Sélection de variables en grande dimension par le Lasso et tests statistiques - application à la pharmacovigilance

Matthieu Pluntz

université PARIS-SACLAY

Sélection de variables en grande dimension par le Lasso et tests statistiques – application à la pharmacovigilance

High-dimensional variable selection with the Lasso and statistical testing – application to pharmacovigilance

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 570 : Santé publique (EDSP)
Spécialité de doctorat : Biostatistiques et data sciences
Graduate School : Santé publique
Référent : Université de Versailles – Saint-Quentin-en-Yvelines (UVSQ)

Thèse préparée dans l'unité de recherche **CESP (Université Paris-Saclay, UVSQ, Inserm)**, sous la direction de **Pascale TUBERT-BITTER**, directeur de recherche, et le co-encadrement de **Cyril DALMASSO**, maître de conférences, et d'**Ismaïl AHMED**, chargé de recherche

Thèse soutenue à Villejuif, le 18 mars 2024, par

Matthieu PLUNTZ

Composition du jury

Membres du jury avec voix délibérative

Chantal GUIHENNEUC Professeur des universités, Université Paris-Cité	Présidente
David CAUSEUR Professeur, Institut Agro Rennes-Angers	Rapporteur & Examinateur
Vivian VIALLON Maître de conférences, HDR, Université Claude Bernard Lyon 1	Rapporteur & Examinateur
Christophe GIRAUD Professeur des universités, Université Paris-Saclay	Examinateur

THESE DE DOCTORAT

NNT : 2024UPASR002

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre : Sélection de variables en grande dimension par le Lasso et tests statistiques – application à la pharmacovigilance

Mots clés : Sélection de variables, Lasso, critère d'information, p-value empirique, p-value conditionnelle, tests multiples

Résumé : La sélection de variables dans une régression de grande dimension est un problème classique dans l'exploitation de données de santé, où l'on cherche à identifier un nombre limité de facteurs associés à un évènement parmi un grand nombre de variables candidates : facteurs génétiques, expositions environnementales ou médicamenteuses.

La régression Lasso (Tibshirani, 1996) fournit une suite de modèles parcimonieux où les variables apparaissent les unes après les autres suivant la valeur du paramètre de régularisation. Elle doit s'accompagner d'une procédure du choix de ce paramètre et donc du modèle associé. Nous proposons ici des procédures de sélection d'un des modèles du chemin du Lasso qui font partie, ou s'inspirent, du paradigme des tests statistiques. De la sorte, nous cherchons à contrôler le risque de sélection d'au moins un faux positif (Family-Wise Error Rate, FWER), au contraire de la plupart des méthodes existantes de post-traitement du Lasso qui acceptent plus facilement des faux positifs.

Notre première proposition est une généralisation du critère d'information d'Akaike (AIC) que nous appelons AIC étendu (EAIC). La log-vraisemblance du modèle considéré y est pénalisée par son nombre de paramètres affecté d'un poids qui est fonction du nombre total de variables candidates et du niveau visé de FWER, mais pas du nombre d'observations. Nous obtenons cette fonction en rapprochant la comparaison de critères d'information de sous-modèles emboîtés d'une régression en grande dimension, de tests multiples du rapport de vraisemblance sur lesquels nous démontrons un résultat asymptotique.

Notre deuxième proposition est un test de la significativité d'une variable apparaissant sur le chemin du Lasso. Son hypothèse nulle dépend d'un ensemble A de variables déjà sélectionnées et énonce qu'il contient toutes les variables actives. Nous cherchons à prendre comme statistique de test la valeur du paramètre de régularisation à partir de laquelle une première variable en dehors de A est sélectionnée par le Lasso. Ce choix se heurte au fait que l'hypothèse nulle n'est pas assez spécifiée pour définir la loi de cette statistique et donc sa p-value. Nous résolvons cela en lui substituant sa p-value conditionnelle, définie conditionnellement aux coefficients estimés du modèle non pénalisé restreint à A. Nous estimons celle-ci par un algorithme que nous appelons simulation-calibration, où des vecteurs réponses sont simulés puis calibrés sur les coefficients estimés du vecteur réponse observé. Nous adaptons de façon heuristique la calibration au cas des modèles linéaires généralisés (binaire et de Poisson) dans lesquels elle est une procédure itérative et stochastique. Nous prouvons que l'utilisation du test permet de contrôler le risque de sélection d'un faux positif dans les modèles linéaires, à la fois lorsque l'hypothèse nulle est vérifiée mais aussi, sous une condition de corrélation, lorsque A ne contient pas toutes les variables actives.

Nous mesurons les performances des deux procédures par des études de simulations extensives, portant à la fois sur la sélection éventuelle d'une variable sous l'hypothèse nulle (ou son équivalent pour l'EAIC) et sur la procédure globale de sélection d'un modèle. Nous observons que nos propositions se comparent de façon satisfaisante à leurs équivalents les plus proches déjà existants, BIC et ses versions étendues pour l'EAIC et le test de covariance de Lockhart et al. (2014) pour le test par simulation-calibration. Nous illustrons également les deux procédures dans la détection d'expositions médicamenteuses associées aux pathologies hépatiques (drug-induced liver injuries, DILI) dans la base nationale de pharmacovigilance (BNPV) en mesurant leurs performances grâce à l'ensemble de référence DILIrank d'associations connues.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Title : High-dimensional variable selection with the Lasso and statistical testing – application to pharmacovigilance

Keywords : Variable selection, Lasso, information criterion, empirical p-value, conditional p-value, multiple testing

Abstract : Variable selection in high-dimensional regressions is a classic problem in health data analysis. It aims to identify a limited number of factors associated with a given health event among a large number of candidate variables such as genetic factors or environmental or drug exposures.

The Lasso regression (Tibshirani, 1996) provides a series of sparse models where variables appear one after another depending on the regularization parameter's value. It requires a procedure for choosing this parameter and thus the associated model. In this thesis, we propose procedures for selecting one of the models of the Lasso path, which belong to or are inspired by the statistical testing paradigm. Thus, we aim to control the risk of selecting at least one false positive (Family-Wise Error Rate, FWER) unlike most existing post-processing methods of the Lasso, which accept false positives more easily.

Our first proposal is a generalization of the Akaike Information Criterion (AIC) which we call the Extended AIC (EAIC). We penalize the log-likelihood of the model under consideration by its number of parameters weighted by a function of the total number of candidate variables and the targeted level of FWER but not the number of observations. We obtain this function by observing the relationship between comparing the information criteria of nested sub-models of a high-dimensional regression, and performing multiple likelihood ratio test, about which we prove an asymptotic property.

Our second proposal is a test of the significance of a variable appearing on the Lasso path. Its null hypothesis depends on a set A of already selected variables and states that it contains all the active variables. As the test statistic, we aim to use the regularization parameter value from which a first variable outside A is selected by Lasso. This choice faces the fact that the null hypothesis is not specific enough to define the distribution of this statistic and thus its p-value. We solve this by replacing the statistic with its conditional p-value, which we define conditional on the non-penalized estimated coefficients of the model restricted to A. We estimate the conditional p-value with an algorithm that we call simulation-calibration, where we simulate outcome vectors and then calibrate them on the observed outcome's estimated coefficients. We adapt the calibration heuristically to the case of generalized linear models (binary and Poisson) in which it turns into an iterative and stochastic procedure. We prove that using our test controls the risk of selecting a false positive in linear models, both when the null hypothesis is verified and, under a correlation condition, when the set A does not contain all active variables.

We evaluate the performance of both procedures through extensive simulation studies, which cover both the potential selection of a variable under the null hypothesis (or its equivalent for EAIC) and on the overall model selection procedure. We observe that our proposals compare well to their closest existing counterparts, the BIC and its extended versions for the EAIC, and Lockhart et al.'s (2014) covariance test for the simulation-calibration test. We also illustrate both procedures in the detection of exposures associated with drug-induced liver injuries (DILI) in the French national pharmacovigilance database (BNPV) by measuring their performance using the DILIrank reference set of known associations.

Remerciements

Cette thèse n'aurait pas vu le jour sans le concours de beaucoup de personnes. Mes remerciements vont d'abord à mes directeurs et encadrant de thèse : Pascale Tubert-Bitter, Cyril Dalmasso et Ismaïl Ahmed, avec qui ce fut toujours un plaisir de travailler, et qui m'ont apporté beaucoup. Merci en particulier à Pascale d'avoir été assez littéralement à mes côtés lors d'une grande partie de la rédaction de ce manuscrit.

Merci aux rapporteurs MM. David Causeur et Vivian Viallon pour leurs retours très pertinents sur cette thèse. Merci également à M. Christophe Giraud et Mme Chantal Guihenneuc d'avoir acccepté de faire partie du jury.

Sur le très long terme, s'agissant d'une thèse ayant d'importants aspects mathématiques, je dois mes remerciements à mes enseignants de mathématiques puis de statistiques du collège-lycée, de prépa, de l'ENSAE et de prépa agrégation qui ont fait grandir ce qui a toujours été une passion pour moi.

Merci aux personnes avec qui j'ai travaillé à l'Anses juste avant et au début de cette thèse : Juliette Bloch, Sandra Sinno-Tellier, Serge Faye, aini que les toxicologues des CAP. Cette expérience très positive m'a encouragé à poursuivre dans la biostatistique quoique sur des aspects différents.

Merci à Raphaëlle Varraso et à Florence Menegaux de l'école doctorale Santé publique pour leur soutien et leurs encouragements en fin de thèse. Merci aux administrateurs de Wikipédia qui ont accepté de bloquer temporairement mon compte, me permettant de limiter mon investissement dans une activité enrichissante, mais chronophage.

Merci à mes collègues Ana, Anne, Anne-Louise, Astelle, Élise, Émeline, Étienne, Juliette, Lucas, Lucas, Hervé, Hong, Romain, Philippe, Sidi, Sidonie, Stéphanie,

Sylvie, Yanis. Ce fut un plaisir de les côtoyer et d'échanger avec eux sur tous sujets scientifiques ou non. Je dois aussi à Sylvie d'avoir été bien nourri par la cantine du CNRS.

Merci à mes amis. Merci à ceux (notamment Bruno) qui déjà docteurs, m'ont inspiré à franchir finalement le pas. Merci à Stéphane d'avoir été mon compagnon de voyages (et de rêves de voyages) y compris pendant la période du Covid qui s'y prêtait peu. Merci enfin à ma famille qui a toujours été à mes côtés pendant ces années.

Valorisation scientifique

Articles

Sur l'AIC étendu

- Matthieu Pluntz, Cyril Dalmasso, Pascale Tubert-Bitter et Ismaïl Ahmed. A simple information criterion for variable selection in high-dimensional regression. *En révision.*

Sur la simulation-calibration

- *Article en cours de rédaction.*

Communications orales

Sur l'AIC étendu

- Matthieu Pluntz, Cyril Dalmasso, Pascale Tubert-Bitter et Ismaïl Ahmed. An FWER-controlling information criterion for high-dimensional variable selection. International Biometric Conference, Riga (Lettonie), 12 juillet 2022.

Sur la simulation-calibration

- Matthieu Pluntz, Cyril Dalmasso, Pascale Tubert-Bitter et Ismaïl Ahmed. A simulation-based significance test for the LASSO in generalized linear models. Channel Network, Rothamsted Research (Harpenden, Royaume-Uni), juillet 2019.

Table des matières

1 - Introduction

1.1 .Régression en grande dimension

Cette thèse porte sur la sélection de variables dans le cadre de modèles de régression en grande dimension. Les données de grande dimension sont de plus en plus fréquentes en santé publique et en recherche clinique. Cela inclut les données de génomique et des autres -omiques [1], d'imagerie médicale, d'épidémiologie numérique, ou encore les données médico-administratives. Nous nous intéresserons en particulier à la pharmacovigilance, c'est-à-dire à la détection des effets secondaires délèteres de médicaments.

De manière générale, la régression est le problème d'estimer le lien entre une grandeur numérique y (la réponse), dont on observe n réalisations $y_1, .., y_n$, et un ensemble de p grandeurs numériques (les régresseurs, ou covariables) dont on observe également n réalisations sur les mêmes individus. Elles sont représentées par une matrice X de dimensions (n, p). Le modèle régression multiple le plus simple est le modèle linéaire. L'effet de chaque variable X_j y est quantifié par un coefficient β_j :

$$\forall\, i \in \{1, .., n\},\ y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \epsilon_i,\ \ \forall i,\ \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

soit en écriture matricielle, $y = \beta_0 + X\beta + \epsilon,\ \epsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$. Une formulation équivalente plus compacte est $Y = X^*\beta^* + \epsilon$ où $\beta^* = (\beta_0, \beta) \in \mathbb{R}^{p+1}$ et X^* est la matrice de dimensions $(n, p+1)$ obtenue en concaténant à X un vecteur colonne constant égale à 1.

1. sciences des données moléculaires qui déterminent le phénotype d'une personne, telles que la protéomique ou la métabolomique.

Le modèle linéaire est adapté aux situations où y est une grandeur continue. Dans d'autres cas, on préfère utiliser un modèle linéaire généralisé, dont la forme générale est

$$\mathrm{E}[Y] = f(X^*\beta^*)$$

où f est une fonction d'une variable réelle, la *fonction de lien*. Cela inclut le modèle binaire où y prend ses valeurs dans $\{0, 1\}$ ce qui indique la présence où l'absence d'un phénomène, qui peut être une pathologie. f doit alors être une fonction croissante à valeurs dans $]0, 1[$, le choix le plus classique étant la fonction logistique $x \to e^x/(1 + e^x)$.

L'ajustement d'un modèle de régression multiple permet de tester la significativité du paramètre β_j associé à chaque covariable j. On peut ensuite réaliser une sélection de variables à partir de tels tests, en concluant que les covariables dont le coefficient estimé est significativement différent de 0 sont actives. Il est aussi possible de considérer la sélection de modèle de façon plus globale, en évaluant les modèles restreints à des sous-ensembles de variables et en sélectionnant les variables qui constituent celui qui optimise un critère d'intérêt. En pharmacovigilance, qui est le domaine d'application de cette thèse, la sélection de variables, exploratoire, joue un rôle de détection de signal. X regroupe un grand nombre d'expositions médicamenteuses pouvant induire l'évènement indésirable d'intérêt y. Les variables sélectionnées par l'analyse statistique sont des signaux qui devront ensuite faire l'objet d'analyses pharmacologiques ou épidémiologiques plus poussées.

Les problèmes de grande dimension, c'est-à-dire lorsque p est élevé (éventuellement, mais pas nécessairement, supérieur à n), sont particulièrement difficiles. Cette tendance générale, dite « fléau de la dimension » ou *curse of dimensionality* (Giraud, 2021) se manifeste à la fois dans l'estimation de modèles non paramétriques, dans celle de modèles paramétriques où il est nécessaire

d'explorer un espace de paramètres de grande dimension pour trouver une solution optimale. Même dans le modèle linéaire pour lequel on dispose d'une solution analytique lorsque $p < n$ (l'estimateur des moindres carrés), celle-ci est affectée d'une importante variance. Lorsque $p \geq n$ le modèle n'est de plus pas identifiable. En outre, l'estimation d'un trop grand nombre de paramètres peut rendre les modèles peu utiles, en particulier difficilement interprétables et peu fiables du point de vue de la prédiction comme de la sélection de variables. La recherche de modèles parcimonieux est donc un problème central.

La sélection de variables doit donc s'adapter au contexte de la grande dimension. Il n'est en général plus possible de la fonder sur l'estimation classique du modèle multivarié dans son ensemble. En génomique, il est courant de réaliser de multiples régressions univariées du vecteur réponse sur chacune des covariables candidates, chacune fournissant un test de significativité. On effectue ensuite une procédure de correction des tests multiples sur les résultats de ces tests. Cette technique ne tient cependant pas compte de la structure de corrélation entre covariables, ou alors seulement lors d'une élimination préliminaire de covariables très fortement corrélées. Une variable peut donc être sélectionnée alors que son association à la réponse est uniquement portée par sa corrélation à des variables actives.

La section suivante présente le Lasso, une technique de régression paramétrique particulièrement adaptée à la sélection de variables en grande dimension.

1.2 .Régression Lasso

Tibshirani (1996) a introduit le Lasso (*least absolute shrinkage and selection operator*), une méthode de sélection de variables et d'apprentissage automatique qui s'est avérée très féconde par les applications pratiques et les déve-

loppements théoriques qu'elle a occasionnés. Le Lasso généralise l'approche très générale en statistique paramétrique qui consiste à trouver le vecteur de paramètres $\beta^* = (\beta_0, \beta)$ maximisant $L(\beta^*; X, y)$, sa vraisemblance étant données les observations X et y. La quantité optimisée dans le Lasso, la fonction de perte l_λ, est une log-vraisemblance pénalisée par la norme L_1 du vecteur de paramètres ($||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$). Le poids de cette pénalité est donné par un paramètre λ, le *paramètre de régularisation* du Lasso :

$$\begin{aligned} \widehat{\beta^*}^{Lasso}(\lambda; X, y) &= \underset{\beta^* \in \mathbb{R}^{p+1}}{\operatorname{argmin}} l_\lambda(\beta^*) \\ \text{où } \; l_\lambda(\beta^*) &= -\log L(\beta^*; X, y) + \lambda ||\beta||_1. \end{aligned}$$

Dans le modèle linéaire, le Lasso se ramène à l'estimation des moindres carrés pénalisés :

$$\begin{aligned} L(\beta^*, \sigma; X, y) &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{-\frac{||y - X^*\beta^*||^2}{2\sigma^2}} \\ \widehat{\beta^*}^{Lasso}(\lambda) &= \underset{\beta^* \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \frac{||y - X^*\beta^*||^2}{2\sigma^2} + \lambda ||\beta||_1 \\ \widehat{\beta^*}^{Lasso}(\tilde{\lambda}) &= \underset{\beta^* \in \mathbb{R}^{p+1}}{\operatorname{argmin}} ||y - X^*\beta^*||^2 + \tilde{\lambda} ||\beta||_1 \end{aligned} \tag{1.1}$$

où $\tilde{\lambda} = 2\sigma^2\lambda$. L'équation (1.1) est parfois donnée comme définition du Lasso dans le cas linéaire. Elle peut aussi être vue comme la minimisation de l'erreur quadratique sous une contrainte de la forme $||\beta||_1 \leq t$, dont elle est la forme lagrangienne (Hastie *et al.* (2021), chapitre 1).

Le Lasso possède également une interprétation bayésienne. En effet, si β est un paramètre aléatoire qui suit une loi a priori de densité $(2\lambda)^p e^{-\lambda||\beta||_1}$ (c'est-à-dire si chaque β_j suit a priori une loi de Laplace de paramètre λ), alors la loi a posteriori possède une densité de probabilité proportionnelle à $e^{-l_\lambda(\beta^*)}$. L'estimateur Lasso est donc un mode a posteriori.

La forme mathématique du Lasso lui permet de réaliser directement une sélection de variables. En effet, pour tout $j \geq 1$, en tout vecteur de paramètres β tel que $\beta_j \neq 0$, la fonction de perte admet la dérivée partielle :

$$\frac{\partial l_\lambda(\beta)}{\partial \beta_j} = -\frac{\partial \log L(\beta)}{\partial \beta_j} + \text{signe}(\beta_j)\lambda$$

donc il suffit que $|\frac{\partial \log L(\beta)}{\partial \beta_j}| < \lambda$ pour que l_λ ne puisse pas admettre de minimum en β. En revanche, si cette inégalité est vraie en un β tel que $\beta_j = 0$, alors l_λ peut y admettre un minimum avec des dérivées partielles à gauche et à droite en β_j non nulles de signes opposés. Si λ est suffisamment important, la plupart des coefficients estimés $\widehat{\beta}_j^{\,Lasso}(\lambda)$ sont donc nuls. L'intercept β_0 n'étant pas pénalisé, il est par contre toujours estimé. Les indices $j \geq 1$ où $\widehat{\beta}_j^{\,Lasso}(\lambda)$ est non nul sont ceux des variables sélectionnées par le Lasso.

La pénalité a plus généralement pour effet de biaiser l'estimation des coefficients tout en diminuant sa variance. Elle produit une contraction (*shrinkage*) des $\widehat{\beta}_j^{\,Lasso}(\lambda)$ dans la direction de 0, même lorsqu'elle n'est pas assez importante pour forcer leur valeur à 0.

La sélection de variables par le Lasso est gouvernée par son paramètre de régularisation. À l'extrême, il existe une valeur de λ suffisamment élevée pour que le Lasso ne sélectionne aucune variable. Diminuer λ à partir de ce point conduit à sélectionner plus de variables, jusqu'à $\lambda = 0$ qui correspond au maximum de vraisemblance où, en général, tous les paramètres estimés sont non nuls. La fonction $\lambda \to \widehat{\beta}^{Lasso}(\lambda)$ constitue le « chemin du Lasso ». On dit que la variable d'indice j « apparaît sur chemin du Lasso » au paramètre limite $\sup\{\lambda : \widehat{\beta}_j^{\,Lasso}(\lambda) \neq 0\}$. Le chemin du Lasso définit une suite de sous-ensembles de $\{1, .., p\}$, les supports $S(\lambda) = \left\{j \in \{1, .., p\} : \widehat{\beta}_j^{\,Lasso}(\lambda) \neq 0\right\}$, qui représentent chacun un modèle plus parcimonieux que le modèle complet. Bien que l'allure générale du chemin du Lasso soit celle de modèles de

moins en moins parcimonieux à mesure que λ diminue, il peut arriver (en particulier en cas de fortes corrélations entre covariables) qu'une variable apparaisse sur le chemin puis en ressorte, c'est-à-dire qu'il existe j, λ, λ' tels que $\lambda > \lambda'$, $\widehat{\beta}_j^{Lasso}(\lambda) \neq 0$ et $\widehat{\beta}_j^{Lasso}(\lambda') = 0$. La sélection de modèle à l'aide du Lasso consiste à choisir l'un de ces $S(\lambda)$, ou, de façon équivalente, à déterminer une valeur optimale de λ et à sélectionner le modèle correspondant.

Le Lasso peut être estimé efficacement par plusieurs algorithmes. Nous avons utilisé le package `glmnet` de R, qui se fonde sur l'algorithme de descente de coordonnées, *coordinate descent* (Hastie *et al.*, 2023). Les $\widehat{\beta}^{Lasso}$ sont estimés de proche en proche le long du chemin du Lasso à des valeurs faiblement espacées du paramètre de régularisation, chaque estimation étant initialisée à celle du λ voisin (Bühlmann et van de Geer (2011), chapitre 2).

1.3 .Méthodes de sélection de variables sur le chemin du Lasso

La sélection de variables par Lasso est très dépendante de son paramètre de régularisation λ. L'utilisation du Lasso pour réaliser une sélection de variables doit donc intégrer une procédure de sélection plus complète. Nous présentons dans cette section plusieurs de ces procédures.

Certaines d'entre elles sont des critères de choix du paramètre λ (validation croisée, cf 1.3.1, et sélection par permutations, cf 1.3.3). D'autres permettent de choisir un des modèles apparaissant sur le chemin du Lasso. Cela peut être fait via un critère de comparaison entre ces modèles (critères d'information, cf 1.3.2) ou bien, en se plaçant dans le paradigme des tests d'hypothèses, en testant la significativité de chacune des variables apparaissant sur le chemin du Lasso (test de covariance, cf 1.3.4). Enfin, des méthodes combinent le Lasso à un rééchantillonnage pour construire l'ensemble des variables que l'on sélectionne, sans que cet ensemble apparaisse nécessaire-

ment sur le chemin du Lasso (*stability selection* et ses variantes, cf 1.3.3).

1.3.1 . Validation croisée

La validation croisée (*cross-validation*) est une technique générale d'estimation des performances de prédiction d'un modèle statistique, qui permet de sélectionner le modèle qui optimise ces performances (Hastie *et al.* (2017), chapitre 7). Elle est très souvent utilisée pour choisir le paramètre de régularisation d'un Lasso.

On se donne une famille $(\lambda_1, .., \lambda_L)$ de paramètres de régularisation, parmi lesquels on souhaite déterminer celui qui minimise l'erreur de prédiction. Pour cela, la validation croisée partitionne les données en K sous-ensembles (typiquement $K = 5$ ou 10). Pour $k = 1, .., K$, on réalise le Lasso sur tous les individus sauf ceux appartenant au k-ième sous-ensemble, ce qui fournit pour chaque paramètre de régularisartion testé un vecteur de paramètres estimés : $\widehat{\beta^*}^{Lasso(k)}(\lambda_1), .., \widehat{\beta^*}^{Lasso(k)}(\lambda_L)$. On calcule ensuite les prédictions fournies par ces estimés sur le k-ième sous-ensemble, qui avait été exclu de l'ajustement, et leur erreur de prédiction sur cet ensemble (erreur quadratique moyenne dans le modèle linéaire, ou plus généralement la moyenne de moins la log-vraisemblance). Le λ_l retenu est celui qui minimise la moyenne des erreurs de prédiction sur les K régressions Lasso effectuées. Dans une variante plus conservative, le λ_l retenu est le plus grand à posséder une erreur de prédiction moyenne inférieure à l'erreur moyenne minimale plus son écart-type estimé.

