Exploitation conjointe de plusieurs bases de données dans la détection de signaux en pharmacovigilance via l'utilisation du lasso pondéré

Émeline Courtois [1] & Ismaïl Ahmed [2] & Pascale Tubert-Bitter [3]

[1] ***UMR 1181 - Inserm : Biostatistique, Biomathématique, Pharmaco-Épidémiologie et Maladies Infectieuses***
Hôpital Paul Brousse, 16 avenue Paul Vaillant-Couturier, Villejuif
emeline.courtois@inserm.fr
[2] ***UMR 1181 - Inserm : Biostatistique, Biomathématique, Pharmaco-Épidémiologie et Maladies Infectieuses***
Hôpital Paul Brousse, 16 avenue Paul Vaillant-Couturier, Villejuif
ismail.ahmed@inserm.fr
[3] ***UMR 1181 - Inserm : Biostatistique, Biomathématique, Pharmaco-Épidémiologie et Maladies Infectieuses***
Hôpital Paul Brousse, 16 avenue Paul Vaillant-Couturier, Villejuif
pascale.tubert@inserm.fr

Résumé. La pharmacovigilance a pour objectif de détecter le plus précocement possible les effets indésirables des médicaments commercialisés. Ce travail de détection s'apparente à une problématique de sélection de variables en grande dimension. Classiquement il s'effectue sur les bases de notifications spontanées, mais récemment un intérêt croissant s'est porté sur l'exploitation des bases médico-administratives. Nous proposons dans ce travail d'intégrer l'information issue d'une stratégie de détection réalisée à partir d'un référentiel de témoins fournis par les bases médico-administratives, dans les modèles d'analyses de la base des notifications spontanées. Cette intégration de l'information est effectuée via l'utilisation de pénalités différentiées pour chaque covariable médicament dans un lasso pondéré afin de guider la sélection de variables opérée. Ces pénalités différentiées sont obtenues à partir de deux types d'informations : des odds ratio et des p-valeurs corrigées pour les tests multiples. Les performances des méthodes basées sur le lasso pondéré sont comparées à celle d'un lasso classique et sont évaluées empiriquement grâce à un ensemble de signaux de référence concernant l'évènement indésirable "lésion hépatique aiguë".

Mots-clés. détection de signal, pharmacovigilance, intégration de données, grande dimension, lasso pondéré

Abstract. The objective of pharmacovigilance is to detect adverse reactions of marketed drugs as early as possible. This detection process is a variable selection issue. Usually it relies on spontaneous reports databases, but recently there has been a growing interest

in the use of medico-administrative databases. In this work, we propose to integrate information from a detection strategy based on controls provided by medico-administrative databases into the analysis models of the spontaneous reports database. This integration of information is achieved through the use of different penalties for each drug covariate in a weighted lasso to guide the covariates selection. These penalties are obtained from two types of information : odds ratios and p-values corrected for multiple testing. The performance of weighted lasso methods is compared to that of a conventional lasso and is empirically evaluated using a set of reference signals for the adverse event "acute liver injury".

Keywords. signal detection, pharmacovigilance, multiple data integration, high dimensional data, weighted lasso

1 Introduction

La pharmacovigilance a pour objectif de détecter le plus précocement possible les effets indésirables des médicaments mis sur le marché. Elle repose le plus souvent sur l'exploitation des bases de notifications spontanées, qui sont constituées de déclarations de la survenue d'évènements indésirables (EIs) par les praticiens de santé, et dont l'origine suspectée est médicamenteuse. A l'échelle nationale, ces notifications spontanées représentent de grands ensemble de données : sur la période 2000-2016 la base française de notifications spontanées comptabilise environ 380 000 notifications, qui comprennent 5 900 EI et 2 300 médicaments différents. Il a donc été proposé depuis une quinzaine d'années un certain nombre de méthodes statistiques de fouille de données visant à détecter des associations statistiques suspectes entre médicaments et EIs. On parle de détection automatisée de signaux, les signaux statistiques générés devant être évalués par des experts. Les méthodes classiques de détection de signaux sont les méthodes de disproportionnalités qui consistent, pour chaque couple (médicament p, EI j), à (1) déterminer le nombre attendu de notifications qui mentionnent le médicament p et l'EI j si la la mention de l'EI j est indépendante de la mention du médicament p ; (2) construire une mesure de "disproportionnalité" pour quantifier l'excès du nombre observé n_{pj} face au nombre attendu de notifications mentionnant le médicament p et l'EI j. Finalement, un signal est généré si cette mesure dépasse un certain seuil critique.

