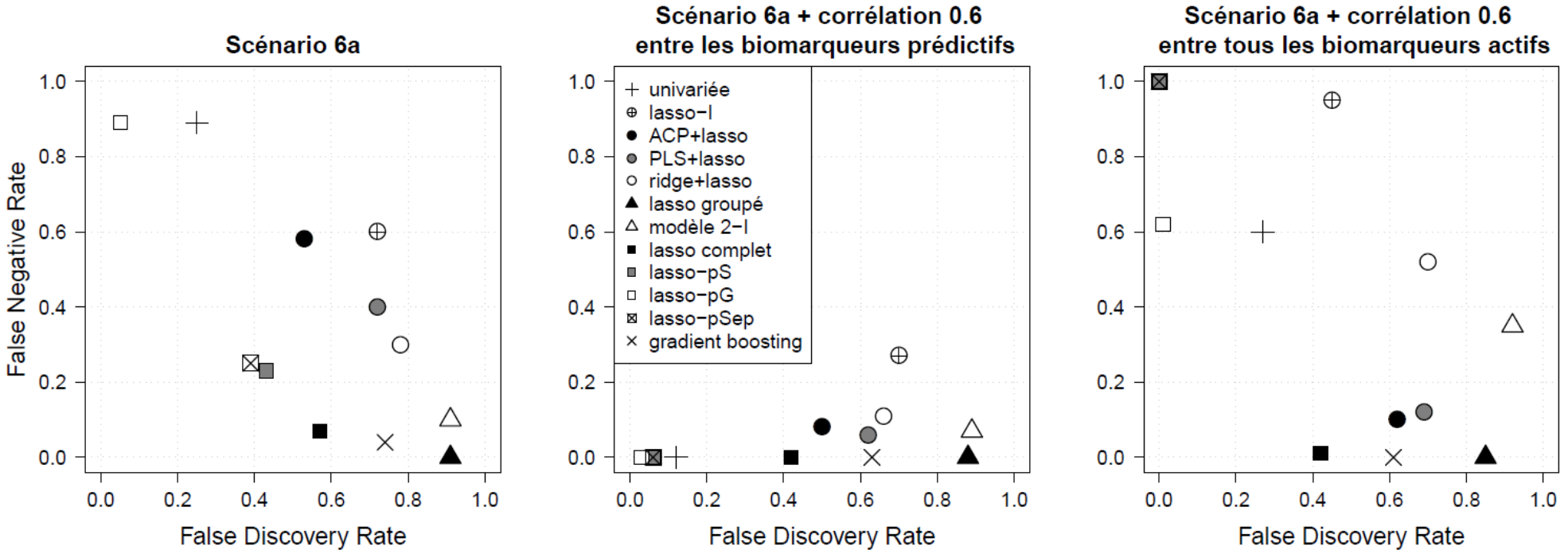
q	***p***	***n***	**lasso-*cvl***	**lasso-*1se***	**lasso-*pcvl***	**lasso-*AIC***	**lasso-*RIC***	**lasso adaptatif**	***stability selection***
		100	0,37 / 0,41	0,01 / 0,90	0,21 / 0,49	0,02 / 0,84	0,13 / 0,67	0,27 / 0,44	0,08 / 0,76
2	*10*	*500*	0,54 / 0,03	0,01 / 0,50	0,09 / 0,11	0,16 / 0,14	0,09 / 0,12	0,34 / 0,04	0,01 / 0,40
		1000	0,58 / 0,00	0,01 / 0,31	0,03 / 0,06	0,18 / 0,03	0,08 / 0,03	0,29 / 0,01	0,00 / 0,32
		100	0,47 / 0,65	0,04 / 0,95	0,35 / 0,74	0,00 / 0,98	0,13 / 0,93	0,42 / 0,69	0,40 / 0,74
10	*100*	*500*	0,70 / 0,03	0,27 / 0,14	0,28 / 0,11	0,31 / 0,12	0,05 / 0,42	0,58 / 0,04	0,13 / 0,16
		1000	0,72 / 0,00	0,26 / 0,02	0,19 / 0,03	0,36 / 0,01	0,13 / 0,05	0,59 / 0,01	0,11 / 0,04
		100	0,48 / 0,92	0,01 / 1,00	0,43 / 0,94	0,00 / 1,00	0,29 / 0,98	0,47 / 0,93	0,34 / 0,98
20	*1000*	*500*	0,80 / 0,11	0,48 / 0,24	0,50 / 0,22	0,2 / 0,51	0,00 / 0,92	0,74 / 0,13	0,48 / 0,27
		1000	0,84 / 0,02	0,59 / 0,04	0,39 / 0,07	0,41 / 0,07	0,09 / 0,25	0,80 / 0,02	0,52 / 0,06