Cette méthode optimise les performances de prédiction du modèle ajusté par le Lasso. Il a été observé qu'elle conduit à une sélection de variables souvent trop libérale (Chen et Chen, 2008), c'est-à-dire au choix d'un λ trop faible. En effet, diminuer λ peut conduire à sélectionner des variables inactives en leur affectant un faible $\widehat{\beta}_j^{Lasso}$, ce qui n'a qu'un faible effet sur les perfor-

mances de prédiction, tout en diminuant le biais de contraction dans l'estimation des coefficients associés aux variables actives, ce qui a un effet positif sur les performances de prédiction pouvant compenser l'effet négatif dû à la sélection de faux positifs. La sélection de variable par validation croisée ne constitue pas une procédure de test de la significativité des variables qu'elle sélectionne. Elle ne fournit aucun contrôle de l'erreur de première espèce.

1.3.2 . Critères d'information

Les critères d'information, ou IC (de l'anglais *Information criteria*) sont des critères de sélection de modèles fondés sur un compromis entre la qualité de l'ajustement d'un modèle et sa complexité. Un critère d'information sélectionne un modèle parmi une famille de modèles paramétriques en minimisant une fonction du nombre de paramètres du modèle A, noté $|A|$, de sa log-vraisemblance maximale sur les données $l\left(\widehat{\theta_A}\right)$, et du nombre d'observations n.

Fondements théoriques

Konishi et Kitagawa (1996) ont proposé une théorie générale des critères d'information en les définissant à partir de la divergence de Kullback-Leibler entre la véritable loi G des données, de densité g, et la loi paramétrique estimée, de densité $f(\cdot, \hat{\theta}_A)$:

$$D_{KL}(g, f(\cdot, \hat{\theta}_A)) = \int g(x) \log \left(\frac{g(x)}{f(x; \hat{\theta}_A)} \right) dx.$$

x représente ici l'ensemble des données, y compris y dans un modèle de régression. Dans cette approche, on cherche la loi paramétrique qui minimise la divergence de Kullback-Leibler, ce qui se ramène à maximiser l'intégrale :

$$\eta(\hat{\theta}_A) = \int g(x) f(x; \hat{\theta}_A) dx$$

qui est elle-même inconnue, g n'étant pas connue, mais peut être estimée par une quantité proportionnelle à la log-vraisemblance :

$$\frac{1}{n} l(\hat{\theta}_A) = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i; \hat{\theta}_A).$$

Cette estimation consiste à remplacer la loi véritable G par la loi empirique (discrète) $\hat{G}$ fournie par les données. Or, puisque $l(\hat{\theta}_A)$ est estimé par maximum de vraisemblance à partir de $\hat{G}$, cette dernière est plus proche de la loi donnée par $\hat{\theta}_A$ que ne l'est G. L'estimation est donc généralement biaisée dans le sens d'une surestimation de $\eta(\hat{\theta}_A)$. Akaike (1973) a prouvé que ce biais est asymptotiquement équivalent à $|A|/n$ si la famille paramétrique $f(\cdot, \theta_A)$ inclut la loi véritable G. Généralement, la quantité que l'on cherche à maximiser est la log-vraisemblance pénalisée d'un terme qui permet de débiaiser l'estimation de la divergence de Kullback-Leibler.

Forme typique des IC et exemples classiques

Dans les exemples les plus classiques, la pénalité retenue est proportionnelle au nombre de paramètres du modèle. La forme typique des IC et la suivante :

$$IC(A) = w * |A| - 2\, l(\widehat{\theta_A}) \tag{1.2}$$

où w est le poids qui pénalise l'ajout de nouveaux paramètres dans A. Les IC les plus utilisés sont le critère d'information d'Akaike (Akaike, 1973) et le critère d'information bayésien (Schwarz, 1978). Ils sont définis par les valeurs suivantes de w :

- AIC : $w = 2$
- BIC : $w = w(n) = \log(n)$

En sélection de variables en grande dimension, on dispose d'un modèle de régression comportant un grand nombre de paramètres, p, et on cherche

à sélectionner un sous-modèle parcimonieux de ce modèle en comparant leurs IC. Pour des raisons combinatoires, le nombre de sous-modèles est trop important pour permettre le calcul de l'IC de chacun d'entre eux. On utilise donc un algorithme de sélection préliminaire des variables, tel que le Lasso, qui produit une famille plus restreinte de sous-modèles candidats. On sélectionne ensuite le modèle dont l'IC (calculé sur les paramètres estimés sans pénalisation) est le plus faible parmi cette famille. La combinaison du BIC à un Lasso préliminaire a été utilisée et comparée à d'autres méthodes, par exemple dans Sabourin *et al.* (2015) et Courtois *et al.* (2021).

Le nombre de paramètres d'un sous-modèle peut être vu comme une sorte de « norme » L_0, car $|A| = \sum_{j=1}^{p} |\beta_j|^0$ (Viallon (2016), annexe A2). De ce point de vue, les IC s'inscrivent donc dans la famille des régressions pénalisées qui inclut le Lasso (norme L_1) et la régression Ridge (norme L_2).

En grande dimension, l'AIC et le BIC peuvent produire un grand nombre de faux positifs (Giraud (2021), chapitre 2). Ils ne permettent pas le contrôle des critères statistiques de tests multiples tels que le taux d'erreur à l'échelle de la famille (*Family-wise error rate*, FWER) ou le taux de fausses découvertes (*False discovery rate*, FDR).

Pour résoudre ce problème, Chen et Chen (2008) ont proposé le critère d'information bayésien étendu (EBIC), une extension du BIC. Il s'agit d'un critère d'information pour les modèles de régression en grande dimension qui prend en compte à la fois n et p :

$$\mathrm{EBIC}(A) = \log(n)|A| + 2\log\binom{p}{|A|} - 2\,l(A). \tag{1.3}$$

Lorsque p est grand, l'ajout d'un nouveau paramètre dans le modèle est plus pénalisé pour tenir compte du grand nombre de modèles candidats. L'EBIC garantit le contrôle asymptotique du FDR et est asymptotiquement consis-

tant. Il ne contrôle en revanche pas le FWER.

1.3.3 . Méthodes par rééchantillonnage

Stability selection

Meinshausen et Bühlmann (2010) proposent la « sélection par stabilité » (*stability selection*), une approche générale de sélection de variable fondée sur le rééchantillonnage. Elle consiste à appliquer le Lasso (ou une autre méthode de sélection de variables dépendant d'un paramètre) à un grand nombre B de rééchantillonages sans remise de taille $\lfloor n/2 \rfloor$ des données, ce qui fournit des estimés $\widehat{\beta}_j^{Lasso-b}$, $b = 1, .., B$. On calcule ensuite pour chaque variable j sa proportion de sélections parmi les rééchantillonnages :

$$\hat{\pi}_j^{\lambda} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1} \left\{ \widehat{\beta}_j^{Lasso-b}(\lambda) \neq 0 \right\}.$$

Les variables sélectionnées sont celles dont le $\hat{\pi}_j^{\lambda}$ maximal sur un ensemble de valeurs de λ dépasse un certain seuil, qui peut être choisi pour contrôler le FWER. Cette méthode est une généralisation du Bolasso Bach (2008), ou Lasso amélioré par bootstrap, où le seuil est fixé à 1. Ahmed *et al.* (2018) ont proposé le Lasso par rééchantillonnage déséquilibré selon la classe (*Class-Imbalanced Subsampling Lasso*, CISL), où l'on s'intéresse à la distribution de proportions de sélections calculées le long du chemin du Lasso. Le rééchantillonage y tient compte du caractère déséquilibré de la réponse y que l'on rencontre en pharmacovigilance.

Sélection par permutations

Sabourin *et al.* (2015) proposent un critère d'arrêt sur le chemin du Lasso fondé sur un rééchantillonage : la sélection par permutations (*permutation selection*). Cette méthode, comme celle de la validation croisée, consiste à dé-

terminer à partir des données une valeur optimale du paramètre de régularisation λ. Les variables retenues seront celles sélectionnées par le Lasso à cette valeur du paramètre.

L'idée de la sélection par permutations est de déterminer la valeur de λ en-dessous de laquelle il est probable que le Lasso sélectionne fortuitement des variables en réalité indépendantes de Y. Pour cela, on utilise des rééchantillonages pour imposer l'indépendance entre la réponse et chacun des X_j, en permutant aléatoirement le vecteur y. Pour un nombre N de rééchantillonages, pour tout $l \in \{1, .., N\}$,

1. On tire une permutation aléatoire π_l de l'ensemble $\{1, .., n\}$.
2. On réalise la régression Lasso du vecteur réponse permuté $y_{\pi_l} = \left(y_{\pi_l(1)}, .., y_{\pi_l(n)}\right)$ sur les X_j.
3. On mesure $\lambda_0(\pi_l)$, la plus petite valeur de λ à laquelle cette régression Lasso ne sélectionne aucune variable (c'est-à-dire le λ d'apparition d'une première variable sur le chemin du Lasso).

Le paramètre de régularisation choisi est alors la médiane $\hat{\lambda}_{\text{perm}}$ de l'ensemble $\{\lambda_0(\pi_1), .., \lambda_0(\pi_N)\}$.

Par construction, chacun des vecteurs permutés y_{π_l} présente la même distribution que y (mêmes moyenne, variance, etc.) tout en étant « indépendant » des X_j (dans le sens où la loi de $y_{\pi_l(i)}$ est la même pour toute observation i, même si les $X_{i,j}$ varient). Les $\lambda_0(\pi_l)$ indiquent donc à quel point du chemin du Lasso le risque de sélection de variable fortuite apparaît, sous une hypothèse d'indépendance entre Y et toutes les X_j.

En revanche, la sélection par permutation ne permet pas d'explorer l'hypothèse d'une réponse associée à certaines variables X_j seulement. La loi asymptotique de $\hat{\lambda}_{\text{perm}}$ dépend de $||\beta||$, qui mesure la force du signal, un signal plus fort produisant un $\hat{\lambda}_{\text{perm}}$ plus élevé. Un fort signal porté par certaines

variables actives va donc entraîner une plus faible sensibilité à la détection du signal porté par d'autres variables actives.

1.3.4 . Test de covariance

Lockhart *et al.* (2014) proposent d'adapter au Lasso le paradigme des tests d'hypothèse. Ils construisent un test de significativité de j_k la k-ième variable sélectionnée sur le chemin du Lasso en mesurant l'impact de son apparition dans le modèle comparée à la sélection par le Lasso de l'ensemble A des $k-1$ premières variables seulement. La statistique de test est calculée à partir du résultat du Lasso à la plus petite valeur du paramètre de régularisation où l'ensemble des variables sélectionnées est $A \cup \{j_k\}$, c'est à-dire la valeur λ_{k+1} où apparaît la variable suivante; ce résultat est comparé à celui d'un Lasso restreint aux variables appartenant à A. La statistique de test est une différence de covariances :

$$T_k(y) = \left(\mathrm{Cov}\left(y, X^* \widehat{\beta^*}^{Lasso}(\lambda_{k+1}; y, X) \right) - \mathrm{Cov}\left(y, X_A^* \widehat{\beta^*}^{Lasso}(\lambda_{k+1}; y, X_A) \right) \right) / \sigma^2.$$

Lockhart et al. prouvent que dans le modèle linéaire, sous l'hypothèse nulle qui affirme que les $k-1$ premières variables sélectionnées par le Lasso contiennent toutes les variables actives, T_k suit la loi connue $\mathrm{Exp}(1)$. Si, comme c'est généralement le cas en pratique, la variance σ^2 est inconnue, elle est remplacée par son estimée $\hat{\sigma}^2$ dans le calcul de la statistique de test et celle-ci suit alors une autre loi connue, une loi de Fisher $F_{2,n-p}$. Dans les deux cas, il est donc possible de calculer exactement la p-value associée à ce test. En revanche, dans le modèle binaire, la statistique de test ne suit pas exactement la loi $\mathrm{Exp}(1)$ ni une autre loi connue, bien que les auteurs conjecturent que ce soit sa loi asymptotique.

Dans le test de covariance, le rejet ou non de l'hypothèse nulle relative à la

k-ième variable dépend du rôle de la $k+1$-ième variable. Si cette dernière est assez fortement associée au vecteur réponse pour que $\lambda_{k+1} \approx \lambda_k$, les deux $\widehat{\beta^*}^{Lasso}$ seront proches (car ils sont des fonctions continues de λ et coïncident en $\lambda = \lambda_k$) donc la différence de covariance sera faible et l'hypothèse nulle ne sera pas rejetée, c'est-à-dire l'impact de la k-ième variable jugé non significatif. Ce test présente donc une situation de faible puissance qui semble évitable, car la présence de deux variables ou plus fortement actives ne devrait pas induire une méthode de sélection à n'en détecter aucune.

1.4 .Objectifs de la thèse

Cette thèse vise à proposer deux procédures de sélection de variables en grande dimension. Ils répondent au problème du choix d'un des modèles proposés par le Lasso.

Le premier de ces deux algorithmes, qui fait l'objet du chapitre 2, est un critère d'information qui appartient à la famille de critères présentée en 1.3.2. Nous proposons une formule simple qui suit la forme générale (1.2) sans faire intervenir le nombre d'observations n. Il s'agit donc d'une généralisation de l'AIC et non du BIC, d'où son nom d'AIC étendu ou EAIC. En revanche, comme le BIC étendu de Chen et Chen, notre critère s'adapte au contexte de la sélection de variable en grande dimension en prenant en compte le nombre p de covariables candidates.

Notre approche consiste à aborder la comparaison d'un critère d'information entre des sous-modèles d'un modèle de grande dimension comme des réalisations multiples d'un test de rapport de vraisemblance. Nous prouvons une propriété asymptotique portant sur les tests de rapport de vraisemblance multiples et indépendants. Elle nous permet de construire l'EAIC de façon à ce que la sélection de variable à l'aide de ce critère contrôle approximative-

ment le family-wise error rate lorsque les dimensions du problème sont importantes et les variables à sélectionner sont importantes.

Notre seconde proposition, exposée au chapitre 3, est, comme le CovTest de Lockhart et al., un test de la significativité d'une variable sélectionnée par le Lasso. À la différence du CovTest, notre proposition se base sur la valeur du paramètre de régularisation associée à la variable dont on teste la significativité (λ_k dans le paragraphe 1.3.4). L'hypothèse nulle de ce test n'étant pas assez spécifiée pour obtenir sa p-value, nous en estimons une « p-value conditionnelle » à l'aide d'une procédure de Monte-Carlo impliquant la simulation de vecteurs sortie sous l'hypothèse nulle, puis une étape de calibration qui leur fait satisfaire la condition sous laquelle est définie la p-value conditionnelle. Il s'agit donc d'une procédure de *simulation-calibration*. Nous démontrons que dans le modèle linéaire, cette procédure assure le contrôle de l'erreur de première espèce sous l'hypothèse nulle, selon laquelle toutes les variables actives appartiennent à un ensemble connu A de taille réduite. Nous développons des méthodes de calibration dans les modèles non linéaires. Nous démontrons de plus un théorème « étendu » de contrôle sous certaines hypothèses de la probabilité de sélectionner un faux positif même lorsque l'hypothèse nulle n'est pas vérifiée, c'est-à-dire lorsqu'il existe des variables actives n'appartenant pas A.

Pour chacune de ces propositions, nous mesurons les performances de sélection de variables de notre méthode à l'aide d'études de simulations poussées, en les comparant à celles des méthodes existantes dont elles s'inspirent.

Nous illustrons de plus ces procédures par une application à la détection des signaux en pharmacovigilance : l'exploitation de la base nationale de pharmacovigilance (BNPV), qui regroupe des notifications spontanées d'expositions médicamenteuses coïncidant à des évènements indésirables.

2 - Critère d'information d'Akaike étendu (EAIC)

2.1 .Introduction

Ce chapitre porte sur les critères d'information (en anglais *information criteria*, IC, introduits dans la section 1.3.2) et leur utilisation en sélection de modèle dans le contexte d'une régression de grande dimension. Nous nous concentrons sur les IC qui suivent la forme la plus fréquente, celle de la log-vraisemblance affectée d'une pénalité proportionnelle aux nombre de paramètres du modèle A considéré :

$$IC(A) = w * |A| - 2\, l(\widehat{\theta_A}). \tag{2.1}$$

Nous proposons une version étendue du critère d'information d'Akaike (AIC, dans lequel $w = 2$) que nous appelons l'AIC étendu où *Extended Akaike Information Criterion* (EAIC). Ce critère suit la forme générale (2.1). Son poids w est indépendant de n (comme dans l'AIC) et dépend de p (comme dans l'EBIC de Chen et Chen (2008)). Il est construit de façon à ce que lorsque les p covariables sont indépendantes et n et p sont tous les deux élevés, le critère contrôle le FWER. Ce contrôle repose sur une propriété de la distribution asymptotique du rapport de vraisemblance entre deux modèles imbriqués. La section 2.2 présente cette propriété et sa signification en termes de tests de rapport de vraisemblance multiples. Sa démonstration est l'objet de la sous-section 2.2.3.

La section 2.3 discute de la sélection des variables à l'aide de critères d'information. Elle décrit les critères existants en termes de tests de rapport de vraisemblance, définit l'EAIC sur la base des résultats de la section 2.2 et dé-

crit sa combinaison avec le Lasso. Nous avons mené une étude de simulation pour évaluer les performances des différents critères d'information, dont l'EAIC, dans un grand nombre de scénarios de régression linéaire ou logistique en grande dimension définis par une large gamme de valeurs de n, de p, de l'intensité du signal et de la corrélation entre les variables. Ces simulations et leurs résultats sont décrits dans la section 2.4. Dans la section 2.5, nous illustrons notre approche par une application à la détection de signaux dans la base nationale de pharmacovigilance (BNPV). Nous comparons les performances des différents IC à l'aide de DILIrank, un ensemble de référence qui recense l'association de médicaments aux lésions hépatiques d'origine médicamenteuse (*drug-induced liver injuries*, DILI) (Chen *et al.*, 2016).

2.2 . Contrôle du FWER lors de tests de rapport de vraisemblance multiples et indépendants

Cette section présente une propriété des tests du rapport des vraisemblance lors de l'insertion d'un paramètre supplémentaire dans un modèle paramétrique. Elle permet de contrôler approximativement le FWER en cas de tests multiples indépendants en choisissant comme valeur critique du test une fonction simple du nombre approximatif de tests et du FWER souhaité.

2.2.1 . Position du problème et notations

On se donne un modèle statistique de log-vraisemblance $l(\theta; X)$. Soient Θ_0 un sous-espace de l'espace des paramètres de dimension d, et Θ_1 un sous-espace de dimension $d+1$ contenant Θ_0. On note l_0 le maximum de la log-vraisemblance sur Θ_0, l_1 son maximum sur Θ_1, et $\Delta l = l_1 - l_0$. Soit H_0 l'hypothèse énonçant que le vrai vecteur de paramètres appartient à Θ_0 et H_1 l'hypothèse alternative d'après laquelle il appartient à $\Theta_1 \backslash \Theta_0$.

Le théorème de Wilks (Wilks, 1938) affirme que sous H_0, $2\Delta l$ suit asymptotiquement une loi du χ^2 à un degré de liberté, autrement dit, Δl suit asymptotiquement une loi gamma de paramètres $(\frac{1}{2}, 1)$. Soit f la densité de probabilité de cette loi, et F sa fonction de répartition. Dans un test du rapport de vraisemblance, la statistique Δl est utilisée pour tester H_0 contre H_1. Une procédure qui rejette H_0 en faveur de H_1 si $\Delta l > x$, pour une certaine valeur critique x, a une probabilité asymptotique $1 - F(x)$ de rejeter à tort H_0 lorsque celle-ci est vérifiée.

Considérons maintenant q sous-espaces $\Theta_1^1, ..., \Theta_1^q$ de dimension $d+1$ de l'espace des paramètres, contenant tous Θ_0. Soient H_i l'hypothèse énonçant que le vrai vecteur de paramètres appartient à $\Theta_1^i \backslash \Theta_0$, l_1^i le maximum de la log-vraisemblance sur Θ_1^i, et $\Delta l^i = l_1^i - l_0^i$.

On réalise les tests de H_0 contre chaque H_i, pour $i = 1, ...q$ à l'aide d'une procédure qui rejette H_0 en faveur de H_i si $\Delta l^i > x$ pour une certaine limite x qui ne dépend pas de i. Ces tests sont un outil de sélection de modèles : si H_0 n'est rejeté pour aucun des i, le modèle dont les paramètres appartiennent à Θ_0 est sélectionné. Si elle est rejetée au moins une fois, alors un modèle avec plus de paramètres sera sélectionné. Nous cherchons à contrôler la probabilité de cet événement sous H_0.

Nous supposons que les q tests du rapport des vraisemblances sont indépendants, c'est-à-dire que les Δl^i sont des variables aléatoires indépendantes. Alors sous H_0 la probabilité asymptotique que le modèle correct soit sélectionné, c'est-à-dire qu'aucun des q tests ne rejette H_0, est $F^q(x) = F(x)^q$. F^q est la fonction de répartition asymptotique de $\max(\Delta l^1, .., \Delta l^q)$. Nous allons démontrer une propriété asymptotique de F^q aux grandes valeurs de q qui se traduit par un contrôle du FWER.

2.2.2 .Énoncé dans le cas simple

Définition 1 (suite des valeurs critiques)**.** Pour tout $\alpha \in]0,1[$, la suite des valeurs critiques au niveau α est la suite $(x_{q,\alpha})_{q\in\mathbb{N},\ q\geq 2}$ où :

$$x_{q,\alpha} = \log q - \frac{1}{2}\log\log q - \log\left(-\log(1-\alpha)\right) - \frac{1}{2}\log\pi.$$

La suite des valeurs critiques vérifie la propriété suivante, que nous démontrerons sous une forme plus générale en section 2.2.3 :

$$\lim_{q\to\infty} F^q(x_{q,\alpha}) = 1-\alpha. \tag{2.2}$$

Cela signifie que lorsqu'on effectue un nombre q de tests indépendants du rapport des vraisemblances de H_0 contre H_i pour $i = 1, ...q$, si les Δl^i suivent leur distribution asymptotique, alors le fait de rejeter H_0 en faveur de H_i si $\Delta l^i > x_{q,\alpha}$ permet de contrôler le FWER à un niveau tendant vers α lorsque q tend vers l'infini.

L'interprétation que nous en faisons en pratique est que pour toute famille de tests indépendants, lorsque le nombre d'observations n et de tests q sont tous deux élevés, le FWER est contrôlé approximativement au niveau α. Il faut cependant remarquer que le résultat que nous prouvons ne détaille pas le comportement asymptotique en n : il suppose que n est suffisamment grand pour remplacer la loi des Δl^i par leur loi asymptotique. Pour démontrer une résultat analogue à (2.2) sous la forme d'une limite en $n, q \to \infty$ sans avoir à faire une telle hypothèse, il faudrait disposer de résultats sur la vitesse de convergence en loi des Δl^i vers leur loi asymptotique.

Compte tenu de la définition (1), la suite des valeurs critiques prend sa forme la plus simple lorsque $\alpha = \alpha_0 = 1 - e^{-\frac{1}{\sqrt{\pi}}} \approx 0.43118$. La borne $x_{q,\alpha_0} = \log q - \frac{1}{2}\log\log q$ contrôle le FWER au niveau α_0.

2.2.3 . Cas général et démonstration

En pratique, la valeur critique x ne doit pas varier en fonction du nombre exact de tests q. En effet, comme nous le verrons dans la section 2.3, l'utilisation d'un critère d'information revient à effectuer plusieurs procédures de tests multiples dans lesquelles q varie légèrement, mais x reste le même, pour toutes les familles de tests. Par conséquent, nous remplaçons q par une quantité qui l'approche, q', dans la définition de x. Nous disposons du lemme suivant :

Lemme 1. *Pour tout $\alpha \in]0,1[$, pour toute suite $q'(q)$ (notée simplement q') telle que $q' \sim_\infty q$, la suite $x_{.,\alpha}$ définie en* (1) *satisfait :*

$$\lim_{q\to\infty} F^q(x_{q',\alpha}) = 1 - \alpha.$$

Cela signifie que lorsqu'on effectue un grand nombre q de tests du rapport de vraisemblance indépendants où les rapports de vraisemblance suivent leur distribution asymptotique, pour toutes les valeurs de q' approximativement égales à q, la valeur $x_{q',\alpha}$ contrôle le FWER approximativement au niveau α. L'équation (2.2) dont nous avons discuté les implications est le cas particulier de ce lemme en $q' = q$.