Plus récemment, des approches de détection de signaux basées sur des régressions logistiques multiples ont été proposées [1, 2]. Au lieu de considérer les données agrégées comme les méthodes de disproportionnalité, ces nouvelles approches sont directement appliquées aux notifications spontanées : l'observation devient la notification individuelle, la réponse est la présence ou l'absence d'un EI donné, et les covariables sont toutes des indicatrices de présence de médicaments. En raison du très grand nombre de covariables potentielles, ces méthodes reposent sur des versions pénalisées de régressions logistiques multiples.

Pour toutes ces approches, les expositions médicamenteuses des cas d'un EI donné sont comparées à celles d'un groupe témoin constitué par l'ensemble des individus notifiés pour un autre EI que celui d'intérêt. En effet, par construction, on ne dispose pas dans les notifications spontanées d'individus témoins représentatifs de la population générale.

En outre, depuis plusieurs années un intérêt croissant s'est porté sur l'exploitation des bases médico-administratives pour la détection de signaux en pharmacovigilance. En France, le SNDS (Système National des Données de Santé) regroupe les données des bases hospitalières et de remboursement de soins, et couvre la quasi-totalité de la population française. Il a également été créé à partir du SNDS l'Échantillon Généraliste des Bénéficiaires (EGB), échantillon au $1/97^{\text{ème}}$. La difficulté dans l'exploitation de ces grandes bases de données tient au fait qu'elles n'ont pas été conçues pour répondre à des questions de santé. Par ailleurs, les EIs pouvant être étudiés sont nécessairement graves puisque requérant une hospitalisation. De plus, malgré une taille très importante, l'exploitation de l'EGB (plus facilement accessible à la communauté des chercheurs) n'est réalisable que pour des EIs relativement fréquents.

L'objectif général de ce travail est de proposer et d'évaluer des stratégies de détection de signaux qui intègrent l'information sur l'exposition médicamenteuse de témoins issus des données de l'EGB. Cette intégration d'information est rendue possible par l'utilisation du lasso pondéré [3], dont le lasso [4] peut être vu comme un cas particulier, où l'on tient compte de pénalités individuelles pour chaque covariable. Ici, ces pénalités individuelles sont déterminées à partir des résultats d'un travail de détection de signal dans une base dite hybride, composée des cas des notifications spontanées et des témoins de l'EGB. Cette méthode d'intégration de données est comparée à une méthode basée sur un lasso classique dans la base des notifications spontanées seule, dénommée ci-après base principale. Une évaluation empirique avec une application à l'évènement indésirable "lésions hépatiques aiguës" (ALI) est réalisée, en utilisant l'ensemble de signaux de référence établi par l'Observational Medical Outcome Partnership (OMOP) [5].

2 Méthode

2.1 Lasso et lasso pondéré

En considérant un EI donné, on introduit les notations suivantes. Soit N le nombre de notifications (i.e. d'observations) et P le nombre de covariables médicaments. On note Y la variable réponse, indicatrice de la présence ou de l'absence de l'EI considéré et X_p la variable binaire de l'exposition au médicament p. Pour $i \in \{1, ..., N\}$, le modèle logistique pour l'EI considéré est défini par :

$$\text{logit}(\Pr(Y_i = 1)) = \beta_0 + \sum_{p=1}^{P} \beta_p X_{ip}. \tag{1}$$

Le nombre de covariables P étant très grand, on a recours à la régression pénalisée de type lasso, qui permet de forcer certains coefficients β_p à valoir exactement zéro. Ce type de régression consiste à maximiser la vraisemblance du modèle (1) moins un terme de pénalité défini par

$$pen(\lambda) = \lambda \sum_{p=1}^{P} |\beta_p|. \tag{2}$$

Le lasso pondéré est une généralisation de la pénalisation définie en (2) qui permet d'introduire pour chaque covariable une valeur de pénalité différente. Le terme de pénalité s'écrit dans ce cas

$$pen(\lambda) = \lambda \sum_{p=1}^{P} w_p |\beta_p|. \tag{3}$$

Au final, la pénalité appliquée à la covariable p est définie par $\lambda_p = \lambda w_p$. L'intérêt d'une telle différentiation de pénalité est de guider la sélection de variables opérée par le lasso : plus la valeur du coefficient w_p est grande, plus la variable p est pénalisée, moins la variable est susceptible d'être incluse dans le modèle. Dans la suite, on parlera de poids pour désigner le coefficient w_p et de pondération pour définir une stratégie visant à déterminer ces poids.