Légende. q : nombre de biomarqueurs actifs, p : nombre de biomarqueurs, n : taille d'échantillon. Quantités moyennes basées sur 250 réplications.

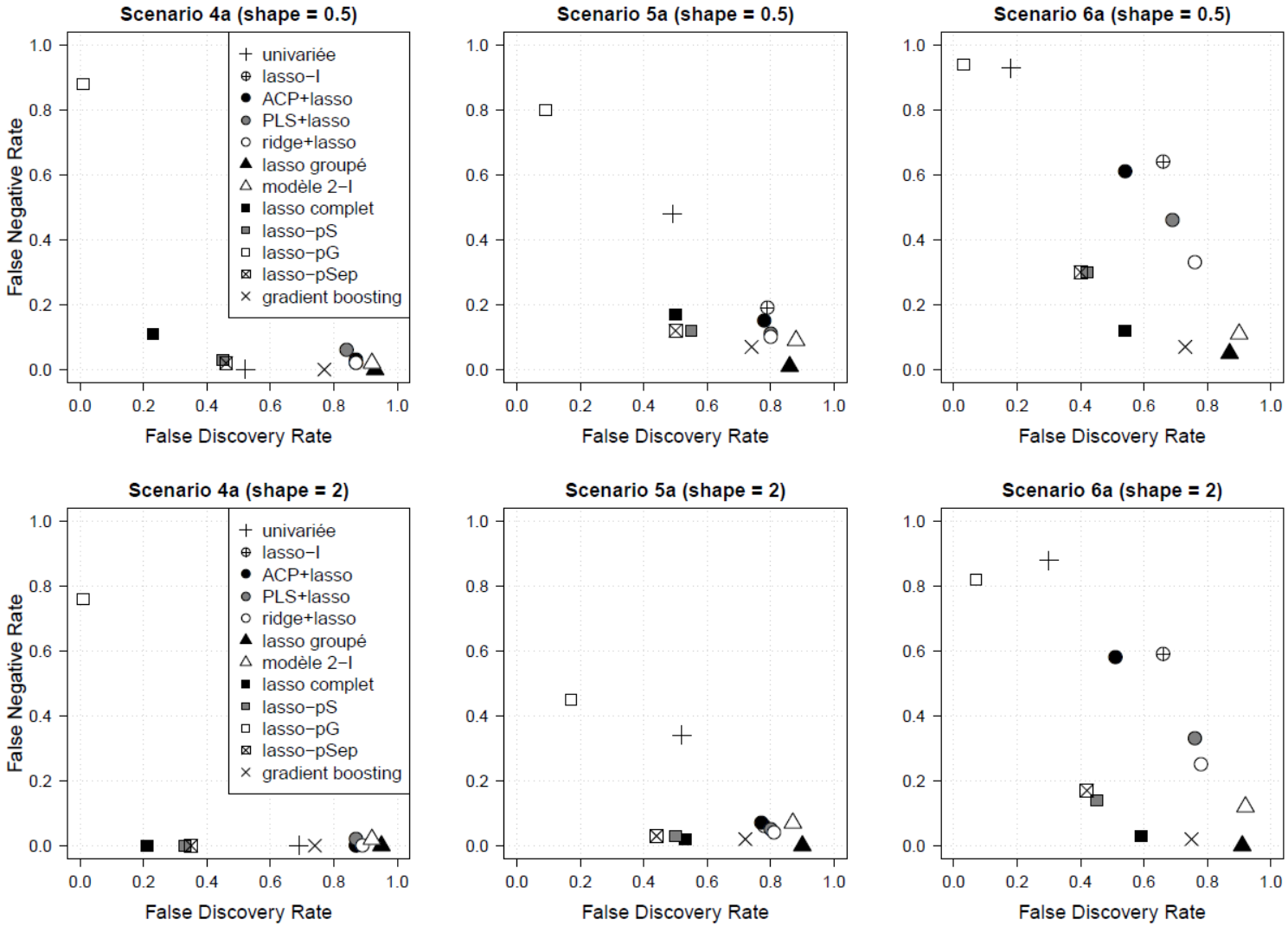
Annexe A10 : Chapitre 4. Impact de la corrélation sur l'approche univariée en matière de taux de fausses découvertes et taux de faux négatifs (*FDR* / *FNR*) dans l'identification d'interactions