Démonstration. Nous supposons que chaque Δl^i suit sa distribution asymptotique : une loi gamma de densité de probabilité $f(x) = \frac{1}{\sqrt{\pi}} x^{-\frac{1}{2}} e^{-x}$. Prouvons d'abord que la fonction de répartition associée, F, possède les bornes suivantes :

$$1 - f(x) \leq F(x) \leq 1 - \left(1 - \frac{1}{2x}\right) f(x) \tag{2.3}$$

Soient $G(x) = 1 - f(x)$ et $H(x) = 1 - \left(1 - \frac{1}{2x}\right) f(x)$. Comme f tend vers 0, à

la fois G et H tendent vers 1 en $+\infty$. Donc pour tout x,

$$\begin{aligned} F(x) &= 1 - \int_x^\infty F'(t)\, dt \\ G(x) &= 1 - \int_x^\infty G'(t)\, dt \\ H(x) &= 1 - \int_x^\infty H'(t)\, dt. \end{aligned}$$

De plus,

$$\begin{aligned} G'(t) &= -f'(t) = \left(1 + \frac{1}{2t}\right) f(t) \\ F'(t) &= f(t) \\ H'(t) &= -\frac{1}{2t^2} f(t) - \left(1 - \frac{1}{2t}\right) f'(t) \\ &= -\frac{1}{2t^2} f(t) + \left(1 - \frac{1}{4t^2}\right) f(t) = \left(1 - \frac{3}{4t^2}\right) f(t) \\ H' &\leq F' \leq G'. \end{aligned}$$

En combinant cette inégalité avec les intégrales ci-dessus, $G(x) \leq F(x) \leq H(x)$. (2.3) est démontré.

Nous utilisons maintenant (2.3) pour borner $\log F(x)$. Nous nous servons de l'inégalité suivante, qui est une conséquence du développement de Taylor du logarithme au voisinage de 1 :

$$\forall t \in [0, \frac{1}{2}],\ -t - t^2 \leq \log(1-t) \leq -t.$$

$$\begin{aligned} \forall x \geq 1,\ \log F(x) &\geq \log\left(1 - f(x)\right) \geq -f(x) - f(x)^2 \\ \log F(x) &\leq \log\left(1 - \left(1 - \frac{1}{2x}\right) f(x)\right) \leq -\left(1 - \frac{1}{2x}\right) f(x) \\ -\left(1 + f(x)\right) f(x) &\leq \log F(x) \leq -\left(1 - \frac{1}{2x}\right) f(x) \end{aligned}$$

$$\log\left(1-\frac{1}{2x}\right) \leq \log\frac{-\log F(x)}{f(x)} \leq \log\left(1+f(x)\right).$$

On définit :

$$T(x) = \log\frac{-\log F(x)}{f(x)}.$$

D'après les inégalités ci-dessus,

$$\lim_{x\to\infty} T(x) = 0. \tag{2.4}$$

Transformons maintenant $T(x)$ pour y faire apparaître les composantes de $x_{q',\alpha}$. D'après l'expression de $f(x)$,

$$T(x) = \log\left(-\log F(x)\right) + x + \frac{1}{2}\log x + \frac{1}{2}\log\pi.$$

Pour tout q,

$$T(x) = \log\left(-\log F^q(x)\right) - \log q + x + \frac{1}{2}\log x + \frac{1}{2}\log\pi.$$

Pour tous q, q', α,

$$\begin{aligned} T(x) = \quad & \log\left(-\log F^q(x)\right) \\ + \quad & \log q' - \log q \\ + \quad & \frac{1}{2}\left(\log x - \log\log q'\right) \\ + \quad & x - \log q' + \frac{1}{2}\log\log q' + \frac{1}{2}\log\pi. \end{aligned}$$

D'après la définition

$$x_{q',\alpha} = \log q' - \frac{1}{2}\log\log q' - \log\left(-\log(1-\alpha)\right) - \frac{1}{2}\log\pi,$$

on a :

$$\begin{aligned} T(x_{q',\alpha}) &= \log\left(-\log F^q(x_{q',\alpha})\right) \\ &+ \log q' - \log q \\ &+ \frac{1}{2}\left(\log x_{q',\alpha} - \log\log q'\right) \\ &- \log\left(-\log(1-\alpha)\right). \end{aligned} \tag{2.5}$$

Et lorsque q' tend vers l'infini,

$$\begin{aligned} \log\log q' - \log\left(-\log(1-\alpha)\right) - \frac{1}{2}\log\pi &= o\left(\log q'\right) \\ x_{q',\alpha} &\sim \log q' \\ \lim_{q'\to\infty} \log x_{q',\alpha} - \log\log q' &= 0. \end{aligned}$$

De plus,

$$\lim_{\substack{q,q'\to\infty \\ q\sim q'}} \log q' - \log q = 0.$$

Dans la décomposition de $T(x_{q',\alpha})$ (2.5), la deuxième et la troisième ligne tendent vers 0 lorsque $q, q' \to \infty,\ q \sim q'$. Et $x_{q',\alpha}$ tend vers $+\infty$ donc d'après (2.4), $T(x_{q',\alpha})$ tend également vers 0. Donc :

$$\lim_{\substack{q,q'\to\infty \\ q\sim q'}} \log\left(-\log F^q(x_{q'})\right) = \log\left(-\log(1-\alpha)\right)$$

$$\lim_{\substack{q,q'\to\infty \\ q\sim q'}} F^q(x_{q'}) = 1-\alpha.$$

□

2.3 .Sélection de variables dans les régressions en grande dimension

Nous allons utiliser les résultats de la section 2.2 pour quantifier le FWER des méthodes de sélection de modèles parcimonieux par des critères d'information dans le cas simple de la comparaison d'un modèle avec tous ceux qui ont un paramètre supplémentaire. Cela nous permettra construire un critère d'information contrôlant le FWER, l'EAIC. Nous nous plaçons dans le cadre des modèles de régression linéaire généralisée ; le même raisonnement pent cependant s'appliquer à n'importe quel modèle paramétrique de grande taille.

2.3.1 .AIC et BIC

Considérons le modèle linéaire généralisé :

$$\mathrm{E}[Y] = f\left(\beta_0 + \Sigma_{j=1}^{p}\beta_j X_j\right)$$

où p est un grand nombre entier. On appelle A^* l'ensemble des indices variables actives, c'est-à-dire un sous-ensemble inconnu de $\{1, ..., p\}$ de petite taille tel que β_j soit nul pour tout $j \geq 1$ n'appartenant pas à A^*. L'objectif de la sélection des variables est de déterminer A^* sur la base des observations de Y et des X_j. Pour chaque sous-ensemble A de $\{1, .., p\}$, il est possible de calculer la log-vraisemblance $l(\beta)$ du sous-modèle suivant (également appelé modèle A) :

$$\mathrm{E}[Y] = f\left(\beta_0 + \Sigma_{j\in A}\beta_j X_j\right).$$

On séletionne ensuite un modèle en minimisant un critère d'information dépendant de $l(A)$ et du nombre $|A|$ de variables dans A, tel que l'AIC. La comparaison des IC entre modèles peut être interprétée en termes de test de rapport de vraisemblance. Soit B un sous-ensemble de $\{1, .., p\}$ contenant A

et ayant une variable de plus que A ; alors

$$\mathrm{AIC}(B) - \mathrm{AIC}(A) = 2 - 2\Delta l.$$

B est sélectionné plutôt que A si $\Delta l > 1$. Si $A^* \subset A$, la probabilité asymptotique de cet évènement est $1 - F(1) \approx 0.1573$, où F est la fonction de répartition d'une loi $\mathrm{Gamma}(\frac{1}{2}, 1)$. L'AIC est donc un moyen de fixer le taux d'erreur de première espèce à un niveau connu.

Cependant, dans ce problème de grande dimension, l'utilisation de l'AIC ou de tout autre critère similaire s'apparente à la réalisation de tests multiples du rapport de vraisemblance. Puisque $p - |A|$ variables ne sont pas incluses dans A, il existe $p-|A|$ sous-ensembles B qui contiennent A et possèdent une variable de plus que A. Lorsque nous comparons les AIC de tous ces sous-ensembles, la probabilité de choisir l'un d'entre eux plutôt que A si ce dernier est le vrai modèle (c'est-à-dire le FWER) est égale à $1 - F(1)^{p-|A|}$ si chaque $\Delta l^j,\ j \in \{1, .., p\} \setminus A$ suit la fonction de répartition asymptotique F et si les tests sont indépendants. Cette probabilité d'erreur approche 1 lorsque p est grand.

La condition d'indépendance des tests est rarement exactement vérifiée. Elle est d'autant plus proche d'être vérifiée que les différentes covariables X_j ne sont pas corrélées entre elles, ou sont faiblement corrélées. Une corrélation entre X_j créera typiquement une corrélation positive entre les Δl^j et par conséquent une probabilité d'erreur de sélection de modèle inférieure à sa valeur dans le cas indépendant, mais tendant néanmoins vers 1.

En utilisant le BIC plutôt que l'AIC, si B possède une variable de plus que A :

$$\mathrm{BIC}(B) - \mathrm{BIC}(A) = \log(n) - 2\Delta l.$$

En supposant à nouveau que chaque Δl^j suit la fonction de répartition F et que les tests sont indépendants, le FWER est :

$$\mathrm{FWER}_{BIC}(n,p) = 1 - F(\frac{1}{2}\log n)^{p-|A|}.$$

Si les modèles que nous comparons sont suffisamment parcimonieux pour que $|A| << n\log(n)$, alors :

$$\mathrm{FWER}_{BIC}(n,p) \approx 1 - e^{-p/\sqrt{\pi n \log(n)/2}}.$$

Si n tend vers l'infini et p est de l'ordre de n ou plus petit, alors le FWER du BIC converge vers 0. Cependant, comme il s'agit d'une régression en grande dimension, nous ne pouvons pas faire cette hypothèse. Par exemple, si p est de l'ordre de n^k pour un k supérieur à 1, le taux d'erreur du BIC converge vers 1. La sélection de variables fondée sur le BIC n'offre pas de borne asymptotique sur le FWER en général.

2.3.2 .AIC étendu

Les résultats de la section 2.2 conduisent à un moyen de contrôler le FWER en utilisant un critère similaire à l'AIC. Nous définissons l'AIC étendu ou EAIC (*Extended Akaike Information Criterion*) :

Définition 2 (EAIC)**.** Pour tout $\alpha \in]0,1[$, pour tout sous-ensemble A de $\{1,..,p\}$,

$$\mathrm{EAIC}_\alpha(A) = 2x_{p,\alpha}|A| - 2l(A)$$

où, comme défini en (1) :

$$2x_{p,\alpha} = 2\log p - \log\log p - 2\log\left(-\log(1-\alpha)\right) - \log\pi.$$

Contrairement à l'AIC et au BIC, l'EAIC ne dépend pas uniquement des caractéristiques propres au modèle A. Il prend également en compte le fait que A est considéré parmi tous les autres sous-modèles d'un modèle de régression à p dimensions.
Comme pour l'AIC et le BIC, cette définition se traduit par la différence suivante pour deux modèles emboîtés A et B, où B possède un paramètre de plus que A :

$$\mathrm{EAIC}_\alpha(B) - \mathrm{EAIC}_\alpha(A) = 2x_{p,\alpha} - 2\Delta l.$$

B est sélectionné plutôt que A si et seulement si $\Delta l > x_{p,\alpha}$. Par conséquent, en minimisant l'EAIC parmi tous les sous-ensembles $p - |A|$ contenant le vrai sous-ensemble actif A et possédant une variable de plus que A, si les tests sont indépendants nous obtenons la situation de la section 2.2.3 où $q = p - |A|$ et $q' = p$. Puisqu'il s'agit d'un problème de grande dimension (à la fois p et n sont élevés) et que les sous-ensembles que nous comparons sont de relativement petite taille ($|A| << p$), la situation asymptotique de la section 2.2.3 ($q \to \infty$ et $q' \sim q$) est bien approchée. Nous faisons donc l'hypothèse que les Δl^j suivent leur distribution asymptotique avec la fonction de répartition F. Par conséquent, le FWER s'approche de

$$1 - F_{\max}^{p-|A|}(x_{p,\alpha}) \approx \alpha.$$

Cas particuliers de l'EAIC

Nous avons observé en 2.2.2 que $x_{p,\alpha}$ se simplifie lorsque $\alpha = \alpha_0 = 1 - e^{-\frac{1}{\sqrt{\pi}}}$. L'EAIC possède donc la version plus simple suivante, dont le FWER approche $\alpha_0 \approx 0.4312$:

$$\mathrm{EAIC}_{\alpha_0}(A) = (2\log p - \log\log p)\,|A| - 2l(A).$$

De plus, en raison du comportement asymptotique de $x_{p,\alpha}$ lorsque $\alpha \to 0$, l'EAIC admet l'approximation suivante pour de petites valeurs de α :

$$\mathrm{EAIC}_\alpha(A) \approx (2\log p - \log\log p - 2\log\alpha - \log\pi)\,|A| - 2l(A).$$

2.3.3 . Procédure de sélection de modèle grâce à l'EAIC

Nous avons défini un nouveau critère d'information qui peut être calculé pour tout sous-modèle donné. Comme les autres critères d'information, il pourrait en principe être utilisé pour sélectionner un modèle d'intérêt en le calculant pour chaque sous-modèle et en choisissant celui dont le critère d'information est le plus faible. Bien que les propriétés des sections 2.3.1 et 2.3.2 portant sur le FWER ne se traduisent pas automatiquement par des résultats équivalents pour cette procédure, elles donnent une bonne indication de la tendance des critères d'information à produire des faux positifs.

Dans la pratique, en raison de la dimension élevée du problème de régression, le nombre de sous-ensembles de variables est en fait trop important pour calculer la vraisemblance de chacun d'entre eux. Une solution est d'utiliser une procédure de présélection qui produit une famille plus restreinte de modèles candidats peu nombreux qui expliquent particulièrement bien le résultat, puis à minimiser l'IC entre ces modèles. Il s'agit d'une approximation computationnellement plus accessible de la minimisation de l'IC parmi tous les sous-modèles. Dans ce travail, nous utilisons comme procédure de présélection la régression Lasso présentée en 1.2, où les variations du paramètre de régularisation fournissent une liste de modèles de taille réduite. Il faut néanmoins garder à l'esprit que les IC n'ont rien de spécifique au Lasso et peuvent être combinés à toute autre méthode de préselection.

La log-vraisemblance utilisée dans le calcul des critères d'information n'est pas pénalisée. Pour combiner le Lasso et un critère d'information, il est néces-

saire d'estimer chacun des modèles sélectionnés par le Lasso sans pénalité, et de ne pas seulement réutiliser les modèles ajustés avec la pénalité du Lasso.

Variante : premier minimum local du critère d'information

La minimisation de l'EAIC ou d'un autre critère d'information sur une suite ($(A_k)_{k\in\mathbb{N}}$) de modèles emboîtés admet une variante qui produit une sélection de variables plus conservative. Elle consiste à sélectionner le premier modèle A_k de cette suite tel que $EAIC(A_k) < EAIC(A_{k+1})$.

Puisque le modèle qui minimise globalement l'EAIC sur la suite vérifie aussi cette inégalité, le minimum local correspond à un k inférieur ou égal à celui du minimum global. Si chaque A_k est un sous-modèle de A_{k+1} (comme c'est souvent mais pas nécessairement le cas dans le Lasso), le modèle du premier minimum local est un sous-modèle de celui du minimum global.

Du point de vue de l'interprétation des comparaisons d'EAIC comme des tests de rapport de vraisemblance, la suite des comparaisons de $EAIC(A_k)$ à $EAIC(A_{k+1})$ est une suite de tests — chacun issu d'une famille de test corrigée de sa multiplicité — dont les hypothèses nulles sont imbriquées. Le premier minimum local revient à sélectionner le premier modèle où l'on ne rejette pas l'hypothèse nulle, de manière analogue à ce que l'on fera avec le test par simulation-calibration (section 3.9).

Le premier minimum local n'est défini que dans le contexte d'une suite présélectionnée de modèles (généralement) imbriqués. Contrairement au minimum global, il n'est pas en principe applicable à l'ensemble des sous-modèles du modèle de grande dimension.

2.4 .Études de simulation

Nous avons mené deux études de simulation. La première vise à observer le contrôle du FWER permis par l'EAIC dans un cadre simple, où il teste tous les modèles ayant une variable active par rapport au modèle nul (sans variables actives) lorsque celui-ci est correct, c'est-à-dire sous ce qui correspond à l'hypothèse nulle de tests multiples du rapport de vraisemblance. La seconde étude de simulations explore les performances de la procédure complète de sélection de modèles fondée sur Lasso de la section 2.3.3 dans une grande variété de scénarios où il existe des variables actives.

2.4.1 .Contrôle du FWER sous l'hypothèse nulle

Nous utilisons l'EAIC pour comparer le modèle nul $\mathrm{E}[Y \mid X] = f(\beta_0)$ à chacun des p modèles avec une variable active $\mathrm{E}[Y \mid X] = f\left(\beta_0 + \beta_j X_j\right),\ j = 1, ..., p$. Nous avons simulé des données sous le modèle nul, calculé l'EAIC de chaque modèle, et sélectionné le modèle à une variable active ayant l'EAIC le plus petit si celui-ci était également inférieur à l'EAIC du modèle nul. Cela équivaut à des tests de rapport de vraisemblance multiples où l'EAIC réalise une correction de la multiplicité des tests. Dans chaque scénario défini par une combinaison de paramètres, nous avons évalué le contrôle du FWER en calculant le FWER estimé, qui est la proportion de jeux de données simulés pour lesquels l'un des modèles non nuls est sélectionné.

Cela peut être considéré comme la première étape de la procédure complète de sélection de modèle : nous minimisons l'EAIC parmi les modèles de taille 0 ou 1 seulement au lieu d'une séquence entière de modèles. Contrairement à la section 2.3.3, la minimisation est effectuée sur tous ces modèles candidats, sans utiliser le Lasso ou toute autre méthode de présélection.

Table 2.1 – Étude de simulation sous l'hypothèse nulle : FWER observé par jeu de paramètres, moyenné sur 1000 simulations.

paramètres			$p = 10^2$	$p = 10^3$
Linéaire	$\rho = 0$	$n = 10^2$	.046	.045
		$n = 10^3$	.030	.030
		$n = 10^4$	.043	.040
	$\rho = 0.5$	$n = 10^2$	.039	.049
		$n = 10^3$	.033	.050
		$n = 10^4$	.032	.048
Binaire	$\rho = 0$	$n = 10^2$	.047	.045
		$n = 10^3$	.037	.037
		$n = 10^4$	.029	.046
	$\rho = 0.5$	$n = 10^2$	.039	.050
		$n = 10^3$	.036	.044
		$n = 10^4$	.029	.050

Plan de simulation

Nous avons simulé 1000 jeux de données dans chacun des 24 scénarios que nous avons explorés. Ils sont définis par les paramètres suivants :

- La famille du modèle de régression : linéaire ou binaire à fonction de lien logistique.
- Le nombre d'observations : $n = 10^2, 10^3$, or 10^4.
- Le nombre de regrésseurs : $p = 10^2$ or 10^3
- La matrice de corrélation utilisée pour simuler les régresseurs : une matrice de Toeplitz $\rho_{(i,j)} = \rho^{|i-j|}$, où $\rho = 0$ ou $\rho = 0.5$.

Nous avons simulé les covariables X_j suivant la loi normale centrée réduite, et la réponse Y, indépendamment des X_j, suivant la loi normale centrée réduite dans les scénarios de modèle linéaire et une loi de Bernoulli de probabilité 0.5 dans les scénarios de modèle binaire. Nous avons utilisé l'EAIC avec $\alpha = 0.05$.

Résultats de simulation

Les FWER que nous avons observés dans les 24 scénarios se répartissent de 0.029 à 0.05 (table 2.1), ce qui est conforme au niveau visé.

2.4.2 .Simulation de la procédure complète

Plan de simulation

Nous avons comparé les performances de l'EAIC à celles d'autres méthodes de sélection de variables fondées sur des critères d'information dans 308 scénarios jeux de paramètres du modèle linéaire généralisé $\mathrm{E}[Y \mid X] = f\left(\beta_0 + X\beta\right)$. Les scénarios sont définis par les paramètres suivants :

- Le type de modèle de régression : linéaire ou binaire logistique.
- Le nombre d' observations : $n = 10^2, 10^3$, ou 10^4.
- Le nombre de régresseurs : $p = 10^2, 10^3$, ou 10^4.
- La matrice de corrélation utilisée pour simuler les régresseurs : une matrice de Toeplitz $\rho_{(i,j)} = \rho^{|i-j|}$, avec $\rho = 0$ ou $\rho = 0.5$.
- Le rapport signal-bruit empirique. Cette quantité est inspirée du rapport signal-bruit utilisé dans Sabourin *et al.* (2015). Contrairement à ces auteurs nous utilisons une version empirique qui est définie pour chaque modèle linéaire généralisé. Il s'agit du rapport entre la variance empirique des signaux ($E_\beta[Y_i \mid X],\ i = 1, .., n$) et la moyenne empirique des variances des bruits ($\mathrm{Var}_\beta(Y_i \mid X),\ i = 1, .., n$) :

$$\mathrm{SNR}(X, \beta) = \frac{\frac{1}{n-1}\Sigma_{i=1}^{n}\left(E_\beta[Y_i \mid X] - \frac{1}{n}\Sigma_{i=1}^{n}E_\beta[Y_i \mid X]\right)^2}{\frac{1}{n}\Sigma_{i=1}^{n}\mathrm{Var}_\beta(Y_i \mid X)}.$$

  Un SNR plus élevé signifie que l'impact de chaque variable active sur Y est plus facilement observable. On peut s'attendre à ce que la sélection des variables soit plus efficace dans les scénarios où le SNR est élevé. Nous fixons $\mathrm{SNR}(X, \beta) = 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5$ou 10.

Nous avons simulé 1000 ensembles de données dans chacun de ces scénarios. Dans chaque simulation, nous avons d'abord généré la matrice des régresseurs X selon une loi normale standard (avec ou sans corrélation selon

le paramètre ρ). Ensuite, nous avons tiré 10 régresseurs actifs uniformément parmi les p régresseurs. Nous avons déterminé leurs coefficients β_j avec l'algorithme suivant :

1 Générer des coefficients non normalisés $\tilde{\beta}_j$ avec $P(\tilde{\beta}_j > 0) = \frac{1}{2}$ et des $|\beta_j|$ qui suivent une loi uniforme sur $[\frac{1}{4}, 1]$, indépendamment les uns des autres et de leur signe ;

2 Définir les coefficients normalisés β_0 et β_j par : $\beta_j = k(X)\tilde{\beta}_j$ et calculer $k(X)$ et β_0 via la résolution numérique d'équations (fonction `multiroot` dans le package R `rootSolve` version 1.8.2.3, Soetaert et Herman (2009)) de sorte que $\mathrm{SNR}(X, \beta)$ soit la valeur souhaitée.

Nous avons ensuite simulé le vecteur de résultat de la simulation y en suivant le modèle linéaire généralisé basé sur X et β simulés. Dans le modèle linéaire, $\mathrm{Var}_\beta(Y \mid X) = 1$.

À cause de difficultés computationnelles dans la simulation de X, les scénarios où $(p = 10^4, \rho = 0.5)$ ou $(p = 10^4, n = 10^4)$ ne sont pas inclus.

Comparaison des procédures

Pour chaque jeu de données simulé, nous avons exécuté l'algorithme Lasso à l'aide du package `glmnet` de R version 4.1-2 (Friedman *et al.*, 2010) avec les paramètres par défaut et un maximum de 100 degrés de liberté (paramètre `dfmax`). Nous avons appliqué sept méthodes différentes de sélection de variables qui consistent toutes à choisir l'un des modèles apparaissant sur le chemin du Lasso. Six méthodes sélectionnent le modèle qui minimise (globalement) un IC :

- l'AIC
- le BIC
- l'EBIC « original » de Chen et Chen (2008)

- une simplification de l'EBIC (Chen et Chen, 2012) :

$$\mathrm{EBIC}^s(B) = (\log n + 2\log p)\,|B| - 2l(B).$$

  Contrairement à l'EBIC original, ce critère s'inscrit dans la forme typique (2.1) d'un critère d'information.
- l'EAIC à $\alpha = 0.5$
- l'EAIC à $\alpha = 0.05$.

Ces méthodes nécessitent d'estimer la version non pénalisée de chacun des modèles sur le chemin du Lasso, puis d'utiliser leur log-vraisemblance dans le calcul des critères d'information. La septième méthode, que nous appelons « oracle », suppose connu le fait qu'il y a exactement 10 variables actives. Elle consiste à sélectionner le premier modèle comportant au moins 10 variables sur le chemin du Lasso. Son intérêt est de donner une idée des meilleures performances que l'on peut attendre d'une sélection de variables sur le chemin du Lasso dans chaque scénario.