2.2 Pondérations proposées

Suivant l'idée de Bergensen *et al.* [3], nous proposons ici deux pondérations différentes qui permettent de prendre en compte l'information provenant d'une base de données externe. On suppose qu'à partir de cette source d'information externe, on dispose pour chaque covariable médicament p du résultat d'un test d'association entre cette covariable et la réponse. A l'issue de cet ensemble de tests, chaque covariable a une mesure d'association ainsi qu'une p-valeur associée. Comme on s'intéresse ici uniquement à des associations délétères, les test réalisés sont unilatéraux et la mesure d'association, ici l'odds ratio (OR), est jugée pertinente si elle excède 1. Pour un seuil de significativité α, seul un sous ensemble de ces tests sont significatifs. Les pondérations que nous proposons dans la suite prennent en compte cette sélection de variables.

Poids 1 : Odds-Ratio A partir des odds ratios des covariables significativement associées à la réponse au seuil α, nous proposons d'introduire les poids suivant dans le lasso pondéré.

$$w_p^{(1)} = \frac{1}{\log(\mathrm{OR}_p)}. \tag{4}$$

Ainsi, plus une covariable est associée fortement à la réponse, moins elle est pénalisée.

Poids 2 : P-valeurs Soit s le nombre de covariables qui ont une p-valeur associée inférieure à α. Nous proposons d'attribuer à ces covariables un poids dans le lasso pondéré déterminé à partir de la fonction de répartition empirique de ces p-valeurs

$$w_p^{(2)} = \frac{1}{s} \sum_{i=1}^{s} \mathbb{1}_{\text{p-valeur}_i \leq \text{p-valeur}_p}. \tag{5}$$

Avec cette transformation monotone, les covariables avec les p-valeurs les plus faibles sont très favorisées dans le processus de sélection.

Pour ces deux pondérations, les covariables non significativement associées à la réponse d'après la source externe sont exclues des covariables candidates à la sélection dans le lasso pondéré effectué dans la base principale. Les poids associés aux variables significatives sont normalisés, de telle sorte à ce que leur moyenne soit égale à 1 pour chacune des pondérations.

3 Application

3.1 Matériel

La première étape de ce travail a consisté à construire pour les cas de ALI des notifications spontanées un groupe de contrôles issus de l'EGB. Pour des raisons de disponibilité des données d'hospitalisation dans l'EGB, nous avons considéré la période 2006-2016.

L'évènement ALI est défini selon un ensemble de *Preferred Term*(PT) de la terminologie MedDRA (Medical Dictionary for Regulatory Activities) qui est utilisé pour coder les EIs dans les notifications spontanées. Un cas de ALI est un individu dont la notification fait mention d'au moins un des ces termes. Sur la période considérée, et pour les notifications où l'âge et le sexe sont renseignés, les notifications spontanées comprennent 11 618 cas et 240 500 non cas de ALI.

Les témoins de l'EGB ont été appariés aux cas des notifications spontanées sur le sexe et par classes d'âge de 10 ans. Sont considérés comme témoins les individus qui n'ont pas été hospitalisés pour un quelconque problème hépatique l'année de leur appariement ou les années antérieures. Nous avons cherché à apparier jusqu'à 100 témoins par cas, tout en interdisant le fait qu'un témoin soit apparié avec plusieurs cas. Au final, nous obtenons 525 506 témoins uniques de l'EGB. Les consommations médicamenteuses des témoins ont été regardées dans le mois précédant la date de l'EI du cas auquel ils sont appariés. En se restreignant aux expositions médicamenteuses communes aux cas et aux témoins des deux différentes bases, et qui sont notifiées plus de 10 fois dans la base hybride, nous considérons 1 020 covariables médicaments.