	Approche univariée	
	Corrélation	**Sans corrélation**
Scénario 1a	0,07 / -	0,04 / -
Scénario 2a	0,06 / -	0,07 / -
Scénario 3a	0,06 / -	0,10 / -
Scénario 4a	0,63 / 0,00	0,06 / 0,00
Scénario 5a	0,52 / 0,41	0,08 / 0,49
Scénario 6a	0,25 / 0,89	0,06 / 0,93

Annexe A11 : Chapitre 4. Taux de faux négatifs (*FNR*) en fonction du taux de fausses découvertes (*FDR*) dans les scénarios alternatifs pour différentes structures de corrélation entre les biomarqueurs actifs



Annexe A12 : Chapitre 4. Taux de faux négatifs (*FNR*) en fonction du taux de fausses découvertes (*FDR*) dans les scénarios alternatifs pour un risque décroissant (1ère ligne) et un risque croissant (2ème ligne) au cours du temps



Annexe A13 : Chapitre 5. Biais et variabilité de l'estimation ponctuelle de la probabilité de survie à 5 ans et taux de couverture des intervalles de confiance pour les différentes approches *à partir de coefficients non pénalisés* estimés par la pénalisation lasso adaptatif

	Estimation ponctuelle de la probabilité de survie à 5 ans						**Intervalle de confiance à 95% de la prédiction**				
	Biais moyen			Erreur standard			Taux de couverture empirique				
	Point	Spline		Point	Spline		Point		Spline		
	Anly.1cv	Anly.1cv	Anly.2cv	Anly.1cv	Anly.1cv	Anly.2cv	Anly.1cv	Boot	Anly.1cv	Anly.2cv	Boot
Coefficients de régression non pénalisés											
(**1**) Aucun effet	0,00	0,00	0,01	0,10	0,09	0,06	0,68	0,96	0,71	0,85	0,99
(**2**) Effet du traitement seul	0,00	0,00	0,01	0,11	0,10	0,06	0,68	0,96	0,71	0,84	1,00
(**3**) 20 biomarqueurs pronostiques	0,00	0,01	0,01	0,12	0,12	0,11	0,73	0,96	0,73	0,79	0,99
(**4**) 15 biomarqueurs prédictifs	0,00	0,01	0,03	0,16	0,18	0,15	0,68	0,96	0,65	0,70	0,95
(**5**) Effet du traitement + (4)	0,01	0,01	0,03	0,16	0,19	0,16	0,68	0,96	0,65	0,70	0,96
(**6**) 20 biomarqueurs pronostiques + (5)	0,00	0,01	0,03	0,17	0,20	0,18	0,71	0,96	0,68	0,71	0,95

Légende. 1cv and 2cv: simple et double validation croisée, Anly: approche analytique, Boot: approche non paramétrique par bootstrap. Quantités moyennes basées sur 250 réplications.

Annexe A14 : Chapitre 5. Biais et variabilité de l'estimation ponctuelle de la probabilité de survie à 5 ans et taux de couverture des intervalles de confiance pour les différentes approches à partir de coefficients non pénalisés estimés par la pénalisation lasso adaptatif ($S_0 \approx 50\%$)