Nous avons calculé des estimations du FWER, du taux de fausses découvertes (FDR) et de la sensibilité de chacune de ces méthodes dans chacun des 308 scénarios en moyennant les résultats obtenus sur les 1000 jeux de données simulés.

Bien que nous nous soyons en général concentrés sur le minimum global des différents IC sur le chemin du Lasso, nous avons également exploré la variante décritye en 2.3.3, en comparant le premier minimum local des EAIC aux niveaux $\alpha = 0.5$ et 0.05 à leur minimum global dans les scénarios linéaires non corrélés.

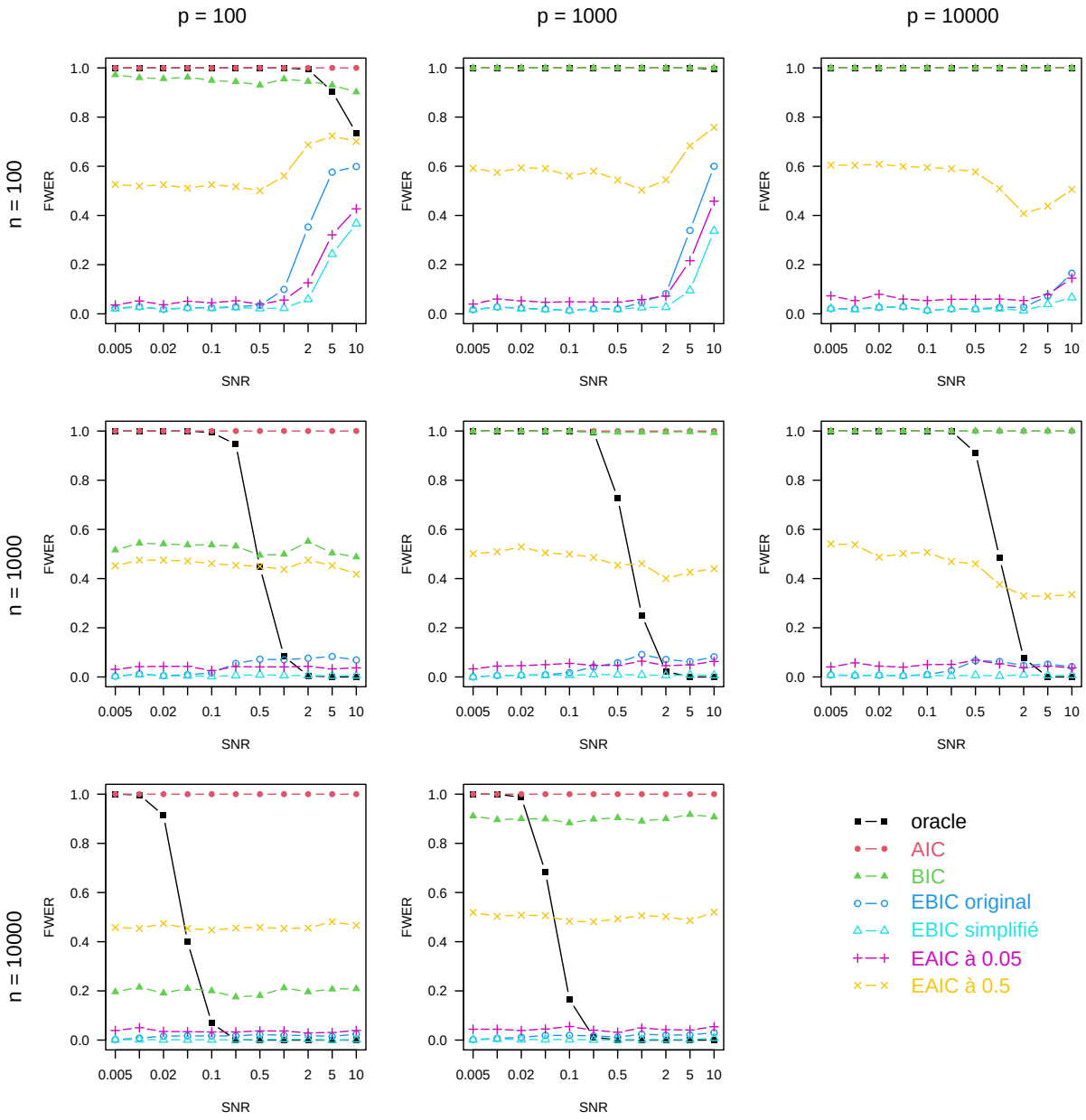

Figure 2.1 – Étude de simulation de la procédure complète : FWER par paramètre, moyenné sur 1000 simulations. Modèle linéaire, $\rho = 0$.

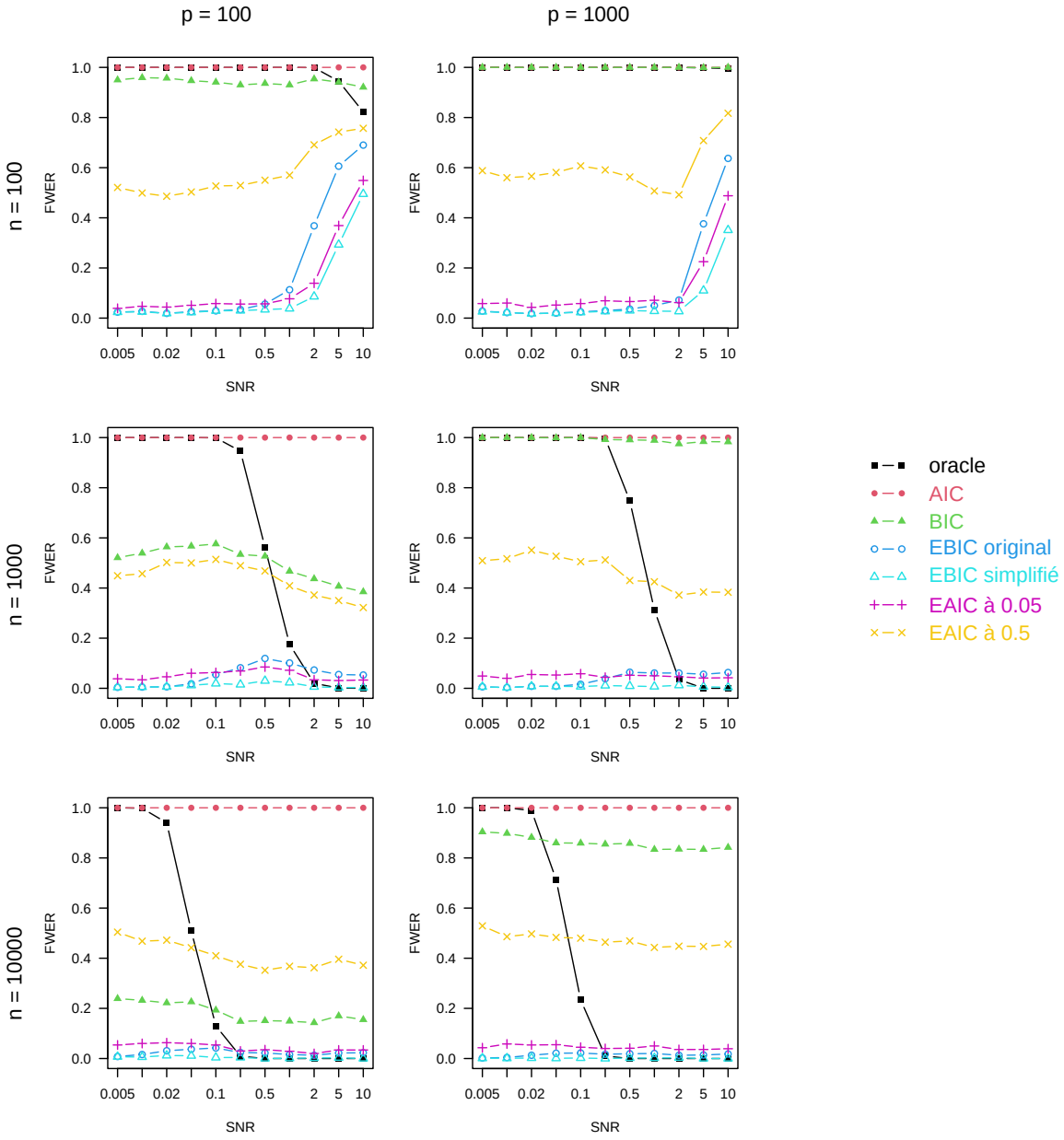

Figure 2.2 – Étude de simulation de la procédure complète : FWER par paramètre, moyenné sur 1000 simulations. Modèle binaire, $\rho = 0$.

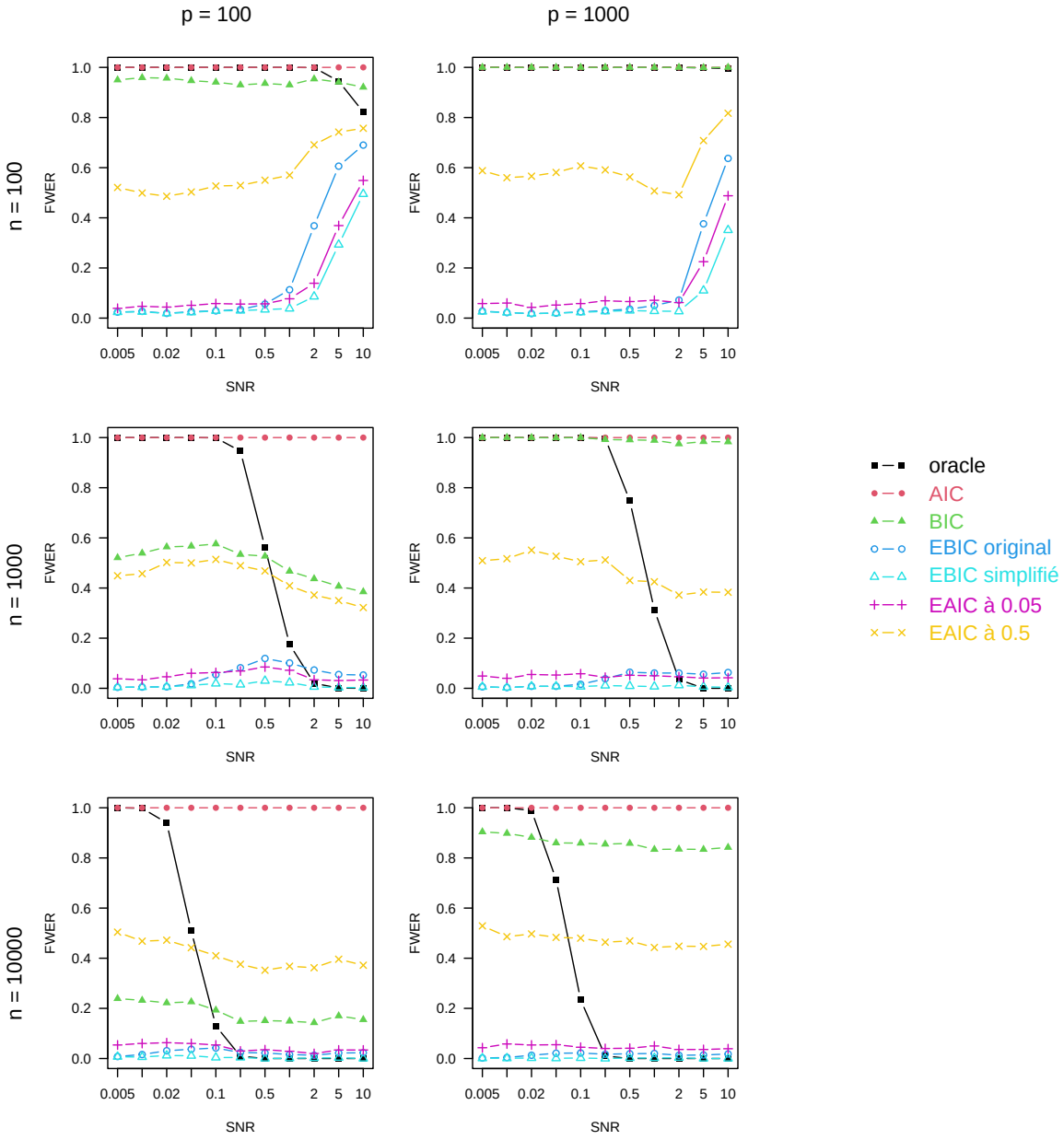

Figure 2.3 – Étude de simulation de la procédure complète : FWER par paramètre, moyenné sur 1000 simulations. Modèle linéaire, $\rho = 0.5$.

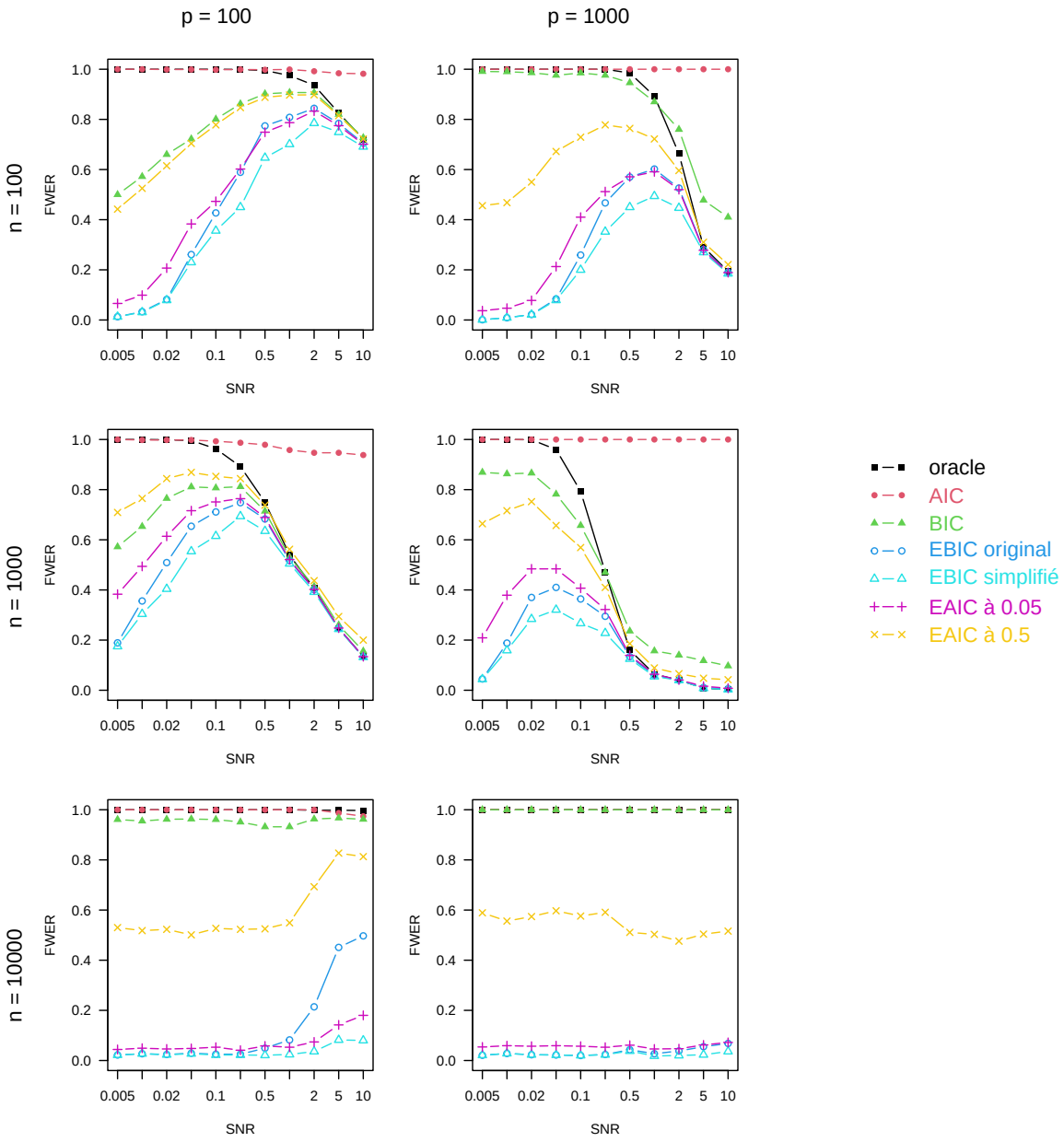

Figure 2.4 – Étude de simulation de la procédure complète : FWER par paramètre, moyenné sur 1000 simulations. Modèle binaire, $\rho = 0.5$.

Résultats

Pour chaque combinaison de n, p, ρ et de la famille du modèle, nous avons tracé le graphique du FWER de chaque méthode (figures 2.1 à 2.4), le nombre de vrais positifs (figures A.1 à A.4 en annexe) et le FDR (figures A.5 à A.8 en annexe) variant avec le SNR.

Dans les figures 2.1 à 2.4, le FWER de l'oracle indique les modèles dans lesquels il est possible de capturer les 10 variables actives exactes en s'arrêtant à un nœud du chemin du Lasso. Lorsque $n = 100$ ou que le SNR est faible, cela n'est presque jamais possible puisque le FWER de l'oracle est égal ou proche de 1. Dans les scénarios non corrélés avec $n \geq 100$ et un SNR suffisamment élevé (Figures 2.1 et 2.2), le FWER de l'oracle est proche de 0, ce qui montre que l'ensemble des 10 variables véritablement actives se trouve sur le chemin du Lasso.

Le FWER de l'AIC, presque toujours égal à 1, montre que cette méthode produit presque toujours des faux positifs.

Dans les scénarios sans corrélation (figures 2.1 et 2.2), le FWER du BIC tend à ne pas dépendre du rapport signal-bruit (sauf dans les modèles binaires à SNR élevé et à n élevé, où il y a une diminution), mais il dépend fortement de n et de p. Dans les modèles linéaires avec peu d'observations ($n = 100$) ou beaucoup de variables ($p \geq 1000$), il est égal ou proche de 1, ce qui signifie que le BIC produit presque toujours des faux positifs. Ce n'est que lorsque n est grand et p est petit que le BIC a un FWER modéré.

Les figures 2.1 et 2.2 montrent que, conformément à sa base théorique, l'AIC étendu contrôle approximativement le FWER au niveau souhaité dans la plupart des scénarios non corrélés. Les seules exceptions sont les scénarios avec à la fois peu d'observations ($n = 100$), un SNR élevé et un nombre modéré de variables ($p \leq 1000$ dans le modèle linéaire, $p = 100$ dans le modèle

binaire). Dans ces conditions, le SNR élevé et le p modéré impliquent qu'une grande partie des variables actives sont capturées (comme le montrent les courbes de sensibilité des figures A.1 et A.2). L'EAIC est donc utilisé pour comparer des modèles d'une taille de l'ordre de 10. A cette taille, la convergence de la distribution garantie par le théorème de Wilks est plus lente que lors de la comparaison de modèles de très petite taille. Par conséquent, à $n = 100$, le régime asymptotique n'est pas atteint et la base théorique du contrôle du FWER par l'EAIC n'est pas vérifiée. De plus, dans les scénarios non corrélés, le FWER de l'EAIC ne descend pas beaucoup en dessous de sa valeur nominale, sauf dans les modèles binaires à fort SNR (figure 2.2).

Le FWER de l'EBIC dans sa formule originale est généralement faible, sauf dans les scénarios où le contrôle du FWER par l'EAIC échoue également : $n = 100$, SNR élevé et nombre modéré de variables. Il présente une allure similaire au FWER de l'EAIC à 0.05 avec plus de variation, étant parfois proche de 0 et ayant des valeurs maximales plus élevées. L'EBIC simplifié est la méthode la plus conservatrice de toutes, avec des schémas similaires à l'EBIC original, mais un FWER plus faible.

Dans les scénarios avec corrélation (figures 2.3 et 2.4), le FWER varie un peu plus avec le SNR pour toutes les méthodes, à l'exception de l'AIC (qui a un FWER toujours proche de 1). Lorsque $n = 100$, les FWER ont un profil similaire à ceux observés dans des scénarios non corrélés, avec un pic encore plus élevé pour un fort SNR et un nombre modéré de variables. Lorsque $n \geq 1000$, il y a un certain écart par rapport au contrôle du FWER qui a été observé dans les paramètres non corrélés. Pour tous les paramètres $n \geq 1000, \rho = 0.5$, l'EAIC à $\alpha = 0.5$ atteint un FWER maximal de 0.612, l'EAIC à $\alpha = 0.05$ atteint 0.251, et l'EBIC d'origine atteint 0.288. Dans les scénarios avec corrélation les plus favorables, à n et SNR élevés (lorsque le modèle correct se trouve presque

toujours sur le chemin du Lasso, comme le montre le FWER de l'oracle proche de 0), l'EAIC contrôlent le FWER à son niveau nominal ou en dessous.

Les figures A.1 à A.4 montrent le nombre de vrais positifs capturés par chaque méthode dans chaque scénario, ce qui correspond à 10 fois leur sensibilité puisque tous les scénarios sont simulés avec 10 variables actives. Les sensibilités augmentent avec le SNR et avec n, mais elles diminuent avec p. Étant donné que les EBIC (en particulier la version simplifiée) sont les méthodes les plus conservatrices en termes de FWER, elles sont également les moins sensibles. L'EAIC à $\alpha = 0.05$ a une sensibilité proche de celle de l'EBIC original. Cependant, deux types de paramètres où elles diffèrent en termes de sensibilité suggèrent un avantage de l'EAIC à 0.05 par rapport à l'EBIC :

- Dans les scénarios à n élevé et faible SNR, la sensibilité de l'EAIC est supérieure à celle de l'EBIC. Il présente par ailleurs un FWER légèrement plus élevé mais le contrôle toujours à son niveau nominal.
- Au contraire, dans les scénario à n faible et SNR élevé, l'EBIC présente une sensibilité plus élevée que l'EAIC. Le prix à payer est un FWER plus élevé, qui atteint ses valeurs les plus élevées dans ces scénarios.

Les courbes de FDR (figures A.5 à A.8) montrent la même hiérarchie de méthodes, l'AIC étant le moins conservateur et les EBIC étant en moyenne les plus conservateurs. Bien que l'EAIC à $\alpha = 0.05$ ait un FDR plus élevé que l'EBIC original moyenné sur l'ensemble des paramètres, son FDR est plus stable, étant de 0.083 au maximum dans les scénarios sans corrélation et de 0.126 dans les scénarios avec corrélation, tandis que l'EBIC atteint 0.128 dans un scénario sans corrélation et 0.171 dans un scénario avec corrélation. De même, l'EAIC à $\alpha = 0.5$, bien qu'il ne soit généralement pas conservateur, a un FDR qui n'approche jamais 1 (FDR maximal égal à 0.623), contrairement au BIC qui a un FDR proche de 1 dans les scénarios où n est faible, p est élevé et le SNR est

faible.

Les figures A.9 à A.11 en annexe montre que le premier minimum local coïncide pratiquement avec le minimum global si $n \geq 1000$. En $n = 100$, cependant, la différence est notable dans les scénarios à fort SNR et le premier minimum local parvient à y éviter la perte de contrôle du FWER, au prix d'une perte notable de sensibilité.

2.5 .Application aux données de pharmacovigilance

Pour illustrer le comportement des méthodes de sélection de variables sur des données réelles, nous les avons appliquées à la base nationale de pharmacovigilance (BNPV). Nous avons utilisé le même prétraitement des données que celui décrit dans Courtois *et al.* (2021), produisant une base de données de $n = 452\,914$ notifications spontanées d'effets indésirables des médicaments du 1er janvier 2000 au 29 décembre 2017 avec 6617 effets indésirables différents (codés au niveau Preferred Term du Medical Dictionary for Regulatory Activities, MedDRA) et $p = 1692$ médicaments différents (codés au 5ème niveau de la hiérarchie Anatomical Therapeutic Chemical) signalés au moins 10 fois. Nous nous sommes concentrés sur un résultat binaire, l'événement indésirable « pathologie hépatique d'origine médicamenteuse » (*Drug-Induced Liver Injury*, DILI), qui est l'un des évènements indésirables les plus fréquents avec 25187 occurrences soit $5,56\%$ de l'ensemble des notifications spontanées. Nous avons utilisé un modèle de régression logistique sur les expositions médicamenteuses qui constituent des covariables binaires. On considère qu'il y a un signal de pharmacovigilance lorsqu'une variable est sélectionnée avec un coefficient estimé positif.

Pour évaluer les performances des méthodes, comme Courtois *et al.* (2021), nous avons utilisé l'ensemble de référence DILIrank de signaux de pharma-

Table 2.2 – Performance de chaque méthode sur les données de la BNPV en termes de nombre de signaux de pharmacovigilance (variables ayant une association positive aux DILI), proportion de faux positifs (FDP), spécificité et sensibilité.