3.2 Calcul des poids et analyses

Sur la base hybride, nous avons implémenté une méthode de détection de signal proche des méthodes de disproportionnalité classiques : le Reporting Fisher Exact Test (RFET) [6]. Cette approche, de par l'utilisation du test exact de Fisher, est robuste aux situations où les effectifs observés de cas exposés au médicament p sont faibles. Ainsi pour chaque covariable médicament, on obtient avec cette procédure un OR ainsi qu'une p-valeur. Pour prendre en compte la multiplicité des comparaisons effectuées, nous corrigeons ces p-valeurs avec la procédure de correction de tests multiple proposée par Ahmed *et al.* [6], qui repose sur le contrôle du Fase Discovery Rate (FDR) et qui est adaptée aux tests d'hypothèses unilatérales effectués ici. Cette approche amène de fait à une sélection de variables, puisque sont considérés comme signaux dans la base hybride toutes les covariables avec une p-valeur corrigée inférieure à un seuil de significativité $\alpha = 0.05$.

A partir des deux types de pondération présentées dans la partie 2.2, nous avons appliqué dans la base principale les deux lasso pondérés correspondants ainsi qu'une régression lasso classique. Afin de s'affranchir du choix d'une pénalité optimale, nous avons comparé les résultats de ces trois régressions pénalisées à nombre de signaux générés équivalents, pour une grille de valeurs de pénalités. Avec ces méthodes, ce que l'on définit comme un signal est une covariable qui a un coefficient pénalisé strictement supérieur à zéro.

3.3 Paramétrage de la comparaison

Les performances des différentes approches de détection ont été évaluées grâce à un ensemble de signaux de référence concernant l'EI ALI défini par l'OMOP. Parmi les covariables médicament considérées, cet ensemble comprend 52 témoins positifs et 13 témoins négatifs.

4 Résultats

Le figure (1-(a)) présente la répartition des signaux générés par le lasso pondéré basé sur les OR issus de la base hybride et le lasso classique. Le détail de cette répartition est aussi présentée pour les témoins positifs (figure 1-(b)) et les témoins négatifs détectés (figure 2-(c)), et les performances de ces deux méthodes en termes de sensibilité et de proportion de fausses découvertes (FDP) sont représentées (figure 2-(d)). Les figures (3) et (4) présentent les mêmes résultats pour le lasso pondéré basé sur les p-valeurs issues de la base hybride, comparé au lasso classique.

Les performances entre les méthodes de détection basées sur le lasso pondéré et celle basée sur la lasso classique sont proches. Néanmoins les approches proposées ont une meilleure sensibilité, et tout particulièrement pour un nombre de signaux générés importants. Le lasso pondéré basé sur les OR génère son premier et unique faux positif à un

nombre de signaux plus grand que la lasso classique, et le lasso pondéré basé sur les p-valeurs ne génère pas de faux négatif. Les variables sélectionnées par les lasso pondérés sont différentes de celles sélectionnées avec le lasso classique.

5 Discussion

Les méthodes basées sur le lasso pondéré ont dans l'ensemble des performances comparables en terme de détection par rapport à une méthode basée sur une régression lasso classique. Néanmoins, les signaux générés par ces méthodes sont différents, ce qui laisse penser que l'ajout d'une information extérieure dans le processus de détection de signaux apporte une information différente voire complémentaire à celle issue d'un processus de détection plus classique.

Avec les pondérations que nous proposons, les covariables qui ne sont pas jugées significatives à l'issue d'un test basé sur une source d'information extérieure sont exclues de la sélection dans le lasso pondéré. Ce choix étant potentiellement assez fort, il convient de préciser que l'information externe est pensée comme étant une information pertinente Néanmoins, pour relâcher cette contrainte, on peut envisager de considérer une valeur de α plus grande.

Une limite avec l'utilisation du lasso pondéré dans un contexte de sélection de variable réside dans la choix optimal de la pénalité. Une solution serait d'avoir recours à la validation croisée pour déterminer le paramètre λ dans (3), bien que la pénalité définie ainsi ne soit pas la plus adaptée dans le cadre de la sélection de variable.

Références

[1] Ola Caster, G. Niklas Norén, David Madigan, and Andrew Bate. Large-scale regression-based pattern discovery : The example of screening the WHO global drug safety database. *Statistical Analysis and Data Mining*, 3(4) :197–208, 2010.