	Estimation ponctuelle de la probabilité de survie à 5 ans						**Intervalle de confiance à 95% de la prédiction**				
	Biais moyen			Erreur standard			Taux de couverture empirique				
	Point	Spline		Point	Spline		Point		Spline		
	Anly.1cv	Anly.1cv	Anly.2cv	Anly.1cv	Anly.1cv	Anly.2cv	Anly.1cv	Boot	Anly.1cv	Anly.2cv	Boot
Coefficients de régression pénalisés											
(**1**) Aucun effet	-0,01	-0,01	-0,01	0,09	0,08	0,06	0,93	0,98	0,94	0,99	1,00
(**2**) Effet du traitement seul	-0,01	-0,01	-0,01	0,08	0,07	0,06	0,95	0,98	0,96	0,99	1,00
(**3**) 20 biomarqueurs pronostiques	-0,01	-0,01	-0,01	0,12	0,13	0,13	0,92	0,97	0,89	0,92	0,99
(**4**) 15 biomarqueurs prédictifs	-0,02	-0,02	-0,02	0,16	0,21	0,19	0,91	0,97	0,85	0,89	0,94
(**5**) Effet du traitement + (4)	-0,02	-0,02	-0,02	0,16	0,20	0,19	0,90	0,97	0,85	0,89	0,94
(**6**) 20 biomarqueurs pronostiques + (5)	-0,02	-0,02	-0,02	0,16	0,20	0,19	0,90	0,96	0,84	0,87	0,91

Légende. 1cv and 2cv: simple et double validation croisée, Anly: approche analytique, Boot: approche non paramétrique par bootstrap. Quantités moyennes basées sur 250 réplications.

Annexe A15 : Chapitre 5. Précision des modèles sélectionnés à partir du lasso-*cvl* (score de Brier, statistique de concordance)

	Brier score intégré (iBrier)				**Statistique de concordance (C)**				**Δ statistique de concordance (ΔC)**			
	Apprentissage		Validation	Vrai	Apprentissage		Validation	Vrai	Apprentissage		Validation	Vrai
	1cv	2cv	1cv	modèle†	1cv	2cv	1cv	modèle†	1cv	2cv	1cv	modèle†
Coefficients de régression pénalisés												
(**1**) Aucun effet	0,097	0,098	0,098	0,098	0,535	0,495	0,498	0,500	0,003	-0,001	0,000	0,000
(**2**) Effet du traitement seul	0,099	0,100	0,100	0,100	0,592	0,584	0,575	0,558	0,005	0,000	0,000	0,000
(**3**) 20 biomarqueurs pronostiques	0,098	0,104	0,104	0,102	0,717	0,631	0,640	0,665	0,087	-0,005	0,000	0,000
(**4**) 15 biomarqueurs prédictifs	0,099	0,105	0,104	0,101	0,667	0,545	0,564	0,641	0,245	0,139	0,176	0,283
(**5**) Effet du traitement + (4)	0,099	0,106	0,105	0,102	0,692	0,611	0,621	0,675	0,256	0,133	0,179	0,284
(**6**) 20 biomarqueurs pronostiques + (5)	0,097	0,110	0,108	0,104	0,766	0,670	0,680	0,718	0,302	0,180	0,209	0,266

Légende. † modèle contenant uniquement les variables ayant un vrai effet. Quantités moyennes basées sur 250 réplications.

Annexe A16 : Chapitre 5. Biais et variabilité de l'estimation ponctuelle de la probabilité de survie à 5 ans et taux de couverture des intervalles de confiance pour les différentes approches à partir de coefficients pénalisés estimés par la pénalisation lasso-*cvl*

	Estimation ponctuelle de la probabilité de survie à 5 ans						**Intervalle de confiance à 95% de la prédiction**				
	Biais moyen			Erreur standard			Taux de couverture empirique				
	Point	Spline		Point	Spline		Point		Spline		
	Anly.1cv	Anly.1cv	Anly.2cv	Anly.1cv	Anly.1cv	Anly.2cv	Anly.1cv	Boot	Anly.1cv	Anly.2cv	Boot
Coefficients de régression pénalisés											
(**1**) Aucun effet	0,00	0,00	0,00	0,01	0,01	0,01	0,97	0,97	0,98	0,99	1,00
(**2**) Effet du traitement seul	0,00	0,00	0,00	0,01	0,01	0,01	0,95	0,97	0,95	0,99	1,00
(**3**) 20 biomarqueurs pronostiques	0,00	0,00	0,00	0,08	0,09	0,09	0,95	0,96	0,93	0,94	0,98
(**4**) 15 biomarqueurs prédictifs	0,00	0,00	0,00	0,10	0,11	0,11	0,75	0,95	0,72	0,69	0,92
(**5**) Effet du traitement + (4)	0,00	0,00	0,00	0,10	0,12	0,12	0,78	0,95	0,75	0,67	0,92
(**6**) 20 biomarqueurs pronostiques + (5)	-0,01	-0,01	0,00	0,12	0,15	0,15	0,92	0,94	0,88	0,89	0,89

Légende. 1cv and 2cv: simple et double validation croisée, Anly: approche analytique, Boot: approche non paramétrique par bootstrap. Quantités moyennes basées sur 250 réplications.