Méthode	Signaux	Signaux à statut connu	Faux positifs	FDP (%)	Spécificité (%)	Sensibilité (%)
AIC	187	69	5	7.2	97.5	48.1
BIC	170	65	5	7.7	97.5	45.1
EAIC à 50%	170	65	5	7.7	97.5	45.1
EAIC à 10%	150	59	4	6.8	98.0	41.4
EAIC à 5%	142	55	2	3.6	99.0	39.8
EBIC original	142	55	2	3.6	99.0	39.8
EBIC simplifié	112	47	2	4.3	99.0	33.8

covigilance connus concernant les pathologies hépatiques d'origine médicamenteuse Chen *et al.* (2016). Il comprend 203 témoins négatifs (médicaments connus pour ne pas être associés à des DILI) et 133 positifs (médicaments connus pour être associés à des DILI). Ce choix d'ensemble de référence pourrait lui-même être discuté, par exemple le premier faux positif (du point de vue de DILIrank) à apparaître sur le chemin du Lasso est la daunorubicine (de code ATC L01DB02) dont la possible association à des DILI est discutée dans la base LiverTox (noa, 2012).

Nous avons mis en œuvre cinq méthodes qui minimisent un critère d'information sur le chemin du Lasso : l'AIC, le BIC, l'EAIC à différents niveaux de α, l'EBIC original et l'EBIC simplifié. Le tableau 2.2 présente les résultats de ces méthodes, y compris l'EAIC à 50%, 10% et 5%. La proportion de fausses découvertes (FDP), la spécificité et la sensibilité ont été calculées sur des médicaments dont le statut était connu. Comme dans la plupart des simulations, l'AIC a généré le plus de signaux et l'EBIC simplifié en a généré le moins. Les sensibilités des méthodes reflètent cette hiérarchie, allant de 48.1% pour l'AIC à 33.8% pour l'EBIC simplifié. L'EAIC à $\alpha = 0.05$ et l'EBIC original ont sélectionné

le même modèle et ont donc les mêmes caractéristiques sur cet ensemble de données. Avec seulement deux faux positifs connus, ils avaient la meilleure spécificité avec l'EBIC simplifié (99.0%), et le meilleur FDP (3.6%). À $\alpha = 0.5$, l'EAIC sélectionne le même modèle que le BIC (le niveau auquel l'EAIC et le BIC sont mathématiquement équivalents étant $\alpha \approx 0.406$ pour ces valeurs de n et de p).

2.6 .Discussion

L'AIC et le BIC, qui sont les critères d'information les plus classiques, sont souvent utilisés pour la sélection d'un modèle parcimonieux dans un contexte de grande dimension. Or en grande dimension, ils ne possèdent pas de propriétés de consistance et ont tendance à sélectionner des modèles comportant de nombreux faux positifs, comme le confirme notre étude de simulation approfondie. Cependant, nous montrons qu'il est possible de limiter le nombre de faux positifs en grande dimension tout en utilisant un critère d'information, à condition que ce critère tienne compte de la dimensionnalité. C'est le cas du BIC étendu et de notre proposition, l'AIC étendu.

Contrairement à l'EBIC, l'EAIC est conçu pour contrôler le FWER à un niveau spécifié. Notre résultat mathématique (lemme 2.2.3) laisse penser qu'il y parvient lorsque les variables candidates ne sont pas corrélées entre elles et que n et p sont suffisamment grands pour que les approximations asymptotiques soient valables. En pratique, les simulations montrent que le FWER de l'EAIC est en effet proche de son niveau spécifié dans presque tous les scénarios non corrélés (sur une large gamme de n et p, dans les modèles linéaires et binaires), les seules exceptions étant les configurations où à la fois n et p sont petits et les variables actives ont un effet important.

En outre, les critères étendus atteignent des performances satisfaisantes

en termes de FDR, contrairement aux critères classiques. Nos simulations montrent qu'alors que l'AIC et le BIC présentent des situations de grave défaillance, c'est-à-dire des scénarios où leur FDR approche 1, ce n'est pas le cas de l'EAIC ou de l'EBIC. Même lorsque les variables candidates sont fortement corrélées, leur FDR est toujours inférieur à 0.3 (dans le cas de l'EBIC et de l'EAIC à $\alpha = 0.05$) ou au maximum d'environ 0.6 (dans le cas de l'EAIC à $\alpha = 0.5$).

L'EAIC ou l'EBIC utilisés en combinaison avec le Lasso, de même que l'AIC-Lasso et le BIC-Lasso, sont des méthodes déterministes, simples et rapides. Elles demandent de réaliser le Lasso une seule fois, puis d'ajuster une suite de modèles de régression non pénalisés de dimension modérée, sans avoir à ré-échantillonner ou à simuler des données contrairement, par exemple, à la validation croisée (voir la section 1.3.1), stability selection (1.3.3) ou permutation selection (1.3.3). Bien que nos simulations se soient fondées sur le Lasso, une autre procédure de présélection peut être utilisée à sa place tant qu'elle fournit une petite famille de modèles candidats peu denses. C'est le cas d'autres régressions pénalisées telles que SCAD (Fan et Li, 2001), ou elastic net Zou et Hastie (2005).

La formule de l'EAIC elle-même est plus simple que celle de l'EBIC. De même, comparée à celle de l'EBIC, la paramétrisation de l'EAIC est directement interprétable car α est le FWER que l'on obtient dans un cadre non corrélé.

Par conséquent, nous recommandons en cas de besoin d'une méthode simple de sélection de variables en grande dimension basée sur le Lasso ou une autre méthode de présélection, et dont on veut s'assurer qu'elle ne sélectionne pas principalement des faux positifs, d'utiliser l'un des deux critères qui s'adaptent à la dimension du problème : l'EAIC ou l'EBIC. L'EAIC est à préférer si un niveau spécifique de FWER est ciblé.

L'EAIC exploite le nombre p de variables candidates en le considérant comme

une approximation du nombre de tests du rapport de vraisemblance implicitement réalisés, et en faisant l'hypothèse exigeante de l'indépendance de ces tests. En raison de la non indépendance de ces tests en pratique due principalement à la corrélation entre variables, on peut envisager de remplacer dans l'EAIC p par une mesure du « nombre effectif » de tests indépendants auxquels les p tests non indépendants sont approximativement équivalents du point de vue de la correction de la multiplicité des tests. Cette notion est utilisée dans les tests multiples en analyse génomique (Galwey, 2009; Li *et al.*, 2012). Une autre piste d'amélioration serait l'analyse plus systématique, et éventuellement la généralisation à d'autres contextes que le chemin du Lasso, de la sélection de modèle par premier minimum local de l'EAIC ou d'un autre critère d'information.

3 - Test par simulation-calibration

Nous proposons une procédure de test qui répond au même problème que le test de Lockhart *et al.* (2014) : produire, pour chaque variable sélectionnée par le Lasso, une p-value qui mesure sa significativité tout en corrigeant les inconvénients de ce test : non-validité dans les cas non linéaires, et parfois manque de puissance.

Nous conservons l'idée d'exploiter, pour chaque variable, la valeur du paramètre de régularisation à laquelle elle apparaît sur le chemin du Lasso. Toutes choses égales par ailleurs, plus une variable est sélectionnée « tôt » sur le chemin du Lasso — c'est-à-dire à une valeur de λ élevée — plus elle est associée à la réponse donc plus elle est significative. Comme Lockhart et al., nous proposons de tester dans le formalisme des tests d'hypothèse, pour chacune des variables apparaissant sur le chemin du Lasso, si son λ de sélection est plus élevé qu'il ne devrait être si il n'y avait pas de lien entre cette variable et la variable réponse.

Pour cela, notre proposition s'inspire de la sélection par permutation de Sabourin *et al.* (2015). Nous reprenons l'idée de générer des données qui suivent la même distribution que les données à analyser, mais dans lesquelles on a artificiellement rendu les variables explicatives indépendantes de la réponse. Comme dans la sélection par permutation, nous effectuons la régression Lasso de la variable réponse observée sur ces données simulées et nous nous intéressons à λ_0, la plus grande valeur de λ où est sélectionnée une variable explicative qui, par construction, est en réalité indépendante de la réponse. Ces λ_0 obtenus sur plusieurs jeux de données simulées forment une population de référence représentative du cas d'indépendance entre la réponse et

les covariables. Contrairement à la sélection par permutation, qui consiste à appliquer le λ_0 médian de cette population aux données d'intérêt, nous comparons le λ_0 obtenu sur les données d'intérêt à cette population de référence pour produire un test de la significativité de ce λ_0, en estimant la p-value du test par Monte-Carlo à partir de la population de référence. Cela se généralise à l'interprétation d'autres quantités que λ_0 : au λ d'apparition de toute variable sur le chemin du Lasso, pas seulement la première. Là où Sabourin et al. conservent la répartition des vecteurs réponse simulés tout en cassant leur association à tous les X_j au moyen d'une permutation, nous conservons au contraire l'association à certaines des covariables en appliquant une « calibration » à des vecteurs réponse simulés, d'où le nom de test par simulation-calibration.

Dans une optique de sélection de variables, si le test conclut à la significativité du λ d'apparition d'une variable sur le chemin du Lasso, nous sélectionnons cette dernière. En itérant le test par simulation nous obtenons donc une procédure complète de sélection d'un modèle, présentée en section 3.9.

3.1 .Position du problème et notations

Soit $X \in \mathbb{R}^{n \times p}$ une matrice de p covariables et n observations et $y \in \mathbb{R}^n$ un vecteur réponse de taille n. Nous considérons le modèle de régression linéaire :

$$y = \beta_0 + X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right) \tag{3.1}$$

où $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p$ et $\sigma > 0$. Soit $\hat{\beta}^{Lasso}(\lambda, y)$ l'estimateur de β par Lasso, qui dépend du paramètre de régularisation λ et de la réponse y. Il dépend aussi de la matrice des covariables X, mais celle-ci est dans la suite considérée

comme une constante.

Soit A un sous-ensemble de $1, .., p$. Nous supposons que les covariables $X_j, j \in A$, sont actives, et nous cherchons à savoir si d'autres covariables sont actives. Nous définissons donc les hypothèses nulle et alternative, qui dépendent de A :

$$H_0(A) : \forall j \notin A, \beta_j = 0$$
$$H_1(A) : \exists j \notin A, \beta_j \neq 0$$

Bien que A soit *a priori* un sous-ensemble quelconque de $\{1, .., p\}$, l'idée est, comme dans Lockhart et al., de prendre pour A l'ensemble des indices des $k-1$ premières covariables apparaissant sur le chemin du Lasso, pour des valeurs de k relativement petites (même si p peut être élevé). Le cas particulier le plus simple est $A = \emptyset$.

Nous considérons la statistique suivante :

$$\lambda_A(y) = \sup\{\lambda \geq 0 : \exists j \notin A, \hat{\beta}_j^{Lasso}(\lambda, y) \neq 0\}$$

C'est-à-dire le λ d'apparition de la première variable sélectionnée en dehors de A, soit la k-ième variable sélectionnée si A regroupe les $k-1$ premières variables. Nous rejetterons l'hypothèse nulle si l'on observe une valeur trop élevée de λ_A par rapport à ce qui est attendu sous l'hypothèse nulle. Une valeur anormalement élevée du λ d'apparition d'une variable est typiquement le résultat d'une association entre la réponse et cette variable. C'est pourquoi un test de $H_0(A)$ fondé sur λ_A peut, en première approximation, être interprété comme évaluant la significativité de la première variable sélectionnée par le Lasso en dehors de A, dont nous notons l'indice j_A.

Il est cependant possible que $H_0(A)$ soit fausse sans que la variable j_A

soit active, si il existe une variable active n'appartenant pas à A qui est sélectionnée « plus tard » sur le chemin du Lasso, à un λ inférieur à λ_A. Dans ce cas, il est correct rejeter $H_0(A)$ d'un point de vue de test d'hypothèse, mais incorrect de conclure que la variable j_A est significative. Dans la partie 3.10, nous montrerons dans quelles conditions la probabilité de cet évènement est contrôlée.

Par ailleurs, il est possible que plusieurs variables apparaissent en même temps sur le chemin du Lasso au paramètre λ_A, que ce soit mathématiquement (par exemple en cas de X et y tous deux binaires il est possible, et de probabilité non nulle, qu'il y ait une symétrie exacte entre les associations de y à deux covariables distinctes) ou approximativement si la différence entre les λ d'apparition de deux variables distinctes est trop petite pour être mesurée par `glmnet`. Dans les deux cas, le rejet de $H_0(A)$ entraîne la sélection de l'ensemble de ces variables.

3.2 .Le problème du calcul de la p-value

Si λ_A était utilisée comme statistique de test, sa p-value associée serait, par définition :

$$p_A^0(y) = \mathrm{P}_{H_0(A)}\left(\lambda_A(Y) \geq \lambda_A(y)\right)$$

où Y est une variable aléatoire qui suit le modèle 3.1.

Il n'existe en général pas d'expression analytique explicite simple du résultat d'une régression Lasso. Cela empêche de déterminer exactement la distribution de $\lambda_A(y)$ et donc de calculer $p_A^0(y)$ ou d'autres quantités similaires. C'est pourquoi nous cherchons plutôt à estimer cette probabilité par la méthode de Monte-Carlo. Nous voudrions appliquer les étapes suivantes :

1. Simuler des vecteurs Y indépendants qui suivent le modèle 3.1 et satisfont $H_0(A)$
2. Effectuer la régression Lasso de ces Y sur X et calculer leur $\lambda_A(Y)$
3. Mesurer la proportion de vecteurs Y vérifiant $\lambda_A(Y) \geq \lambda_A(y)$.

Or, la première de ces étapes n'est pas possible sans information supplémentaire car même en supposant vraie $H_0(A)$, les paramètres du modèle 3.1 ne sont pas connus : $H_0(A)$ garantit la nullité des $\beta_j, j \notin A$ mais ne dit rien sur la valeur des $\beta_j, j \in A$ ni sur celle de σ. Une tentative naïve de réaliser cette étape, où l'on remplace simplement β_0, les $\beta_j, j \in A$ et σ par leurs valeurs estimées à partir des données, conduit à biaiser l'estimation de la p-value (voir la figure 3.1, section 3.11.2).

C'est pourquoi, à la place de $p_A^0(y)$, nous allons considérer une variante dont la définition comportera toutes les informations nécessaires à son estimation par Monte-Carlo.

3.3 .p-value définie conditionnellement

La solution que nous proposons est d'introduire un conditionnement par les paramètres qui restent inconnus sous $H_0(A)$. Considérons le modèle linéaire suivant :

$$y = X_A\beta_A + \epsilon_A, \quad \epsilon_A \sim \mathcal{N}(0, \sigma_A^2 I_n) \tag{3.2}$$

où X_A est la matrice composée d'un vecteur colonne de 1 et des colonnes de X dont les indices apartiennent à A, et β_A est le vecteur $(\beta_j)_{j\in\{0\}\cup A}$ (Nous utilisons donc pour ce modèle la formulation matricielle compacte introduite en 1.1).

Il s'agit de la forme réduite que prend le modèle 3.1 lorsque $H_0(A)$ est

vraie. On a alors $\sigma = \sigma_A$. On appelle $\theta_A = (\beta_A, \sigma_A)$ le vecteur de paramètres du modèle 3.2, $\Theta_A = \mathbb{R}^{1+|A|} \times \mathbb{R}_+$ l'espace de ses valeurs possibles, et $\widehat{\theta_A}(y)$ l'estimateur par maximum de vraisemblance (non-pénalisée) de θ_A.

Nous définissons la statistique de test suivante :

Définition 3 (statistique p_A)**.**

$$p_A(y) = \mathrm{P}_{H_0(A)}\left(\lambda_A(Y) \geq \lambda_A(y) | \widehat{\theta_A}(Y) = \widehat{\theta_A}(y)\right).$$

Cette formule reprend celle d'une p-value en y ajoutant un conditionnement, ce qui fait de $p_A(y)$ une *p-value conditionnelle*.

$p_A(y)$ est définie de sorte que l'on puisse connaître aussi précisément que possible sa loi de probabilité. Pour cela, nous avons besoin du lemme suivant, qui est un résultat général sur la fonction de répartition d'une variable aléatoire réelle.

Lemme 2. *Soit U une variable aléatoire réelle et F sa fonction de répartition. Alors :*

$$\forall t \in [0,1], \mathrm{P}\left(F(U) \leq t\right) \leq t,$$

et si U suit une loi continue, la probabilité est égale à t.

Dans le cas général d'inégalité, on dit que la loi de $F(U)$ *domine stochastiquement* la loi uniforme sur $[0,1]$. Ce terme issu de la théorie de la décision signifie que la fonction de répartition de la loi de probabilité dite dominante est inférieure en tout point à celle de la loi dite dominée, et donc que la loi dominante est systématiquement décalée vers les valeurs élevées par rapport à la loi dominée.

Dans le cas d'égalité, la loi de $F(U)$ est simplement la loi uniforme sur $[0,1]$, car leurs fonctions de répartition sont égales.

Démonstration. Prouvons d'abord l'inégalité. Soit $t \in [0,1]$ et soit l'ensemble $I = \{u \in \mathbb{R} : F(u) \leq t\}$. On veut prouver que $\mathrm{P}(U \in I) \leq t$. F étant croissante, I est un intervalle non borné à gauche.

L'inégalité est évidemment vraie si $t = 1$. On peut donc supposer que $t < 1$. Comme $\lim_{+\infty} F = 1$, on a $\exists u \in \mathbb{R} : F(u) > t$.

Si $t = 0$ et F est strictement positive, alors $\forall u, F(u) > t$ donc $\mathrm{P}\left(F(U) \leq t\right) = 0$ donc l'inégalité est vraie. On peut donc supposer que soit $\exists u \in \mathbb{R} : F(u) = 0$, soit $t > 0$. Dans le premier cas, un tel u est un élément de I. Dans le second cas, comme $\lim_{-\infty} F = 0$, on a $\exists u \in \mathbb{R} : F(u) \leq t$ c'est-à-dire un élément de I.

Hors des cas triviaux, I est donc un ensemble majoré non vide. Il possède donc une borne supérieure réelle, notée u^+. $I \subset]-\infty, u^+]$ donc

$$\mathrm{P}(U \in I) \leq \mathrm{P}(u \leq u^+) = F(u^+).$$

Donc si $F(u^+) \leq t$, alors on a bien $\mathrm{P}(U \in I) \leq t$.

Si au contraire $F(u^+) > t$, alors $u^+ \notin I$ par définition de I. u^+ étant une borne supérieure, $I =]-\infty, u^+[$ et il existe une suite croissante $(u_p)_{p\in\mathbb{N}}$ d'éléments de I tels que que $\lim_{p\to\infty} u_p = u^+$. Donc $I = \cup_p]-\infty, u_p]$. S'agissant d'une union croissante d'ensembles mesurables, on a

$$\begin{aligned} \mathrm{P}(U \in I) &= \lim_{p\to\infty} \mathrm{P}(U \in]-\infty, u_p]) \\ &= \lim_{p\to\infty} F(u_p). \end{aligned}$$

Or par appartenance à I, $\forall p, F(u_p) \leq t$, donc par passage à la limite :

$$\mathrm{P}(U \in I) \leq t.$$

L'inégalité étant prouvée, déduisons-en l'égalité dans le cas continu. En notant G la fonction de répartition de $-U$, on a $\forall u \in \mathbb{R}$:

$$\begin{aligned} G(-u) &= \mathrm{P}(-U \leq -u) \\ &= \mathrm{P}(U \geq u) \\ &= \mathrm{P}(U > u) + \mathrm{P}(U = u) \\ &= 1 - F(u) + \mathrm{P}(U = u). \end{aligned}$$

Si la loi de U est continue, alors $\mathrm{P}(U = u) = 0$ donc $G(-u) = 1 - F(u)$. Appliquons maintenant l'inégalité du lemme à la variable $-U$ et au seuil $1-t$:

$$\begin{aligned} \mathrm{P}\left(G(-U) \leq 1 - t\right) &\leq 1 - t \\ \mathrm{P}\left(1 - F(U) \leq 1 - t\right) &\leq 1 - t \\ \mathrm{P}\left(F(U) \geq t\right) &\leq 1 - t \\ \mathrm{P}\left(F(U) < t\right) &\geq t \text{ (évènement complémentaire)} \\ \mathrm{P}\left(F(U) \leq t\right) &\geq t \text{ (évènement contenant le précédent)} \end{aligned}$$

Ce qui, combiné avec l'inégalité du lemme, donne $\mathrm{P}\left(F(U) \leq t\right) = t$.

□

Cela permet d'obtenir la propriété suivante portant sur $p_A(y)$:

Lemme 3. *Sous $H_0(A)$, la loi de $p_A(Y)$ domine stochastiquement la loi uniforme sur $[0, 1]$:*

$$\forall t \in [0, 1], \mathrm{P}_{H_0(A)}\left(p_A(Y) \leq t\right) \leq t.$$

Démonstration. Décomposons la quantité à majorer par intégration sur Θ_A,

l'ensemble des valeurs que peut prendre $\widehat{\theta_A}(Y)$:

$$\begin{aligned} \forall t \in [0,1], \quad & \mathrm{P}_{H_0(A)} \quad (p_A(Y) \leq t) = \\ & \int_{\theta'_A \in \Theta_A} \mathrm{P}_{H_0(A)} \left(p_A(Y) \leq t | \widehat{\theta_A}(Y) = \theta'_A \right) d\mathrm{P}_{H_0(A)} \left(\widehat{\theta_A}(Y) = \theta'_A \right). \end{aligned} \tag{3.3}$$

Pour tout vecteur de paramètres θ'_A, notons $F_{\theta'_A}$ la fonction de répartition de la loi de $-\lambda_A(Y)$ conditionnellement à $\widehat{\theta_A}(Y) = \theta'_A$. Alors pour tout y,

$$p_A(y) = \mathrm{P}_{H_0(A)} \left(-\lambda_A(Y) \leq -\lambda_A(y) | \widehat{\theta_A}(Y) = \widehat{\theta_A}(y) \right) = F_{\widehat{\theta_A}(y)} \left(-\lambda_A(y) \right).$$

On applique maintenant le lemme 2 en prenant pour loi de U la loi conditionnelle de $-\lambda_A(Y)$. La probabilité obtenue est donc conditionnelle :

$$\forall t \in [0,1], \forall \theta'_A \in \Theta_A, \ \mathrm{P}_{H_0(A)} \left(p_A(Y) \leq t | \widehat{\theta_A}(Y) = \theta'_A \right) \leq t. \tag{3.4}$$

Cette inégalité s'injecte dans l'intégrale 3.3 :

$$\mathrm{P}_{H_0(A)}(p_A(Y) \leq t) \leq \int_{\theta'_A \in \Theta_A} t \, d\mathrm{P}_{H_0(A)} \left(\widehat{\theta_A}(Y) = \theta'_A \right) = t.$$

□

En pratique, la loi de $\lambda_A(Y)$ conditionnellement à $\widehat{\theta_A}(Y) = \theta'_A$ est typiquement continue pour tout vecteur de paramètres à écart-type non nul, c'est-à-dire tout $\theta'_A \in \Theta^*_A$ avec $\Theta^*_A = \mathbb{R}^{|A|} \times \mathbb{R}^*_+$. On appellera *hypothèse de continuité* le fait de supposer cela.

Cette hypothèse correspond à l'essentiel des cas pratiques car en général, on peut faire varier localement Y tout en maintenant $\widehat{\theta_A}(Y) = \theta'_A$, et cela fait varier chacune des covariances empiriques $\widehat{\mathrm{Cov}}(Y, X_j)$. En particulier, cela fait varier $\widehat{\mathrm{Cov}}(Y, X_{j_A})$, la covariance empirique avec la variable sélectionnée par le Lasso à $\lambda_A(Y)$, et donc fait varier $\lambda_A(Y)$ qui est localement une fonction

strictement croissante de $\left|\widehat{\mathrm{Cov}}(Y, X_{j_A})\right|$. Il n'y a donc aucune constante λ qui soit atteinte par $\lambda_A(Y)$ de façon constante sur un ensemble suffisamment important de valeurs de Y pour que $P\left(\lambda_A(Y) = \lambda\right) > 0$.

L'hypothèse de continuité permet d'obtenir des résultats plus simples. Ainsi :

Corollaire 3.1. *Sous l'hypothèse de continuité,* $p_A(Y)$ *suit une loi uniforme sur* $[0, 1]$.