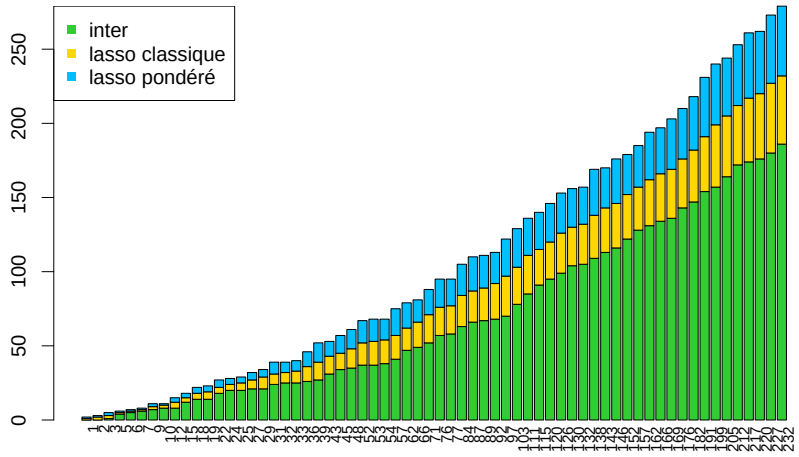
[2] Ismaïl Ahmed, Antoine Pariente, and Pascale Tubert-Bitter. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Statistical Methods in Medical Research*, 27(3) :785–797, 2018.

[3] Linn Cecilie Bergersen, Ingrid K. Glad, and Heidi Lyng. Weighted lasso with data integration. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011.

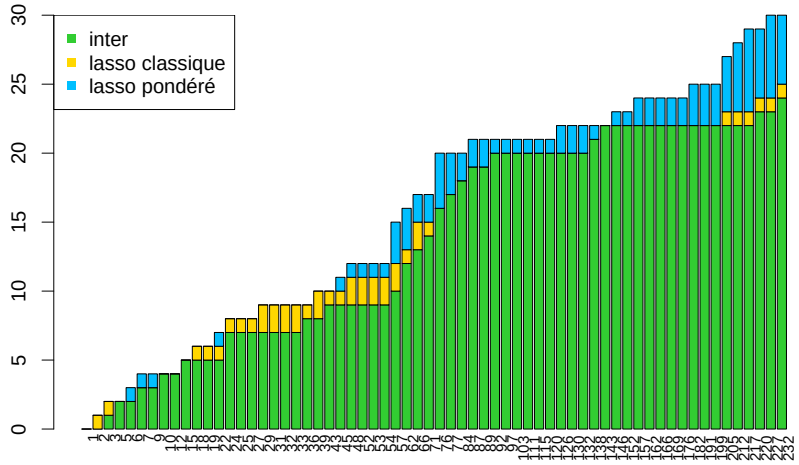
[4] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1) :267–288, 1996.

[5] Patrick B. Ryan, Martijn J. Schuemie, Emily Welebob, Jon Duke, Sarah Valentine, and Abraham G. Hartzema. Defining a Reference Set to Support Methodological Research in Drug Safety. *Drug Safety*, 36(Suppl 1), 2013.

[6] I. Ahmed, C. Dalmasso, F. Haramburu, F. Thiessard, P. Broët, and P. Tubert-Bitter. False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics*, 2010.

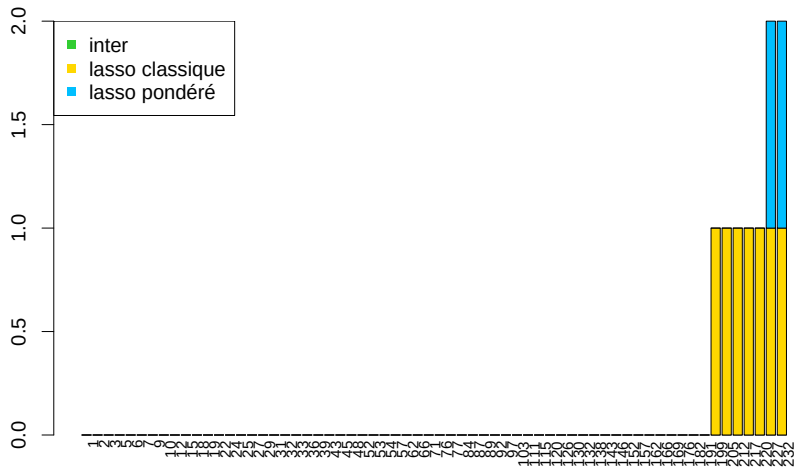


(a)