Annexe A17 : Chapitre 5. Précision des modèles sélectionnés à partir du lasso-*pcvl* (score de Brier, statistique de concordance)

	Brier score intégré (iBrier)				**Statistique de concordance (C)**				**Δ statistique de concordance (ΔC)**			
	Apprentissage		Validation	Vrai	Apprentissage		Validation	Vrai	Apprentissage		Validation	Vrai
	1cv	2cv	1cv	modèle†	1cv	2cv	1cv	modèle†	1cv	2cv	1cv	modèle†
Coefficients de régression pénalisés												
(**1**) Aucun effet	0,097	0,098	0,098	0,098	0,533	0,494	0,498	0,500	0,001	-0,001	0,000	0,000
(**2**) Effet du traitement seul	0,100	0,100	0,100	0,100	0,591	0,584	0,576	0,558	0,002	0,000	0,000	0,000
(**3**) 20 biomarqueurs pronostiques	0,102	0,105	0,105	0,102	0,689	0,627	0,636	0,665	0,003	-0,001	0,000	0,000
(**4**) 15 biomarqueurs prédictifs	0,101	0,105	0,104	0,101	0,645	0,540	0,559	0,641	0,214	0,124	0,157	0,283
(**5**) Effet du traitement + (4)	0,102	0,106	0,106	0,102	0,671	0,608	0,619	0,675	0,220	0,116	0,158	0,284
(**6**) 20 biomarqueurs pronostiques + (5)	0,106	0,111	0,110	0,104	0,715	0,655	0,665	0,718	0,191	0,112	0,144	0,266
Coefficients de régression non pénalisés												
(**1**) Aucun effet	0,097	0,099	0,099	0,098	0,539	0,496	0,498	0,500	0,001	-0,001	0,000	0,000
(**2**) Effet du traitement seul	0,099	0,101	0,100	0,100	0,593	0,582	0,572	0,558	0,002	0,000	0,000	0,000
(**3**) 20 biomarqueurs pronostiques	0,098	0,107	0,106	0,102	0,702	0,626	0,635	0,665	0,003	-0,001	0,000	0,000
(**4**) 15 biomarqueurs prédictifs	0,097	0,109	0,107	0,101	0,660	0,543	0,561	0,641	0,231	0,131	0,166	0,283
(**5**) Effet du traitement + (4)	0,098	0,110	0,109	0,102	0,695	0,599	0,609	0,675	0,239	0,123	0,165	0,284
(**6**) 20 biomarqueurs pronostiques + (5)	0,100	0,114	0,112	0,104	0,741	0,655	0,666	0,718	0,206	0,123	0,152	0,266

Légende. † modèle contenant uniquement les variables ayant un vrai effet. Quantités moyennes basées sur 250 réplications.

Annexe A18 : Chapitre 5. Biais et variabilité de l'estimation ponctuelle de la probabilité de survie à 5 ans et taux de couverture des intervalles de confiance pour les différentes approches à partir de coefficients estimés par la pénalisation lasso-*pcvl*