Démonstration. La preuve est identique à celle du lemme 3 à deux adaptations près :

- Les intégrales sont sur Θ_A^* au lieu de Θ_A, ce qui donne le même résultat car $\mathrm{P}\left(\widehat{\theta_A}(Y) \notin \Theta_A^*\right) = \mathrm{P}(\widehat{\sigma_A}(Y) = 0) = 0$.
- Pour tout $\theta'_A \in \Theta_A^*$, la loi de $-\lambda_A(Y)$ conditionnellement à $\widehat{\theta_A}(Y) = \theta'_A$ est une loi continue donc on est dans le cas d'égalité du lemme 2. L'inégalité (3.4) est donc une égalité. Elle est conservée par intégration.

□

L'hypothèse de continuité est énoncée sur Θ_A^* et non sur Θ_A car les vecteurs de paramètres de type $\theta'_A = (\beta'_A, 0)$ (avec écart-type nul) conduisent à une loi de λ_A conditionnellement à $\widehat{\theta_A}(Y) = \theta'_A$ non continue. En effet, le conditionnement revient alors à supposer que $\widehat{\sigma_A}(Y) = 0$ donc que Y est une combinaison linéaire des variables de A (il n'est donc plus possible de faire varier localement Y en maintenant le conditionnement). Sous ce conditionnement, il existe une probabilité non nulle que le Lasso ne sélectionne aucune variable en dehors de A, soit $\lambda_A(Y) = 0$.

Par ailleurs, il est possible que l'hypothèse de continuité ne soit pas vérifiée. Par exemple, si il existe une variable $X_{j'}$ qui soit une combinaison linéaire des $X_j, j \in A$, alors sa covariance avec toute combinaison linéaire de Y et des

$X_j, j \in A$ est déterminée de façon unique par $\widehat{\theta_A}$. Or, si aucune autre variable que $X_{j'}$ et celles de A n'apparaît sur le chemin du Lasso avant $X_{j'}$, le λ d'apparition de $X_{j'}$ sur le chemin du Lasso est déterminé par ces covariances. Il existe donc une valeur de λ, fonction déterministe de θ'_A, que λ_A prend avec une probabilité non nulle conditionnellement à $\widehat{\theta_A}(Y) = \theta'_A$ (car la probabilité que j' soit la première variable sélectionnée en dehors de A est non nulle). Dans ce type de cas, la loi conditionnelle de $\lambda_A(Y)$ n'est donc pas continue.

Le lemme 3 permet d'utiliser $p_A(y)$ comme statistique de test. Cette quantité est proche de 0 lorsque $\lambda_A(y)$ est élevé, ce qui est représentatif de $H_1(A)$. C'est pourquoi nous souhaitons rejeter $H_0(A)$ si $p_A(y)$ est suffisamment petit. Le lemme assure que pour tout $\alpha \in [0, 1]$, rejeter $H_0(A)$ si et seulement si $p_A(y) \leq \alpha$ garantit que l'erreur de type 1 est au plus α. La p-value associée à la statistique $p_A(y)$ est $p = P_{H_0(A)}(p_A(Y) \leq p_A(y))$ qui est $p_A(y)$ elle-même sous l'hypothèse de continuité, et inférieure à $p_A(y)$ dans le cas général.

3.4 .Loi conditionnelle du vecteur réponse

L'avantage de choisir $p_A(y)$ comme statistique de test plutôt que $\lambda_A(y)$, et donc (au plus) $p_A(y)$ comme p-value plutôt que $p^0_A(y)$, est qu'il est possible d'estimer $p_A(y)$ par la méthode de Monte-Carlo en suivant les étapes listées en 3.2. En effet, ce qui empêche d'appliquer cette méthode à $p^0_A(y)$ est l'impossibilité de simuler Y selon sa loi sous $H_0(A)$, qui n'est pas connue. Grâce au conditionnement qui intervient dans la définition de $p_A(y)$, on a besoin pour estimer cette quantité de simuler Y selon sa loi sous $H_0(A)$ conditionnellement à $\widehat{\theta_A}(Y)$. Cela est possible grâce à une série de lemmes. Le premier décrit la loi conditionnelle de Y sous l'hypothèse nulle :

Lemme 4. *Sous $H_0(A)$, la loi de Y conditionnellement à $\widehat{\theta_A}(Y) = \widehat{\theta_A}(y)$ est la loi uniforme sur l'ensemble $S_A(y) = \left\{y' \in \mathbb{R}^n | \widehat{\theta_A}(y') = \widehat{\theta_A}(y)\right\}$.*

Démonstration. Soit p_Y la loi de probabilité de Y et soit $y' \in \mathbb{R}^n$. Alors, d'après le modèle 3.1 :

$$\log p_Y(y') = -\frac{1}{2\sigma^2}(y' - X\beta)^T(y' - X\beta) + \text{Constante.}$$

Supposons $H_0(A)$ vraie. Alors Y suit le modèle 3.2 :

$$\begin{aligned}\log p_Y(y') &= -\frac{1}{2\sigma^2}(y' - X_A\beta_A)^T(y' - X_A\beta_A) + \text{Constante} \\ &= -\frac{1}{2\sigma^2}\left({y'}^T y' - 2{\beta_A}^T X_A^T y'\right) + \text{Constante.}\end{aligned}$$

$\widehat{\beta_A}$ est l'estimateur par maximum de vraisemblance, c'est-à-dire, le modèle étant linéaire, par moindres carrés :

$$\widehat{\beta_A}(y') = (X_A^T X_A)^{-1} X_A^T y' \quad \text{donc} \quad X_A^T y' = X_A^T X_A \widehat{\beta_A}(y').$$

Comme il s'agit d'un modèle linéaire, la somme totale des carrés de y' se décompose en somme des carrés expliquée par par X_A plus somme des carrés des résidus :

$$\begin{aligned}\sum_{i=1}^n {y_i'}^2 &= \sum_{i=1}^n \left(\sum_{j\in A} \widehat{\beta_A}(y')_j X_{ij}\right)^2 + \sum_{i=1}^n \left(y_i' - \sum_{j\in A} \widehat{\beta_A}(y')_j X_{ij}\right)^2 \\ {y'}^T y' &= \left(X_A\widehat{\beta_A}(y')\right)^T X_A\widehat{\beta_A}(y') + n\widehat{\sigma^2}(y') \\ \log p_Y(y') &= -\frac{1}{2\sigma^2}\left(\left(X_A\widehat{\beta_A}(y')\right)^T X_A\widehat{\beta_A}(y') + n\widehat{\sigma^2}(y') - 2{\beta_A}^T X_A^T X_A\widehat{\beta_A}(y')\right) \\ &+ \text{Constante.}\end{aligned}$$

$p_Y(y')$ est donc une fonction déterministe de $\widehat{\theta_A}(y') = (\widehat{\beta_A}(y'), \hat{\sigma}(y'))$. Par construction, $\widehat{\theta_A}$ est constant sur $S_A(y)$ donc la densité p_Y est constante sur $S_A(y)$. Donc la loi de Y conditionnellement à $Y \in S_A(y)$ est uniforme. □

Mathématiquement, cet énoncé et la dernière phrase de sa démonstration ne sont pas tout à fait complets car nous ne définissons pas ce qu'est une loi uniforme sur le sous-ensemble de $\mathbb{R}^n$ de mesure nulle qu'est $S_A(y)$. Une loi uniforme est une loi de probabilité dont la densité est constante, or la densité de probabilité d'une variable aléatoire U se définit par rapport une mesure μ qui doit vérifier (théorème de Radon-Nikodym) la condition d'*absolue continuité* : $\forall B,\ \mu(B) = 0 \implies \mathrm{P}(U \in B) = 0$. Implicitement, toutes les densités de probabilité que l'on utilise dans $\mathbb{R}^n$ sont définies par rapport à la mesure de Lebesgue de $\mathbb{R}^n$ (généralisation multidimensionnelle de l'aire, du volume, etc.). Or la loi conditionnelle de Y n'est pas absolument continue par rapport à la mesure de Lebesgue puisque $S_A(Y)$ est de mesure de Lebesgue nulle (à la manière de la surface d'une sphère dans l'espace tridimensionnel, qui est de volume nul) mais par définition $\mathrm{P}(Y \in S_A(y) | \widehat{\theta_A}(y') = \widehat{\theta_A}(y)) = 1$. Rendre le raisonnement entièrement rigoureux demanderait donc de définir une mesure sur $S_A(y)$ qui satisfasse la condition d'absolue continuité (analogue de l'aire d'une surface plongée dans l'espace tridimensionnel), sans doute à l'aide d'outils avancés de géométrie différentielle qui sortent du propos de cette thèse.

En pratique, nous admettons que la restriction d'une loi à densité à un ensemble (même de mesure nulle) sur lequel la densité de probabilité est constante produit une loi uniforme sur cet ensemble, et que le caractère uniforme d'une loi de probabilité sur un sous-ensemble de $\mathbb{R}^n$ est conservé par les transformations les plus simples : translation et homothétie.

La définition de $p_A(y)$ (3) s'appuie sur la loi de Y conditionnellement à $\widehat{\theta_A}(Y) = \widehat{\theta_A}(y)$, dont on sait grâce au lemme 4 qu'elle est la loi uniforme sur $S_A(y)$. Donc dans la définition de $p_A(y)$, Y peut être remplacé par toute variable aléatoire suivant cette loi :

Corollaire 4.1. *Toute variable aléatoire Y^u qui suit une loi uniforme sur $S_A(y)$ satisfait $\mathrm{P}(\lambda_A(Y^u) \geq \lambda_A(y)) = p_A(y)$.*

3.5 .Calibration linéaire du vecteur réponse

Le corollaire 4.1 implique que l'on peut estimer $p_A(y)$ par une méthode de Monte-Carlo à condition de savoir simuler des vecteurs réponses aléatoires de façon uniforme sur $S_A(y)$. Il est donc important de pouvoir « imposer » à un vecteur réponse y' la condition $\widehat{\theta_A}(y') = \widehat{\theta_A}(y)$, ou, plus généralement, la condition $\widehat{\theta_A}(y') = \theta_A^{(2)}$ pour tout vecteur de paramètres cible $\theta_A^{(2)}$. Cela est possible grâce aux *fonctions de calibration* que nous définissons de la façon suivante.

Pour tous vecteurs de paramètres $\theta_A^{(1)} = (\beta_A^{(1)}, \sigma_A^{(1)}) \in \Theta_A^*$ et $\theta_A^{(2)} = (\beta_A^{(2)}, \sigma_A^{(2)}) \in \Theta_A$, soit :

$$\begin{aligned} \mathrm{cal}_{\theta_A^{(1)} \to \theta_A^{(2)}} : \mathbb{R}^n &\to \mathbb{R}^n \\ y' &\to X_A \beta_A^{(2)} + \frac{\sigma_A^{(2)}}{\sigma_A^{(1)}} \left(y' - X_A \beta_A^{(1)} \right). \end{aligned}$$

Les fonctions de calibration envoient une valeur donnée de $\widehat{\theta_A}$ sur une autre valeur donnée, c'est-à-dire :

Lemme 5. *Pour tous $\theta_A^{(1)} \in \Theta_A^*, \theta_A^{(2)} \in \Theta_A$, pour tout $y^{(1)} \in \mathbb{R}^n$ vérifiant $\widehat{\theta_A}(y^{(1)}) = \theta_A^{(1)}$, on a :*

$$\widehat{\theta_A} \left(\mathrm{cal}_{\theta_A^{(1)} \to \theta_A^{(2)}}(y^{(1)}) \right) = \theta_A^{(2)}.$$

Démonstration. Soient $\left(\theta_A^{(1)}, \theta_A^{(2)}, y^{(1)} \right) \in \Theta_A^* \times \Theta_A \times \mathbb{R}^n$ avec $\widehat{\theta_A}(y^{(1)}) = \theta_A^{(1)}$, c'est-à-dire $\widehat{\beta_A}(y^{(1)}) = \beta_A^{(1)}$ et $\widehat{\sigma_A}(y^{(1)}) = \sigma_A^{(1)}$.

Soit $y^{(2)} = \mathrm{cal}_{\theta_A^{(1)} \to \theta_A^{(2)}}(y^{(1)})$. Alors :

$$\begin{aligned}
\widehat{\beta_A}(y^{(2)}) &= (X_A^T X_A)^{-1} X_A^T y^{(2)} \\
&= (X_A^T X_A)^{-1} X_A^T \left(X_A \beta_A^{(2)} + \frac{\sigma_A^{(2)}}{\sigma_A^{(1)}} \left(y^{(1)} - X_A \beta_A^{(1)} \right) \right) \\
&= (X_A^T X_A)^{-1} X_A^T X_A \beta_A^{(2)} \\
&+ \frac{\sigma_A^{(2)}}{\sigma_A^{(1)}} \left((X_A^T X_A)^{-1} X_A^T y^{(1)} - (X_A^T X_A)^{-1} X_A^T X_A \beta_A^{(1)} \right) \\
\widehat{\beta_A}(y^{(2)}) &= \beta_A^{(2)} + \frac{\sigma_A^{(2)}}{\sigma_A^{(1)}} \left(\widehat{\beta_A}(y^{(1)}) - \beta_A^{(1)} \right).
\end{aligned}$$

$\widehat{\beta_A}(y^{(1)}) = \beta_A^{(1)}$ donc $\widehat{\beta_A}(y^{(2)}) = \beta_A^{(2)}$. De plus :

$$\begin{aligned}
\widehat{\sigma_A}^2(y^{(2)}) &= \frac{1}{n} \left|\left| y^{(2)} - X_A \widehat{\beta_A}(y^{(2)}) \right|\right|^2 \\
&= \frac{1}{n} \left|\left| X_A \left(\beta_A^{(2)} - \widehat{\beta_A}(y^{(2)}) \right) + \frac{\sigma_A^{(2)}}{\sigma_A^{(1)}} \left(y^{(1)} - X_A \beta_A^{(1)} \right) \right|\right|^2 \\
&= \left(\frac{\sigma_A^{(2)}}{\sigma_A^{(1)}} \right)^2 \frac{1}{n} \left|\left| y^{(1)} - X_A \beta_A^{(1)} \right|\right|^2 \quad \text{car } \widehat{\beta_A}(y^{(2)}) = \beta_A^{(2)} \\
&= \left(\frac{\sigma_A^{(2)}}{\sigma_A^{(1)}} \right)^2 \widehat{\sigma_A}^2(y^{(1)}) \\
\widehat{\sigma_A}^2(y^{(2)}) &= {\sigma_A^{(2)}}^2 \quad \text{car } \widehat{\sigma_A}(y^{(1)}) = \sigma_A^{(1)}.
\end{aligned}$$

On a bien $\widehat{\theta_A}(y^{(2)}) = \theta_A^{(2)}$. □

On en déduit la possibilité de simuler uniformément sur $S_A(y)$:

Lemme 6. *Pour tout $\theta'_A \in \Theta_A^*$, pour toute variable aléatoire $Y' \in \mathbb{R}^n$ qui suit le modèle 3.2 avec les paramètres θ'_A, la variable aléatoire :*

$$Y^{cal} = \mathrm{cal}_{\widehat{\theta_A}(Y') \to \widehat{\theta_A}(y)}(Y')$$

suit une loi uniforme sur $S_A(y)$.

Démonstration. En remplaçant θ_A et Y dans le lemme 4 par θ'_A et Y', on obtient que la loi de Y' conditionnellement à $\widehat{\theta_A}(Y')$ est uniforme sur $S_A(Y')$. De plus, conditionnellement à $\widehat{\theta_A}(Y')$, $X_A\widehat{\beta_A}(Y')$ et $\hat{\sigma}(y)/\hat{\sigma}(Y')$ sont des constantes donc :

$$\mathrm{cal}_{\widehat{\theta_A}(Y')\to\widehat{\theta_A}(y)} : y' \to X_A\widehat{\beta_A}(y) + \frac{\hat{\sigma}(y)}{\hat{\sigma}(Y')}\left(y' - X_A\widehat{\beta_A}(Y')\right)$$

est une composée d'homothétie et de translation. Nous avons admis que ce type de transformation préserve le caractère uniforme d'une loi de probabilité. De plus d'après le lemme 5, elle envoie $S_A(Y')$ sur $S_A(y)$. Donc conditionnellement à $\widehat{\theta_A}(Y')$, Y^{cal} suit une loi uniforme sur $S_A(y)$. Donc par intégration, Y^{cal} suit une loi uniforme sur $S_A(y)$. □

Informellement, la fonction de calibration, en tant que simple composée d'homothétie et de translation, peut être vue comme la modification la moins importante que l'on peut faire subir à un vecteur y' pour lui imposer la condition $\widehat{\theta_A}(y') = \widehat{\theta_A}(y)$. En particulier, la structure de corrélation empirique entre y' et les variables n'appartenant pas à A change peu entre y' et sa version calibrée : elle n'est modifiée que dans la mesure où elle est portée par les variables de A.

3.6 .Algorithme d'estimation de la p-value conditionnelle

Nous possédons maintenant tous les ingrédients pour estimer $p_A(y)$ par la méthode de Monte-Carlo suivante, qui adapte et détaille la méthode proposée en 3.2. Elle est paramétrée par un entier N, le nombre de simulations, qui contrôle la précision de l'estimation.

Algorithme 1 (Estimation de $p_A(Y)$ par simulation-calibration)**.** Les quatre étapes sont les suivantes :

1 Calculer $\widehat{\theta_A}(y)$, le vecteur de paramètres du modèle 3.2 estimé par maximum de vraisemblance.

2 Simuler N vecteurs réponse $y^{(1)}, .., y^{(N)}$ indépendants les uns des autres suivant la loi uniforme sur $S_A(y)$. Cela est réalisé de la façon suivante, pour tout $l = 1, .., N$:

2.1 Simuler y^{sim} suivant le modèle 3.2 avec un vecteur de paramètres $\theta_A^{sim} \in \Theta_A^*$ quelconque :

$$y^{sim} = X_A \beta_A^{sim} + \epsilon \text{ avec } \epsilon \sim \mathcal{N}(0, {\sigma^{sim}}^2 I_n)$$

2.2 Calculer l'estimation du maximum de vraisemblance $\widehat{\theta_A}(y^{sim})$.

2.3 Calculer la version calibrée de y^{sim} :

$$y^{(l)} = \mathrm{cal}_{\widehat{\theta_A}(y^{sim}) \to \widehat{\theta_A}(y)}(y^{sim})$$

3 Pour tout $l = 1, .., N$, réaliser la régression Lasso de $y^{(l)}$ sur X, et calculer $\lambda_A(y^{(l)})$.

4 Calculer la p-value conditionnelle empirique :

$$\widehat{p_A}(y) = \frac{1}{N} \sum_{l=1}^{N} \mathbf{1} \left\{ \lambda_A(y^{(l)}) \geq \lambda_A(y) \right\}.$$

En pratique, à l'étape 2.1. nous utilisons $\theta_A^{sim} = \widehat{\theta_A}(y)$, mais ce choix n'a pas d'importance car toutes les valeurs de θ_A^{sim} produisent des réponses calibrées $y^{(l)}$ qui suivent la même loi.

3.7 .Propriétés théoriques de l'algorithme

3.7.1 .Propriétés pour un vecteur réponse fixé

Remarquons d'abord que les trois sous-étapes de l'étape 2. de l'algorithme produisent bien le résultat annoncé au début de l'étape 2., c'est-à-dire simuler des $y^{(l)}$ suivant la loi uniforme sur $S_A(y)$. Cela est garanti par le lemme 6, puisque y^sim suit le modèle 3.2 et qu'on lui applique la bonne fonction de calibration.

Cela permet de connaître la loi du $\widehat{p_A}(y)$ produit par l'algorithme. Elle est donnée par la loi binomiale :

Lemme 7.

$$N\widehat{p_A}(y) \sim \mathrm{Bin}(N, p_A(y)).$$

Démonstration. Pour tout l, $y^{(l)}$ suit la loi uniforme sur $S_A(y)$. Donc, d'après le corollaire 4.1, à chaque simulation :

$$P\left(\lambda_A(y^{(l)}) \geq \lambda_A(y)\right) = p_A(y).$$

$N\widehat{p_A}(y)$ est donc la somme de N variables aléatoires binaires indépendantes de moyenne $p_A(y)$. Cette quantité suit donc une loi $\mathrm{Bin}(N, p_A(y))$. □

En tant qu'estimateur de $p_A(y)$, $\widehat{p_A}(y)$ a donc les propriétés suivantes :

- Il est non biaisé.
- Sa variance vaut $\mathrm{Var}\left(\widehat{p_A}(y)\right) = p_A(y)(1 - p_A(y))/N$, qui tend vers 0 lorsque N tend vers l'infini.
- D'après la loi des grands nombres, $\widehat{p_A}(y)$ converge presque sûrement vers $p_A(y)$ lorsque N tend vers l'infini.

3.7.2 . Propriétés pour un vecteur réponse aléatoire

Ce qui précède s'applique à un vecteur réponse y donné. Dans le cas d'une vecteur réponse aléatoire Y qui suit le modèle 3.1 (ou modèle 3.2 sous l'hypothèse nulle), il est intéressant de connaître la loi de $\widehat{p_A}(Y)$. En effet la statistique de test théorique $p_A(Y)$, dont la loi uniforme (ou dominant celle-ci) est est garantie par le lemme 3 et le corollaire 3.1, n'est pas connue, elle est seulement estimée par $\widehat{p_A}(Y)$. La décision de sélection ou non d'une variable est donc fondée sur $\widehat{p_A}(Y)$, et tout résulat de contrôle de l'erreur de première espèce doit se déduire de la loi que suit $\widehat{p_A}(Y)$ sous l'hypothèse nulle. Nous déterminons de façon exacte cette loi lorsque l'hypothèse de continuité définie en 3.3 est vérifiée, puis nous en déduisons une domination stochastique de cette loi dans le cas général. Cela garantit le contrôle de l'erreur de première espèce (à un terme correctif de l'ordre de $1/N$ près) lorsque $\widehat{p_A}(Y)$ est utilisée comme une p-value.

La relation entre $p_A(y)$ et $\widehat{p_A}(y)$ (lemme 7) va permettre de transposer les résultats portant sur la loi de $p_A(Y)$ à celle de $\widehat{p_A}(Y)$, en passant d'une loi continue à une loi discrète :

Lemme 8. *Sous $H_0(A)$ et sous l'hypothèse de continuité, la loi de $\widehat{p_A}(Y)$ est la loi uniforme discrète sur $\{0, 1/N, .., 1 - 1/N, 1\}$.*

Démonstration. Soit $k \in 0, .., N$. La loi de $\widehat{p_A}(Y)$ conditionnellement à $p_A(Y)$ est donnée par $N\widehat{p_A}(Y) \mid p_A(Y) \sim \mathrm{Bin}(N, p_A(Y))$, c'est-à-dire :

$$\mathrm{P}\left(\widehat{p_A}(Y) = \frac{k}{N} \mid p_A(Y)\right) = \binom{N}{k} p_A(Y)^k \left(1 - p_A(Y)\right)^{N-k}$$

Le corollaire 3.1 garantit que sous $H_0(A)$ et sous l'hypothèse de continuité,

la loi de $p_A(Y)$ est la loi uniforme sur $[0,1]$. On a donc :

$$\begin{aligned}
\mathrm{P}_{H_0(A)}\left(\widehat{p_A}(y)=\frac{k}{N}\right) &= \int_0^1 \mathrm{P}\left(\widehat{p_A}(y)=\frac{k}{N} \mid p_A(Y)=t\right)dt \\
&= \binom{N}{k}\int_0^1 t^k(1-t)^{N-k}dt \\
&= \binom{N}{k}\frac{\Gamma(k+1)\Gamma(N-k+1)}{\Gamma(N+2)} \\
&= \frac{1}{N+1}.
\end{aligned}$$

Cette probabilité ne dépend pas de k, donc la loi de $\widehat{p_A}(y)$ est uniforme sous $H_0(A)$ et l'hypothèse de continuité. □

Ce résultat, analogue au corollaire 3.1, possède un équivalent qui n'utilise pas l'hypothèse de continuité, analogue au lemme 3. Pour le prouver, nous utilisons le résultat suivant portant sur la dominance stochastique (voir les notes de cours d'Olivier Bos, *Dominance stochastique : théorie et applications*) :

Lemme 9. *Une loi P_1 domine stochastiquement une loi P_2 si et seulement si il existe deux variables aléatoires U_1 et U_2 telles que $U_1 \sim P_1$, $U_2 \sim P_2$ et $U_1 \geq U_2$ presque sûrement.*

Sans détailler la démonstration, remarquons que le sens existence de telles variables vers dominance stochastique se vérifie facilement, et que dans l'autre sens, si P_2 est la loi uniforme sur $[0,1]$ (cas qui nous intéresse ici), U_2 se construit à partir de U_1 par la loi conditionnelle $U_2 \mid (U_1 = u) \sim$ Unif $([\mathrm{P}(U_1 < u), \mathrm{P}(U_1 \leq u)])$.