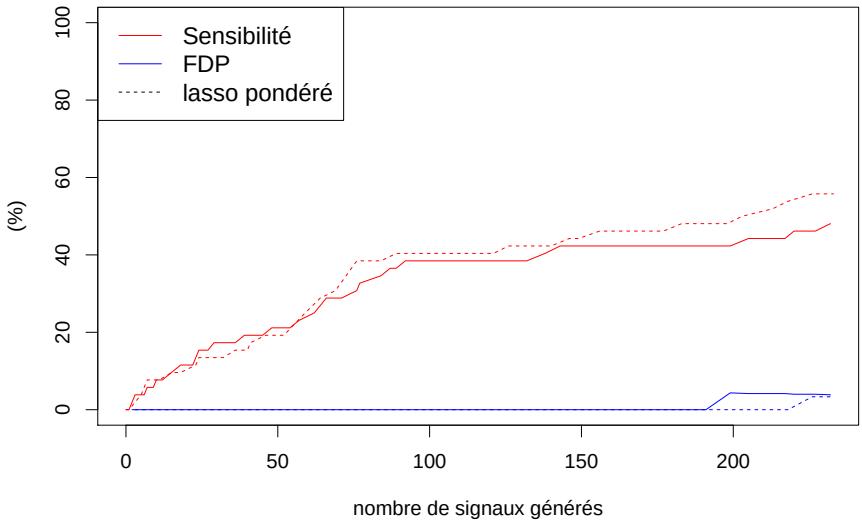


(b)

FIGURE 1 – Comparaison des performances entre le lasso et le lasso pondéré basé sur les **OR de la base hybride**. La figure (a) représente le nombre de signaux détectés par les deux méthodes (vert) ainsi que les signaux détectés uniquement par l'une des méthodes (jaune/bleu) en fonction du nombre de signaux générés. De manière analogue, la figure (b) présente le nombre de témoins positifs détectés par les deux méthodes.

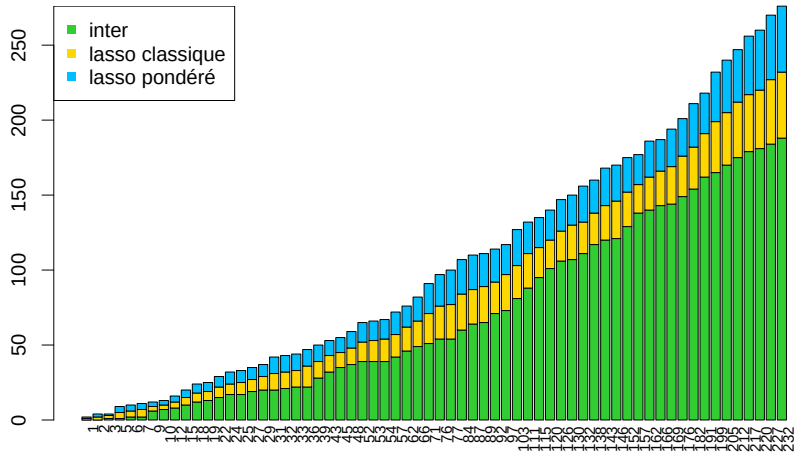


(c)

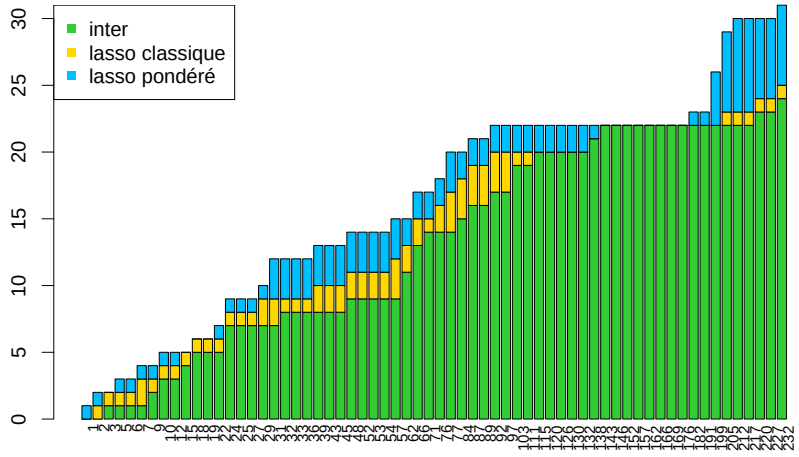


(d)

FIGURE 2 – Comparaison des performances entre le lasso et le lasso pondéré basé sur les **OR de la base hybride**. La figure (c) représente le nombre de témoins négatifs détectés par les deux méthodes (vert) ainsi que les témoins négatifs détectés uniquement par l'une des méthodes (jaune/bleu) en fonction du nombre de signaux générés. La figure (d) présente les résultats des deux approches en terme de sensibilité (%) et de proportion de fausses découvertes (FDP, %) pour les signaux à statut connu, en fonction du nombre de signaux générés.