	Estimation ponctuelle de la probabilité de survie à 5 ans						**Intervalle de confiance à 95% de la prédiction**				
	Biais moyen			Erreur standard			Taux de couverture empirique				
	Point	Spline		Point	Spline		Point		Spline		
	Anly.1cv	Anly.1cv	Anly.2cv	Anly.1cv	Anly.1cv	Anly.2cv	Anly.1cv	Boot	Anly.1cv	Anly.2cv	Boot
Coefficients de régression pénalisés											
(**1**) Aucun effet	0,00	0,00	0,00	0,04	0,03	0,02	0,94	0,98	0,95	0,99	1,00
(**2**) Effet du traitement seul	0,00	0,00	0,00	0,04	0,04	0,03	0,93	0,98	0,94	0,99	1,00
(**3**) 20 biomarqueurs pronostiques	0,00	0,00	0,00	0,08	0,09	0,09	0,73	0,95	0,70	0,75	0,91
(**4**) 15 biomarqueurs prédictifs	0,00	0,00	0,00	0,10	0,11	0,11	0,79	0,95	0,73	0,79	0,90
(**5**) Effet du traitement + (4)	0,00	0,00	0,00	0,10	0,13	0,12	0,78	0,94	0,72	0,78	0,89
(**6**) 20 biomarqueurs pronostiques + (5)	0,00	0,00	0,00	0,13	0,15	0,15	0,64	0,86	0,60	0,64	0,78
Coefficients de régression non pénalisés											
(**1**) Aucun effet	0,00	0,00	0,00	0,08	0,07	0,04	0,65	0,98	0,69	0,87	1,00
(**2**) Effet du traitement seul	0,00	0,00	0,00	0,08	0,07	0,05	0,65	0,98	0,67	0,87	1,00
(**3**) 20 biomarqueurs pronostiques	0,00	0,00	0,00	0,09	0,10	0,09	0,67	0,98	0,64	0,73	0,99
(**4**) 15 biomarqueurs prédictifs	0,00	0,01	0,02	0,13	0,15	0,13	0,63	0,98	0,57	0,67	0,96
(**5**) Effet du traitement + (4)	0,00	0,01	0,02	0,13	0,16	0,14	0,63	0,98	0,57	0,67	0,96
(**6**) 20 biomarqueurs pronostiques + (5)	0,00	0,01	0,01	0,14	0,17	0,16	0,59	0,98	0,54	0,61	0,94

Légende. 1cv and 2cv: simple et double validation croisée, Anly: approche analytique, Boot: approche non paramétrique par bootstrap. Quantités moyennes basées sur 250 réplications.

B Implémentation

Annexe B19 : Choix du package pour la pénalisation lasso

Actuellement, de nombreux packages `R` sont proposés pour implémenter les régressions pénalisées et peuvent être listés à partir du *CRAN Task View* « *Machine Learning* ». Dans le cadre de la pénalisation lasso appliquée à des données de survie, trois packages semblent être couramment utilisés: `glmpath` (Park et Hastie, 2015), `glmnet` (Friedman et al., 2010, 2016) et `penalized` (Goeman, Meijer et Chaturvedi, 2016).

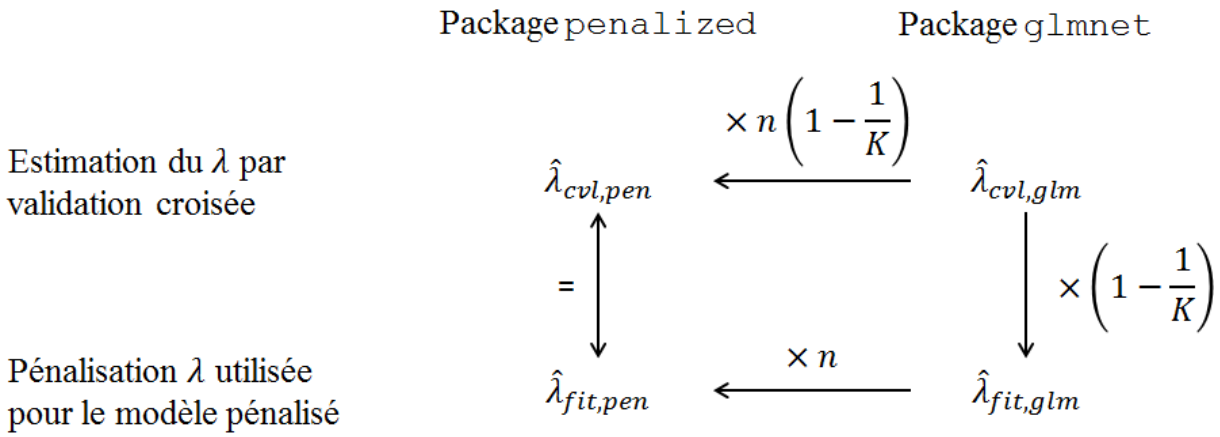
Tableau : Temps de calcul moyen pour estimer $\hat{\lambda}_{cvl}$ avec $K = 5$ sous échantillons

Matrice $n \times p$	**`glmpath`**	**`glmnet`**	**`penalized`**
500 × 10	36	0,09	9,8
1000 × 10	372	0,13	43
500 × 100	477	0,30	13
1000 × 100	4110	0,42	63
500 × 1000	1223	22	25

Légende. temps en secondes, ordinateur : Processeur 3,4GHz et RAM 16,3Mb.