Les lemmes 8 et 9 permettent de prouver le résultat suivant sur la loi de $\widehat{p_A}(Y)$, sans l'hypothèse de continuité :

Lemme 10. *Sous $H_0(A)$, la loi de $\widehat{p_A}(y)$ domine stochastiquement la loi uniforme discrète sur $\{0, 1/N, .., 1-1/N, 1\}$.*

Démonstration. D'après le lemme 9 et la dominance stochastique de la loi de $p_A(Y)$ sur la loi uniforme sur $[0,1]$ (lemme 3), il existe une variable aléatoire p_A^u de loi uniforme sur $[0,1]$ telle que $p_A(Y) \geq p_A^u$ presque sûrement (p. s.).

Soient $U_1, .., U_N$ des variables aléatoires indépendantes toutes de loi uniforme sur $[0,1]$. On définit les variables aléatoires :

$$\begin{aligned} \hat{p}_1 &= \frac{1}{N}\sum_{t=1}^{N} \mathbf{1}\left\{U_t \leq p_A(Y)\right\} \\ \hat{p}_2 &= \frac{1}{N}\sum_{t=1}^{N} \mathbf{1}\left\{U_t \leq p_A^u\right\} \end{aligned}$$

$p_A(Y) \geq p_A^u$ p. s. donc pour tout t, $\{U_t \leq p_A(Y)\} \geq \{U_t \leq p_A^u\}$ p. s. donc $\hat{p}_1 \geq \hat{p}_2$ p. s. Don la loi de $\hat{p}_1$ domine stochastiquement celle de $\hat{p}_2$.

De plus, de même que $\widehat{p_A}(y)$, conditionnellement à $p_A(Y)$ et $p_A^u(Y)$, $\hat{p}_1$ et $\hat{p}_2$ sont des moyennes empiriques de variables aléatoires indépendantes identiquement distribuées selon une loi de Bernoulli. On a donc les lois conditionnelles :

$$\begin{aligned} N\hat{p}_1 \mid p_A(Y) &\sim \mathrm{Bin}(N, p_A(Y)) \\ N\hat{p}_2 \mid p_A^u &\sim \mathrm{Bin}(N, p_A^u) \end{aligned}$$

$\hat{p}_1$ et $\widehat{p_A}(y)$ suivent la même loi conditionnellement à $p_A(Y)$ donc elles suivent la même loi globalement. $\hat{p}_2$ suit la même loi que $p_A(Y)$ dans le lemme 8 : une loi binomiale à paramètre aléatoire uniforme sur $[0,1]$ (divisée par N), or ce lemme indique que c'est la loi uniforme discrète sur $\{0, 1/N, .., 1-1/N, 1\}$.

Comme la loi de $\hat{p}_1$ domine stochastiquement celle de $\hat{p}_2$, la loi de $\widehat{p_A}(y)$ domine stochastiquement loi uniforme discrète sur $\{0, 1/N, .., 1-1/N, 1\}$. □

Ce résultat nous permet de contrôler l'erreur de type 1 : pour tout $\alpha \in [0,1]$, produire un $\widehat{p_A}(y)$ puis rejeter $H_0(A)$ si et seulement si $\widehat{p_A}(y) \leq \alpha$ ga-

rantit une erreur de type 1 inférieure où égale à $\frac{\lfloor \alpha N \rfloor + 1}{N+1} \leq \alpha + \frac{1-\alpha}{N+1}$. Réciproquement, rejeter $H_0(A)$ si et seulement si $\widehat{p_A}(y) \leq \alpha - \frac{1-\alpha}{N}$ garantit une erreur de type 1 inférieure où égale à α.

Les termes résiduels $\frac{1-\alpha}{N+1}$ et $-\frac{1-\alpha}{N}$ sont très petits (plus petits que la granularité d'estimation de $\widehat{p_A}(y)$, qui est $1/N$) et contrôlables par l'utilisateur puisque N peut être choisi ausssi grand que l'on veut (au prix du temps de calcul). Remarquons par ailleurs qu'il est possible d'adapter la définition de $\widehat{p_A}$ pour faire formellement disparaître les termes résiduels. Avec :

$$\widehat{p_A}^+(y) = \frac{1}{N+1}\left(1 + \sum_{l=1}^{N} \mathbf{1}\left\{\lambda_A(y^{(l)}) \geq \lambda_A(y)\right\}\right),$$

le critère de rejet $\widehat{p_A}^+(y) \leq \alpha$ entraîne le contrôle d'erreur au niveau α. $\widehat{p_A}^+(y)$ est cependant un estimateur biaisé de $p_A(y)$.

3.8 .Cas des modèles linéaires généralisés

3.8.1 .Position du problème

L'algorithme proposé est valide dans le cas du modèle linéaire. Nous proposons une adaptation de celui-ci à certains modèles linéaires généralisés. Nous considérons les deux modèles suivants, qui sont les modèles linéaires généralisés les plus classiques :

- modèle binaire (à fonction de lien logistique) :

$$y \in \{0,1\}^n, \;\; y|X \sim \text{Bernoulli}\left(f(X\beta)\right) \;\; \text{avec} \;\; f(t) = \frac{1}{1+e^{-t}}$$

- modèle de Poisson (à fonction de lien exponentielle) :

$$y \in \mathbb{N}^n, \;\; y|X \sim \text{Poisson}\left(f(X\beta)\right) \;\; \text{avec} \;\; f(t) = e^t.$$

Contrairement au modèle linéaire, ces deux modèles n'ont pas de paramètre d'écart-type σ, le vecteur des paramètres θ se réduit donc à β (y compris dans le modèle restreint à A). On écrit donc β_A ou $\widehat{\beta_A}$ à la place de θ_A ou $\widehat{\theta_A}$. Ces deux modèles ont aussi en commun le fait que y suit une loi discrète et non plus continue.

Comme dans le cas linéaire, nous supposons que les variables $X_j, j \in A$ sont actives et nous cherchons à déterminer si d'autres variables sont actives, n'appartenant pas à A. Les hypothèses $H_0(A), H_1(A)$, ainsi que la statistique $\lambda_A(y)$, sont définies de la même façon. Nous voudrions en principe estimer $p_A^0(y) = \mathrm{P}_{H_0(A)}\left(\lambda_A(Y) \geq \lambda_A(y)\right)$, ce qui n'est pas possible car la loi de Y sous $H_0(A)$ n'est pas connue. Comme dans le cas linéaire, nous proposons un algorithme qui calcule $\widehat{p_A}(y)$, une estimation d'une statistique qui approxime $p_A^0(y)$. Cet algorithme est proche de l'algorithme 1, n'en différant que par sa méthode de calibration. Cependant, la statistique estimée par $\widehat{p_A}(y)$ dans les modèles linéaires généralisés n'est pas la même que dans le cas linéaire et elle ne possède pas ses propriétés théoriques.

La solution adoptée dans le cas linéaire était d'estimer, à la place de $p_A^0(y)$, $p_A(y) = \mathrm{P}_{H_0(A)}\left(\lambda_A(Y) \geq \lambda_A(y) | \widehat{\beta_A}(Y) = \widehat{\beta_A}(y)\right)$. Malheureusement, dans les modèles discrets, cette quantité perd son intérêt car l'ensemble E des valeurs que peut prendre y est beaucoup plus restreint : il est fini dans le modèle binaire ($E = \{0,1\}^n$), et dénombrable dans le modèle de Poisson ($E = \mathbb{N}^n$). L'ensemble $\Theta_A = \mathbb{R}^{|A|}$ des paramètres possibles est au contraire non dénombrable. On peut donc s'attendre, en général, à ce que la fonction estimateur $\widehat{\beta_A}$, qui envoie un ensemble fini ou dénombrable vers un ensemble non dénombrable, soit injective : un seul vecteur, y, produit exactement l'estimé $\widehat{\beta_A}(y)$. Donc le conditionnement par $\widehat{\beta_A}(Y) = \widehat{\beta_A}(y)$ implique $Y = y$ d'où $\lambda_A(Y) = \lambda_A(y)$ donc on a nécessairement $p_A(y) = 1$.

Néanmoins, si n est assez grand, le nombre de valeurs que peut prendre y (et donc $\widehat{\beta_A}(y)$) est très élevé (2^n dans le modèle binaire). Donc même si seul $y' = y$ satisfait $\widehat{\beta_A}(y') = \widehat{\beta_A}(y)$, un grand nombre de vecteurs y' distincts peuvent satisfaire $\widehat{\beta_A}(y') \approx \widehat{\beta_A}(y)$. Par conséquent, le conditionnement par $\widehat{\beta_A}(Y) = \widehat{\beta_A}(y)$, qui ne permet pas de définir une grandeur intéressante dans les modèles discrets, peut — informellement — être remplacé par un « conditionnement par $\widehat{\beta_A}(Y) \approx \widehat{\beta_A}(y)$ ». Plus formellement, cela revient à utiliser une loi de probabilité $\mathrm{P}_{\widehat{\beta_A}(y)}$ sur l'ensemble des valeurs que peut prendre y, produisant des vecteurs aléatoires Y compatibles avec $H_0(A)$ et qui satisfont $\widehat{\beta_A}(Y) \approx \widehat{\beta_A}(y)$. Cette dernière condition peut se quantifier en terme d'erreur quadratique moyenne :

$$\mathrm{EQM}\left(\mathrm{P}_{\widehat{\beta_A}(y)}\right) = \mathrm{E}_{Y \sim \mathrm{P}_{\widehat{\beta_A}(y)}}\left[\left(X_A\widehat{\beta_A}(Y) - X_A\widehat{\beta_A}(y)\right)^2\right]$$

qui doit être très proche de 0. On estimera ensuite :

$$\widetilde{p_A}(y) = \mathrm{P}_{Y \sim \mathrm{P}_{\widehat{\beta_A}(y)}}\left(\lambda_A(Y) \geq \lambda_A(y)\right)$$

qui est l'analogue approximatif pour les modèles discrets de la statistique de test $p_A(y)$ du modèle linéaire.

Dans le modèle linéaire, nous simulions des vecteurs $y^{(l)}$ suivant la loi produite par le conditionnement par $\lambda_A(Y) = \lambda_A(y)$ — loi que nous avons déterminé explicitement : il s'agit de la loi uniforme sur l'ensemble $S_A(y)$ — puis nous utilisions la population des $y^{(l)}$ pour estimer $p_A(y)$ par Monte-Carlo. La simulation de vecteurs suivant cette la loi était réalisée en appliquant une fonction de calibration à des vecteurs simulés selon le modèle linéaire sous l'hypothèse nulle. De façon similaire, dans les modèles discrets, nous proposons de simuler des $y^{(l)}$ suivant une loi $\mathrm{P}_{\widehat{\beta_A}(y)}$ qui satisfait aux conditions ci-

dessus mais qui n'est pas déterminée explicitement, et d'utiliser ces $y^{(l)}$ pour estimer $\widetilde{p_A}(y)$ par Monte-Carlo. La simulation de vecteurs suivant $\mathrm{P}_{\widehat{\beta_A}(y)}$ est également réalisée en appliquant à des vecteurs simulés selon le modèle linéaire généralisé sous l'hypothèse nulle une procédure de calibration, qui est cependant plus complexe que celle du modèle linéaire.

3.8.2 . Calibration dans les modèles non linéaires

Étant donnés deux vecteurs de paramètres $\beta_A^{(1)}, \beta_A^{(2)} \in \mathbb{R}^n$, nous cherchons une procédure de calibration de $\beta_A^{(1)}$ vers $\beta_A^{(2)}$ telle que :

Conditions 1 (Propriétés désirables de la calibration)**.** Pour tout $y^{(1)} \in E$, en notant $y^{(2)}$ sa version calibrée :

1. $y^{(2)}$ soit « proche » de $y^{(1)}$ pour conserver autant que possible sa structure de corrélation avec les variables en-dehors de A.
2. Si $\widehat{\beta_A}(y^{(1)}) = \beta_A^{(1)}$ alors $\widehat{\beta_A}(y^{(2)}) \approx \beta_A^{(2)}$.

La condition 1.2. est l'adaptation informelle de la propriété formelle satisfaite par la calibration dans le cas linéaire d'après le lemme 5, l'égalité des paramètres, qui est une propriété trop forte dans les modèles discrets, étant remplacée par une approximation.

Dans les modèles linéaires, la calibration est une fonction $\mathrm{cal}_{\beta_A^{(1)} \to \beta_A^{(2)}} : E \to E$ composée de multiplication par et d'ajout de constantes non entières. Dans les modèles discrets, cela n'est pas possible car la calibration doit produire un vecteur composé de nombre entiers. La solution à cela est de remplacer des valeurs non entières déterministes par des entiers aléatoires de même espérance. C'est pourquoi la calibration dans les modèles non linéaires est une procédure aléatoire qui se décrit par la loi du vecteur calibré conditionnellement au vecteur de départ : $Y^{(2)}|Y^{(1)} \sim \mathrm{P}_{\beta_A^{(1)} \to \beta_A^{(2)}}$.

Algorithme de calibration simple dans les modèles non linéaires

Nous présentons ci-dessous une première procédure simple de calibration, selon une loi conditionnelle notée $\mathrm{P}^{(1)}_{\beta_A^{(1)} \to \beta_A^{(2)}}$. Elle constitue une étape de la procédure complète qui permet de simuler selon $\mathrm{P}_{\beta_A^{(1)} \to \beta_A^{(2)}}$. Dans les deux modèles linéaires généralisés, elle est construite à partir des vecteurs de prédictions produits par les paramètres de départ et cible : $e^{(1)} = f\left(X\beta_A^{(1)}\right)$ et $e^{(2)} = f\left(X\beta_A^{(2)}\right)$.

Algorithme 2 (Calibration non-linéaire : algorithme simple)**.** Simuler $Y^{(2)}|Y^{(1)} \sim \mathrm{P}^{(1)}_{\beta_A^{(1)} \to \beta_A^{(2)}}$ signifie, connaissant le vecteur $Y^{(1)}$, simuler le vecteur $Y^{(2)}$ de la façon suivante, indépendamment pour tout individu i :

- Dans le modèle binaire :

$$\begin{aligned} Y_i^{(2)}|Y_i^{(1)} &\sim \mathrm{Bernoulli}(Z_i) \text{ où :} \\ \text{si } e_i^{(2)} \leq e_i^{(1)},\ Z_i &= \frac{e_i^{(2)}}{e_i^{(1)}} Y_i^{(1)} \\ \text{si } e_i^{(2)} \geq e_i^{(1)},\ Z_i &= 1 - \frac{1 - e_i^{(2)}}{1 - e_i^{(1)}}(1 - Y_i^{(1)}) \end{aligned}$$

- Dans le modèle de Poisson :

$$\begin{aligned} Y_i^{(2)} &= \lfloor Z_i \rfloor + R_i \text{ où :} \\ Z_i &= \frac{e_i^{(2)}}{e_i^{(1)}} Y_i^{(1)} \\ R_i|Y_i^{(1)} &\sim \mathrm{Bernoulli}\left(Z_i - \lfloor Z_i \rfloor\right). \end{aligned}$$

Propriétés de l'algorithme de calibration simple

Par construction, $\mathrm{E}\left[Y^{(2)}|Y^{(1)}\right] = Z$ dans les deux modèles. Les deux éventualités non-exclusives du cas binaire sont cohérentes car si $e_i^{(1)} = e_i^{(2)}$ alors $Y_i^{(2)} = Y_i^{(1)}$ quelle que soit la formule suivie. Dans les deux modèles,

si $\beta_A^{(1)} = \beta_A^{(2)}$ (cas où aucune calibration n'est nécessaire) alors $e^{(1)} = e^{(2)}$ et $Y^{(1)} = Z = Y^{(2)}$.

Si un individu i vérifie $e_i^{(1)} \approx e_i^{(2)}$, alors la probabilité que $Y_i^{(2)} = Y_i^{(1)}$ est élevée. Si $\beta_A^{(1)} \approx \beta_A^{(2)}$, cela est globalement vérifié par les différents i donc, comme désiré (condition 1.1.), la calibration modifie peu le vecteur $Y^{(1)}$.

Par ailleurs, la calibration « transforme un vecteur qui suit les paramètres $\beta_A^{(1)}$ en un vecteur qui suit les paramètres $\beta_A^{(2)}$ » dans le sens où, $Y^{(1)}$ étant aléatoire :

$$\mathrm{E}\left[Y^{(1)}\right] = e^{(1)} \implies \mathrm{E}\left[Y^{(2)}\right] = \mathrm{E}\left[Z\right] = e^{(2)}. \tag{3.5}$$

Cette propriété n'implique cependant pas la condition 1.2., qui porte sur les estimés $\widehat{\beta_A}(Y^{(1)})$ et $\widehat{\beta_A}(Y^{(2)})$, ou, de façon équivalente, sur les vecteurs de prédictions qu'ils produisent : $\hat{e}^{(m)} = f\left(\widehat{\beta_A}(Y^{(m)})\right)$ pour $m = 1$ ou 2. La condition peut être vérifiée mathématiquement seulement dans le cas simple où le modèle est réduit à un intercept, c'est-à-dire $A = \emptyset$. Les vecteurs de prédiction $e^{(1)}$ et $e^{(2)}$ sont alors des constantes.

Lemme 11. *Dans les modèles à un paramètre, si $\widehat{\beta_A}(Y^{(1)}) = \beta_A^{(1)}$ c'est-à-dire $\hat{e}^{(1)} = e^{(1)}$, alors :*

- $\mathrm{Var}\left[\hat{e}^{(2)}|Y^{(1)}\right] = e^{(2)}$
- $\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) \leq \frac{1}{n}\left|e^{(2)} - e^{(1)}\right|$ *avec égalité dans le modèle binaire.*

Démonstration. Dans les modèles à un paramètre, l'estimation de ce paramètre par maximum de vraisemblance se ramène au calcul de la moyenne empirique : $\hat{e}^{(m)} = \overline{Y^{(m)}}$.

En supposant que $\widehat{\beta_A}(Y^{(1)}) = \beta_A^{(1)}$, on a donc $\overline{Y^{(1)}} = e^{(1)}$. Dans le modèle binaire comme dans le modèle de Poisson,

$$\mathrm{E}\left[\hat{e}^{(2)}|Y^{(1)}\right] = \mathrm{E}\left[\overline{Y^{(2)}}|Y^{(1)}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[Y_i^{(2)}|Y^{(1)}\right] = \overline{Z}.$$

Z est défini de façon linéaire donc soit $\overline{Z} = \frac{e^{(2)}}{e^{(1)}}\overline{Y^{(1)}}$ (modèle binaire avec $e^{(2)} \leq e^{(1)}$, et modèle de Poisson), soit $\overline{Z} = 1 - \frac{1-e^{(2)}}{1-e^{(1)}}(1-\overline{Y^{(1)}})$ (modèle binaire avec $e^{(2)} \geq e^{(1)}$). Dans les deux cas, puisque $\overline{Y^{(1)}} = e^{(1)}$, on a bien $\overline{Z} = e^{(2)}$.

Dans le modèle binaire comme dans le modèle de Poisson,

$$\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) = \mathrm{Var}\left(\overline{Y^{(2)}}|Y^{(1)}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(Y_i^{(2)}|Y^{(1)}\right).$$

Pour tout $p \in [0,1]$ la variance d'une $\mathrm{Bernouilli}(p)$ est $p(1-p)$. Donc dans le modèle binaire,

$$\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Z_i(1-Z_i).$$

Si $e^{(2)} \leq e^{(1)}$,

$$\begin{aligned}
\forall i, Z_i(1-Z_i) &= \frac{e^{(2)}}{e^{(1)}}Y_i\left(1 - \frac{e^{(2)}}{e^{(1)}}Y_i\right) \\
&= 0 \text{ si } Y_i = 0,\ 1 - \frac{e^{(2)}}{e^{(1)}} \text{ si } Y_i = 1 \\
\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) &= \frac{1}{n^2}\sum_{Y_i=1}\left(1 - \frac{e^{(2)}}{e^{(1)}}\right) \\
&= \frac{1}{n^2}(ne^{(1)})\left(1 - \frac{e^{(2)}}{e^{(1)}}\right) \text{ car } \overline{Y^{(1)}} = e^{(1)} \\
\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) &= \frac{1}{n}\left(e^{(1)} - e^{(2)}\right).
\end{aligned}$$

Et si $e^{(2)} \geq e^{(1)}$,

$$\begin{aligned}
\forall i, Z_i(1-Z_i) &= \frac{1-e^{(2)}}{1-e^{(1)}}(1-Y_i)\left(1-\frac{1-e^{(2)}}{1-e^{(1)}}(1-Y_i)\right) \\
&= 0 \text{ si } Y_i = 1,\ 1-\frac{1-e^{(2)}}{1-e^{(1)}} \text{ si } Y_i = 0 \\
\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) &= \frac{1}{n^2}\sum_{Y_i=0}\left(1-\frac{1-e^{(2)}}{1-e^{(1)}}\right) \\
&= \frac{1}{n^2}n(1-e^{(1)})\left(1-\frac{1-e^{(2)}}{1-e^{(1)}}\right) \text{ car } \overline{1-Y^{(1)}} = 1-e^{(1)} \\
\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) &= \frac{1}{n}\left(e^{(2)}-e^{(1)}\right).
\end{aligned}$$

Dans les deux cas $\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) = \frac{1}{n}\left|e^{(2)}-e^{(1)}\right|$.

Dans le modèle de Poisson, notons $\forall i,\ T_i = Z_i - \lfloor Z_i \rfloor$. Alors :

$$T_i(1-T_i) \leq \min(T_i, 1-T_i) = \min\left(|Z_i - \lfloor Z_i \rfloor|, |Z_i - \lceil Z_i \rceil|\right).$$

L'entier k qui minimise $|Z_i - k|$ est soit $\lfloor Z_i \rfloor$, soit $\lceil Z_i \rceil$. Donc comme $Y_i^{(1)}$ est un nombre entier,

$$T_i(1-T_i) \leq |Z_i - Y_i^{(1)}| = \left|\frac{e^{(2)}}{e^{(1)}} - 1\right| Y_i^{(1)}.$$

$$\begin{aligned}
\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) &= \frac{1}{n^2}\sum_{i=1}^{n} T_i(1-T_i) \leq \frac{1}{n^2}\sum_{i=1}^{n}\left|\frac{e^{(2)}}{e^{(1)}} - 1\right| Y_i^{(1)} \\
\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) &\leq \frac{1}{n^2}\left|\frac{e^{(2)}}{e^{(1)}} - 1\right| (ne^{(1)}) = \frac{1}{n}\left|e^{(2)} - e^{(1)}\right|.
\end{aligned}$$

□

Dans les modèles à un paramètre, $\mathrm{P}^{(1)}_{\widehat{\beta_A}(Y^{(1)}) \to \beta_A^{(2)}}$ produit donc une loi de $Y^{(2)}$ où $\hat{e}^{(2)}$ est distribué autour de $e^{(2)}$ avec une faible variance donc, par

application de f^{-1}, $\widehat{\beta_A}(Y^{(2)})$ est distribué autour de la valeur cible $\beta^{(2)}$ avec une faible variance. Ce résultat, combiné à la propriété de transfert de $\beta^{(1)}$ vers $\beta^{(2)}$ (3.5) qui est valide dans les modèles à plusieurs paramètres, suggère que la loi de $\widehat{\beta_A}(Y^{(2)})$ sous $\mathrm{P}^{(1)}_{\widehat{\beta_A}(Y^{(1)})\to\beta_A^{(2)}}$ a une propriété similaire dans les modèles à plusieurs paramètres.

Algorithme itéré de calibration

L'inégalité $\mathrm{Var}\left(\hat{e}^{(2)}|Y^{(1)}\right) \leq \frac{1}{n}\left|e^{(2)} - e^{(1)}\right|$ donnée par le lemme 11 indique que, au moins dans les modèles à un paramètre, $\hat{e}^{(2)}$ (ou $\widehat{\beta_A}(Y^{(2)})$) s'approche de sa cible avec d'autant plus de précision que le vecteur de départ ($\hat{e}^{(1)}$ ou $\widehat{\beta_A}(Y^{(1)})$) en est déjà proche. Il est donc utile de réappliquer la procédure de calibration suivant une loi $\mathrm{P}^{(1)}_{\cdot\to\beta_A^{(2)}}$ à un vecteur déjà calibré. Nous proposons donc une définition de $\mathrm{P}_{\beta_A^{(1)}\to\beta_A^{(cal)}}$, la loi conditionnelle décrivant la procédure complète de calibration recherchée en 3.8.2, via l'algorithme de simulation suivant.