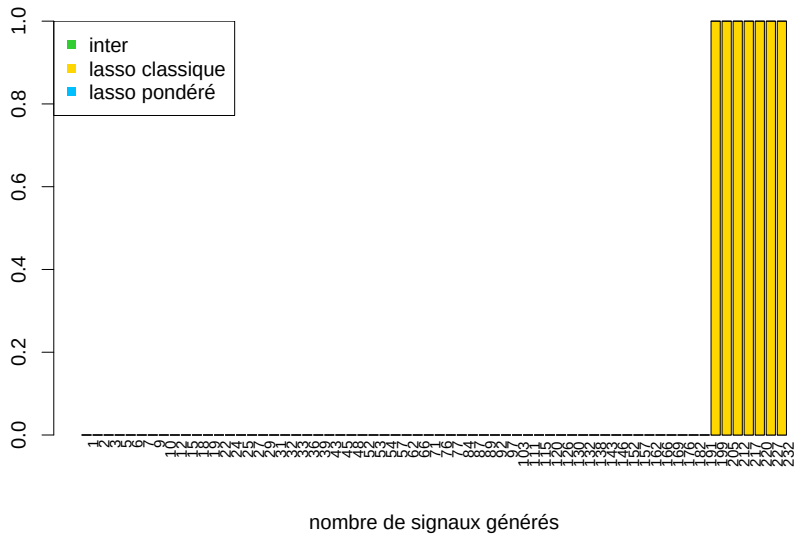


(a)

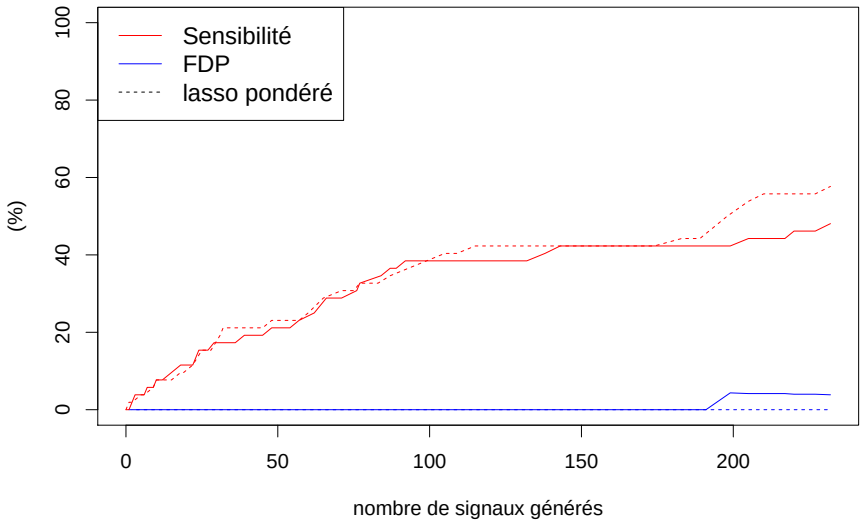


(b)

FIGURE 3 – Comparaison des performances entre le lasso et le lasso pondéré basé sur les **p-valeurs de la base hybride**. La figure (a) représente le nombre de signaux détectés par les deux méthodes (vert) ainsi que les signaux détectés uniquement par l'une des méthodes (jaune/bleu) en fonction du nombre de signaux générés. De manière analogue, la figure (b) présente le nombre de témoins positifs détectés par les deux méthodes.



(c)



(d)

FIGURE 4 – Comparaison des performances entre le lasso et le lasso pondéré basé sur les **p-valeurs de la base hybride**. La figure (c) représente le nombre de témoins négatifs détectés par les deux méthodes (vert) ainsi que les témoins négatifs détectés uniquement par l'une des méthodes (jaune/bleu) en fonction du nombre de signaux générés. La figure (d) présente les résultats des deux approches en terme de sensibilité (%) et de proportion de fausses découvertes (FDP, %) pour les signaux à statut connu, en fonction du nombre de signaux générés.