Les packages `glmnet` et `penalized` sont actuellement mieux documentés que le package `glmpath` grâce notamment à la présentation d'une vignette. De plus, le package `glmpath` a des temps de calcul extrêmement important en comparaison aux deux autres packages lorsqu'on l'utilise avec ces paramètres par défaut (voir Tableau ci-dessus). En raison de la clarté de sa documentation et de sa rapidité de calcul, nous avons choisi d'utiliser le package `glmnet`.

Annexe B20 : Paramétrisation de la pénalisation des packages `glmnet` et `penalized`



Légende. n : taille de l'échantillon, K : nombre de sous-échantillons pour la validation croisée.

Annexe B21 : Transformation de la base de données pour le modèle de Poisson

La transformation de la base de données a été réalisée telle que décrite par Whitehead (1980). Le Tableau ci-dessous montre un exemple de modification de bases. La Figure ci-dessous montre l'estimation de la probabilité de survie au cours du temps entre ces deux techniques et pour différentes tailles d'intervalle.

Tableau : Exemple de transformation d'une base de données

Base initiale (pour Cox)

time	*status*	*treat*
0,57	1	+0,5
1,55	0	+0,5
2,65	1	-0,5
1,31	0	-0,5
4,42	0	-0,5

Base modifiée (pour Poisson)

int	*treat*	*m*	*Rt*	*N*
0	-0,5	61	33,90	242
0	+0,5	95	31,94	258
0,166	-0,5	31	27,33	181
0,166	+0,5	18	25,65	163
0,333	-0,5	11	24,12	150

Légende. *int* : intervalle, *m* : nombre d'événements, *Rt* : personne-années, *N* : nombre de sujets à risque.

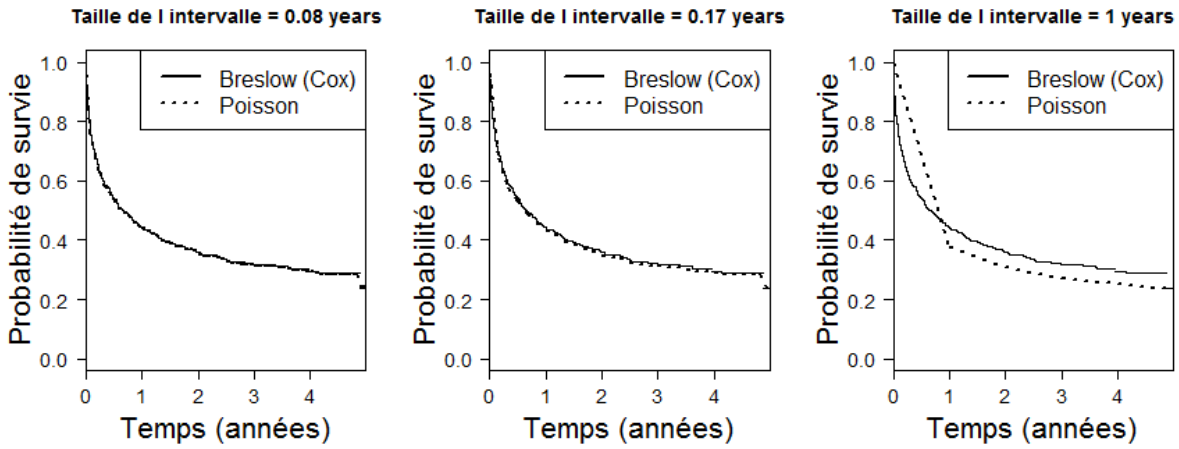


Figure : Illustration des courbes de survie avec les deux modèles (Cox et Poisson) pour différentes tailles d'intervalle de temps