Algorithme 3 (Calibration non-linéaire : algorithme itératif)**.** Soient $\beta_A^{(1)}, \beta_A^{(cal)} \in \Theta$ des vecteurs de paramètres et $y^{(1)}$ un vecteur réponse. Simuler $Y^{(cal)}|y^{(1)} \sim \mathrm{P}_{\beta_A^{(1)}\to\beta_A^{(cal)}}$ signifie, pour un certain nombre d'itérations $k = 1, 2, ..$:

1. Simuler un vecteur $y^{(k+1)}$ à partir de $y^{(k)}$ suivant la loi $Y^{(k+1)}|Y^{(k)} \sim \mathrm{P}^{(1)}_{\beta_A^{(k)}\to\beta_A^{(cal)}}$, selon l'algorithme 2.
2. Calculer l'estimateur du maximum de vraisemblance, et définir $\beta_A^{(k+1)} = \widehat{\beta_A}(y^{(k+1)})$.
3. Calculer l'erreur quadratique moyenne empirique $\mathrm{EQM}(k+1) = ||X_A\beta_A^{(k+1)} - X_A\beta_A^{(cal)}||_2^2$. Si $\mathrm{EQM}(k+1) > \mathrm{EQM}(k)$, rejeter le $y^{(k+1)}$ qui vient d'être simulé et définir à la place $y^{(k+1)} = y^{(k)}$, $\beta^{(k+1)} = \beta^{(k)}$. Si cela se produit 3 fois consécutivement, mettre fin aux itérations.

On prend alors pour $Y^{(cal)}$ la valeur de $y^{(k)}$ au moment de l'arrêt des itérations.

3.8.3 .Algorithme de test par simulation-calibration dans les modèles linéaires généralisés

Grâce aux algorithmes de calibration 2 et 3, nous pouvons finalement proposer l'algorithme de génération, dans les modèles linéaires généralisés binaire et de Poisson, de la p-value conditionnelle empirique $\widehat{p_A}(y)$ qui mesure la significativité de la première variable sélectionnée par le Lasso en-dehors de l'ensemble A. Il est volontairement très proche sur la forme de l'algorithme 1 des modèles linéaires, les deux principales différences étant :

- au plan théorique, que la quantité estimée n'est pas exactement la même p-value conditionnelle, $\widetilde{p_A}(Y)$ au lieu de $p_A(Y)$ (voir 3.8.1) ;
- au plan algorithmique, que l'étape de calibration est plus complexe, d'où l'appel aux sous-algorithmes 2 et 3.

les adaptations algorithmiques nécessaires ayant étés renvoyés aux algorithmes de calibration auquel il fait appel.

Algorithme 4 (Estimation de $\widetilde{p_A}(Y)$ par simulation-calibration)**.** Les quatre étapes sont les suivantes :

1 Calculer $\widehat{\beta_A}(y)$, le vecteur de paramètres du modèle restreint à A estimé par maximum de vraisemblance.

2 Simuler N vecteurs réponse $y^{(1)}, .., y^{(N)}$ indépendants les uns des autres suivant une loi dite $P_{\widehat{\beta_A}(y)}$. Cela signifie, pour tout $l = 1, .., N$:

2.1 Simuler y^{sim} suivant le modèle linéaire généralisé avec le vecteur de paramètres $\widehat{\beta_A}(y)$:

$$\begin{aligned} y^{sim} &\sim \text{Bernoulli}\left(\frac{1}{1+X_A\widehat{\beta_A}(y)}\right) \text{ ou} \\ y^{sim} &\sim \text{Poisson}\left(\exp(X_A\widehat{\beta_A}(y))\right). \end{aligned}$$

2.2 Calculer l'estimation du maximum de vraisemblance $\widehat{\beta_A}(y^{sim})$.

2.3 Générer une version de y^{sim} calibrée vers $\widehat{\beta_A}(y)$ suivant l'algorithme 3 :

$$y^{(l)}|y^{sim} \sim \mathrm{P}_{\widehat{\beta_A}(y^{sim})\to\widehat{\beta_A}(y)}.$$

3 Pour tout $l = 1, .., N$, réaliser la régression Lasso de $y^{(l)}$ sur X, et calculer $\lambda_A(y^{(l)})$.

4 Calculer la p-value conditionnelle empirique :

$$\widehat{p_A}(y) = \frac{1}{N}\sum_{l=1}^{N} \mathbf{1}\left\{\lambda_A(y^{(l)}) \geq \lambda_A(y)\right\}.$$

$\widehat{p_A}(y)$ a vocation à être interprétée de la même manière dans les modèles linéaires et linéaires généralisés. Cependant, les propriétés relatives à sa distribution sous l'hypothèse nulle (lemmes 8 et 10) ne sont démontrés que dans le cas linéaires. La construction de $\widehat{p_A}(y)$ dans les cas linéaires généralisés est conçue pour que ces propriétés y soient approximativement valables.

3.9 .Procédure de sélection de variables

Le test par simulation-calibration mesure la significativité d'une covariable sélectionnée par le Lasso. Lorsque l'on ne connaît a priori aucune variable active et l'on souhaite sélectionner un modèle entier, il est nécessaire d'utiliser le test de façon itérée.

3.9.1 .Notations et algorithme

Pour tout $k \in \{1, .., p\}$, soit j_k l'indice de de la k-ième variable sélectionnée par le Lasso et soit $A_k = \{j_1, .., j_k\}$ l'ensemble des k premières variables sélectionnées, avec $A_0 = \emptyset$. Chaque j_k est la première variable sélectionnée en-dehors de l'ensemble A_{k-1}, soit, d'après la notation introduite en 3.1, $j_k = j_{A_{k-1}}$. Il s'agit d'une légère variante du chemin du Lasso d'où les variables

ne « ressortent » pas : si il existe un λ tel que $\widehat{\beta_j}^{Lasso}(\lambda) \neq 0$ (ce qui se traduit par un k tel que $j \in A_k$), par définition j appartient à tous les $A_{k'}, k' > k$ même lorsqu'il existe des $\lambda' < \lambda$ tels que $\widehat{\beta_j}^{Lasso}(\lambda') = 0$.

Le test par simulation-calibration de $H_0\left(A_{k-1}\right)$ mesure donc la significativité de la variable j_k. De plus, puisque les variables ne « ressortent » pas de la suite d'ensembles, les hypothèses nulles $H_0\left(A_k\right)$ sont de moins en moins fortes : si $k < k'$, alors $A_k \subset A_{k'}$ donc comme les hypothèses portent sur les complémentaires de ces ensembles (sur lesquels elles affirment la nullité de β), $H_0\left(A_k\right)$ implique $H_0\left(A_{k'}\right)$.

Nous appliquons l'algorithme suivant :

Algorithme 5 (Procédure simple de sélection de variables par simulation-calibration)**.** Pour $k = 1, 2, ..$

1. Calculer $p_k = \widehat{p_{A_{k-1}}}(y)$ par simulation-calibration (algorithme 1 dans le cas linéaire, ou 4 dans les cas non linéaires).
2. • Si $p_k \leq \alpha$, continuer l'algorithme ;
   • sinon, sélectionner A_{k-1} et arrêter l'algorithme.

3.9.2 . Choix du critère d'arrêt

Dans cette procédure, plusieurs tests d'hypothèses sont réalisés. Cependant, comme elle s'arrête au premier test dont on ne rejette pas l'hypothèse nulle, elle n'est pas assimilable à une procédure de tests multiples proprement dite, dans laquelle un nombre potentiellement élevé de tests satisfont leur hypothèse nulle. Il n'est donc pas souhaitable de rendre plus exigeant le seuil de rejet des hypothèses comme le font les procédures de Bonferroni et de Benjamini et Hochberg (Benjamini et Hochberg, 1995), qui servent à contrôler les risques de faux positifs dus au grand nombre de p-values générées sous l'hypothèse nulle.

En revanche, il est possible d'adapter le critère d'arrêt de la procédure au

fait qu'elle évalue une suite de tests ordonnés. G'Sell *et al.* (2016) ont proposé, dans le problème général d'une suite de tests ordonnés dont on mesure les p-values $p_1, .., p_m$, le critère *ForwardStop* qui consiste à rejeter les $\hat{k}_F$ premiers tests où :

$$\begin{aligned} \hat{k}_F &= \max\left\{k \in \{1, .., m\} : p_k^{FS} \leq \alpha\right\} \\ p_k^{FS} &= -\frac{1}{k}\sum_{i=1}^{k} \log(1 - p_i). \end{aligned}$$

Ce critère a l'avantage de pouvoir, dans la grande majorité des cas, être évalué en connaissant seulement les premières valeurs de la suite (p_k), car la suite (p_k^{FS}) se calcule à partir des $p_i, i \leq k$ et elle est généralement croissante en k. En pratique, lorsque nous appliquons le ForwardStop, nous sélectionnons donc les $\hat{k}'_F$ premières variables où :

$$\hat{k}'_F = \min\left\{k \in \{1, .., m\} : p_k^{FS} > \alpha\right\} - 1$$

ce qui revient à remplacer la condition de poursuite de l'algorithme à l'étape 2 par $p_k^{FS} \leq \alpha$.

Par opposition au ForwardStop, le critère simple $p_k \leq \alpha$ est appelé *seuillage* (en anglais *thresholding*). Le seuillage permet de contrôler à la fois le FWER et le FDR au niveau α (Marcus *et al.*, 1976), tandis que le ForwardStop, généralement moins conservatif, permet de contrôler le FDR au niveau α (G'Sell *et al.*, 2016).

Il est important de remarquer que ces résultats de contrôle sont valables portent sur l'erreur de première espèce, c'est-à-dire le rejet à tort de l'hypothèse nulle lorsque celle-ci est vérifiée. Or dans la procédure séquentielle, on teste fréquemment des $H_0(A)$ qui ne sont pas vérifées : notamment à la première étape de la procédure, où $A = \emptyset$ donc il suffit qu'il existe une va-

riable active pour que $H_0(A)$ soit incorrecte. Il se peut qu'il existe une ou des variables actives en-dehors de A mais qu'elles ne sont sélectionnées par le Lasso qu'à des faibles valeurs de λ et que la première variable sélectionnée en-dehors de A soit une variable inactive. Dans ce cas, retenir cette variable constitue un faux positif du point de vue de la sélection de variable, mais pas une erreur de première espèce puisque $H_0(A)$ n'est pas vérifiée. La section 3.10 présente un résultat de contrôle de cette occurence.

3.10 .Théorème étendu : contrôle de l'erreur de sélection

3.10.1 .Introduction et énoncé du théorème

On considère que l'ensemble des covariables se divise en trois sous-ensembles disjoints : $\{1, .., p\} = A \cup B \cup C$ avec $\forall j \in C, \beta_j = 0$. On effectue le test par simulation-calibration de l'hypothèse $H_0(A)$ dans une optique de sélection de variables, c'est-à-dire, si le test rejette $H_0(A)$, on sélectionne la variable (presque sûrement unique) $j_A(Y) \in B \cup C$ telle que $\widehat{\beta_{j_A(Y)}}^{Lasso}(\lambda) \neq 0$ pour λ au voisinage de λ_A . Cette sélection constitue un vrai positif si $j_A(Y) \in B$ et un faux positif si $j_A(Y) \in C$. Puisque $H_0(A)$ n'est pas vérifiée (car il peut exister des $j \in B$ tels que $\beta_j \neq 0$), les résultats des sections précédentes ne s'appliquent pas. Cependant, nous avons le résultat suivant :

Théorème 1. *Supposons que les variables actives sont orthogonales aux variables inactives, c'est-à-dire :*

$$X_C^T X_{A \cup B} = 0.$$

Alors, le test par simulation-calibration de $H_0(A)$ au niveau α a une probabilité inférieure à $\alpha + \frac{1-\alpha}{N+1}$ de sélectionner un faux positif.

Comme sous l'hypothèse nulle (section 3.7.2), ce résultat peut également être vu comme un contrôle au niveau α du risque de sélection d'un faux positif

si le critère de sélection est $\widehat{p_A}(y) \leq \alpha - \frac{1-\alpha}{N}$.

Ce théorème rend applicables les propriétés de contrôle du FWER et du FDR de la procédure de sélection de variable avec arrêt par seuillage et ForwardStop respectivement, à condition que les variables actives soient orthogonales aux variables inactives.

3.10.2 . Preuve du théorème

Démonstration. Soit $\widehat{\beta}^{Lasso-A\cup C}(\lambda, Y)$ l'estimé par Lasso au paramètre λ du modèle linéaire restreint à $A \cup CB$, et soit :

$$\lambda_{A-C}(Y) = \sup \left\{ \lambda \geq 0 : \exists j \in C, \widehat{\beta}_j^{Lasso-A\cup C}(\lambda, Y) \neq 0 \right\}.$$

On définit $\beta^{Lasso-A\cup B}(\lambda, Y)$ et $\lambda_{A-B}(Y)$ de façon similaire en remplaçant C par B, et on note $\beta^{Lasso-A}(\lambda, Y)$ l'estimé par Lasso du modèle restreint à A. Puisque les chemins du Lasso de $\widehat{\beta}^{Lasso-A\cup B}$, $\widehat{\beta}^{Lasso-A\cup C}$ et $\widehat{\beta}^{Lasso}$ coïncident tant que seules des variables appartenant à A ont été sélectionnées, on a $\lambda_A(Y) = \max(\lambda_{A-B}(Y), \lambda_{A-C}(Y))$. Le test produit un faux positif lorsqu'à la fois $\widehat{p_A}(Y) \leq \alpha$ (rejet de $H_0(A)$ et sélection d'une variable) et $\lambda_{A-C}(Y) \geq \lambda_{A-B}(Y)$ (la variable sélectionnée est un faux positif). $\widehat{p_A}(Y)$ prend des petites valeurs lorsque $\lambda_A(Y)$ (égal à $\lambda_{A-C}(Y)$ dans le cas d'un faux positif) est élevé. On cherche donc à dominer la loi de $\lambda_{A-C}(Y)$.

Soit $\lambda > 0$. Par définition de λ_{A-C},

$$\lambda_{A-C}(Y) > \lambda \Leftrightarrow \exists \lambda' > \lambda : \exists j \in C : \beta_j^{Lasso-A\cup C}(\lambda', Y) \neq 0$$

En tout point du chemin du Lasso, la covariance des résidus avec chaque variable active est égale à λ'. Donc :

$$\lambda_{A-C}(Y) > \lambda \Leftrightarrow \exists \lambda' > \lambda : \exists j \in C : \left| \mathrm{Cov}\left(Y - X_{A\cup C}\widehat{\beta}^{Lasso-A\cup C}(\lambda', Y), X_j \right) \right| = \lambda'$$

De plus, tant qu'aucune variable en-dehors de A n'est sélectionnée, $\widehat{\beta}^{Lasso-A\cup C}$ et $\widehat{\beta}^{Lasso-A}$ coïncident. D'où :

$$\lambda_{A-C}(Y) > \lambda \Leftrightarrow \exists \lambda' > \lambda : \exists j \in C : \left|\mathrm{Cov}\left(Y - X_A \widehat{\beta}^{Lasso-A}(\lambda', Y), X_j\right)\right| = \lambda'.$$

Comme $\forall j \in C, X_j^T X_A = 0$, on a :

$$\begin{aligned} \lambda_{A-C}(Y) > \lambda &\Leftrightarrow \exists \lambda' > \lambda : \exists j \in C : |\mathrm{Cov}(Y, X_j)| = \lambda' \\ \lambda_{A-C}(Y) > \lambda &\Leftrightarrow \exists j \in C : |\mathrm{Cov}(Y, X_j)| > \lambda. \end{aligned} \tag{3.6}$$

Donc $\lambda_{A-C}(Y) = \max_{j \in C} |\mathrm{Cov}(Y, X_j)|$.

Soient Y_0 et Y_1 deux vecteurs aléatoires qui satisfont le modèle 3.1, avec la même variance résiduelle σ^2 et des vecteurs de coefficients, respectivement β_0 et β_1, qui ne diffèrent que par leurs coefficients sur B : $\beta_{0A} = \beta_{1A}$ est quelconque, $\beta_{0B} = 0$ alors que β_{1B} est quelconque, et $\beta_{0C} = \beta_{1C} = 0$. Autrement dit, Y_0 satisfait $H_0(A)$ et Y_1 ne la satisfait pas nécessairement. On cherche à prouver le théorème sur le vecteur quelconque Y_1, en s'appuyant sur des comparaisons avec le cas connu de l'hypothèse nulle (Y_0).

Soit également $\theta'_A = (\beta'_A, \sigma'_A) \in \Theta_A$ un vecteur de paramètres du modèle restreint à A. Dans les quatre sous-sections suivantes, nous allons prouver que la loi de $\lambda_{A-C}(Y_1)$ conditionnellement à $\widehat{\theta_A}(Y_1) = \theta'_A$ est stochastiquement dominée par la loi de $\lambda_{A-C}(Y_0)$ conditionnellement à $\widehat{\theta_A}(Y_0) = \theta'_A$.

Décomposition du vecteur Y

On considère la régression linéaire de chacune des variables appartenant à B sur celles appartenant à A :

$$X_B = X_A \Pi_{AB} + \widetilde{X}_B, \;\; \widetilde{X}_B^T X_A = 0.$$

Elle permet de définir, en $Y = Y_0$ ou $Y = Y_1$, la composante de l'espérance de Y due exclusivement aux variables de B :

$$y_B = \widetilde{X}_B \beta_B.$$

Par hypothèse, $y_{B0} = 0$ mais y_{B1} peut être non nul. On le met sous la forme $y_{B1} = ||y_{B1}|| u_{B1}$ où u_{B1} est un vecteur unitaire (si $y_{B1} = 0$, on prend pour u_{B1} un vecteur unitaire quelconque dans $\mathrm{Vect}\left(\widetilde{X}_B\right)$).

Le modèle 3.1 s'écrit alors :

$$Y = X_A \left(\beta_A + \Pi_{AB} \beta_B\right) + y_B + \epsilon$$

ou encore, puisque $y_{B0} = 0$, l'écriture suivante valide à la fois en $h = 0$ et $h = 1$, en notant $\gamma_h = ||y_{Bh}||$:

$$Y_h = X_A \left(\beta_A + \Pi_{AB} \beta_{hB}\right) + \gamma_h u_{B1} + \epsilon_h, \quad \epsilon_h \sim \mathcal{N}\left(0, \sigma^2 I_n\right). \tag{3.7}$$

On considère la régression linéaire associée à cette écriture, c'est-à-dire la régression de Y sur les variables X_A et le vecteur u_{B1} :

$$Y = X_A \widehat{\beta}_A + \widehat{\gamma} u_{B1} + r, \;\; r^T X_A = 0, \; r^T u_{B1} = 0. \tag{3.8}$$

En raison de l'orthogonalité entre les variables de cette régression (u_{B1} est une combinaison linéaire des $\widetilde{X}_B$ donc $u_{B1}^T X_A = 0$), le $\widehat{\beta}_A$ obtenu est le même que dans la régression linéaire de Y sur X_A seulement.

$\widehat{\beta}_A$, $\widehat{\gamma}$ et r sont toutes trois des fonctions de Y, donc sont des variables aléatoires. On va caractériser leurs lois, en distinguant si il y a lieu entre $Y = Y_0$ et $Y = Y_1$, puis leurs lois conditionnellement à $\widehat{\theta_A}(Y) = \theta'_A$.

Lois non conditionnelles des composantes de la régression

$X_A\widehat{\beta}_A$, $\widehat{\gamma}u_{B1}$ et r sont les projetés orthogonaux de Y sur les espaces orthogonaux entre eux que sont, respectivement, $\mathrm{Vect}(X_A)$, $\mathrm{Vect}(u_{B1})$, et le supplémentaire orthogonal de ces deux espaces, noté V_r. Comme Y est un vecteur gaussien, ce sont tous trois des vecteurs gaussiens, et ils sont centrés en les projetés orthogonaux de $\mathrm{E}[Y] = X_A\left(\beta_A + \Pi_{AB}\beta_{hB}\right) + \gamma_h u_{B1}$ sur ces trois espaces (autrement dit, la régression linéaire est sans biais) :

$$\begin{aligned} E\left[X_A\widehat{\beta}_A\right] &= \beta_A + \Pi_{AB}\beta_{hB} \\ E\left[\widehat{\gamma}u_{B1}\right] &= \gamma_h u_{B1} \\ E[r] &= 0. \end{aligned}$$

De plus la matrice de variance-covariance de Y, égale à $\sigma^2 I_n$, s'écrit encore $\sigma^2 I_n$ dans une base orthogonale portée par ces trois sous-espaces. Donc les trois vecteurs gaussiens $X_A\widehat{\beta}_A$, $\widehat{\gamma}u_{B1}$ et r sont indépendants (les variables $\widehat{\beta}_A$, $\widehat{\gamma}$ et r sont donc indépendantes) et leurs variance projetée sur toute dimension de leur sous-espace support respectif est toujours σ^2. En particulier,

* r suit la loi normale multivariée centrée de matrice de variance-covariance $I_{\dim(V_r)} = I_{p-|A|-1}$, c'est-à-dire la loi à densité sur sur V_r :

$$f_r(u) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^{p-|A|-1}} e^{-\frac{||u||^2}{2\sigma^2}}.$$

  Cette loi est la même que l'on considère $r(Y_0)$ ou $r(Y_1)$.
* $\widehat{\gamma} \sim \mathcal{N}\left(\gamma_h, \sigma^2\right)$. L'espérance varie donc selon que l'on considère Y_0 ou Y_1 : $\widehat{\gamma}(Y_0) \sim \mathcal{N}\left(0, \sigma^2\right)$ et $\widehat{\gamma}(Y_1) \sim \mathcal{N}\left(\gamma_1, \sigma^2\right)$.

Ce résultat implique une forme forte de dominance de la loi de $|\widehat{\gamma}(Y_1)|$ sur celle de $|\widehat{\gamma}(Y_0)|$:

Définition 4 (propriété du rapport de vraisemblance monotone ou *monotone likelihood ratio property*)**.** Soient P_0 et P_1 deux lois de probabilités sur $\mathbb{R}$ admettant des densités f_0 et f_1 de même support S. On dit que P_1 vérifie la propriété du rapport de vraisemblance monotone sur P_0 si $\frac{f_1}{f_0}$ est croissante sur S.

Lemme 12. *La loi de $|\widehat{\gamma}(Y_1)|$ vérifie la propriété du rapport de vraisemblance monotone sur celle de $|\widehat{\gamma}(Y_0)|$.*

Démonstration. Pour $h = 0$ ou $h = 1$, pour tout $t \in \mathbb{R}_+$,

$$f_{|\widehat{\gamma}|}(t) = f_{\widehat{\gamma}}(t) + f_{\widehat{\gamma}}(-t)$$

d'où :

$$\begin{aligned}
\frac{f_{|\widehat{\gamma}(Y_1)|}}{f_{|\widehat{\gamma}(Y_0)|}}(t) &= \frac{e^{-\frac{(t-\gamma_1)^2}{2\sigma^2}} + e^{-\frac{(t+\gamma_1)^2}{2\sigma^2}}}{2e^{-\frac{t^2}{2\sigma^2}}} \\
&= \frac{1}{2}\left(e^{-\frac{(t-\gamma_1)^2-t^2}{2\sigma^2}} + e^{-\frac{(t+\gamma_1)^2-t^2}{2\sigma^2}}\right) \\
&= \frac{1}{2}e^{-\frac{\gamma_1^2}{2\sigma^2}}\cosh\left(\frac{\gamma_1 t}{\sigma^2}\right)
\end{aligned}$$

La fonction cosinus hyperbolique est croissante sur $\mathbb{R}_+$, donc la propriété du rapport de vraisemblance monotone est vérifiée. □

La propriété du rapport de vraisemblance monotone implique la dominance stochastique de P_1 sur P_0 (Roosen et Hennessy, 2004). De plus, elle est conservée par l'application aux variables aléatoires d'une même fonction dérivable h strictement croissante, car celle-ci revient à introduire un terme multiplicatif qui se simplifie dans le rapport des densités :

$$\frac{f_{h(U_1)}}{f_{h(U_0)}}(h(t)) = \frac{h'(t)f_{U_1}(t)}{h'(t)f_{U_0}(t)} = \frac{f_{U_1}}{f_{U_0}}(t).$$

Donc, en apliquant la fonction carré, la loi de $\widehat{\gamma}(Y_1)^2$ vérifie la propriété du rapport de vraisemblance monotone sur celle de $\widehat{\gamma}(Y_0)^2$.

Loi conditionnelle du vecteur des résidus

Ayant obtenu ces résultats sur les lois non conditionnelles de $\widehat{\beta}_A(Y)$, $\widehat{\gamma}(Y)$ et $r(Y)$, on va en déduire des résultats sur la loi de la troisième de ces composantes, le vecteur des résidus $r(Y)$, conditionnellement à $\widehat{\theta_A}(Y) = \theta'_A$. Le résidu de la régression de Y sur X_A seulement est égal à $\widehat{\gamma}(Y)u_{B1} + r(Y)$, d'où :

$$n\widehat{\sigma_A}^2 = ||\widehat{\gamma}u_{B1} + r(Y)||^2 = \widehat{\gamma}(Y)^2 + ||r(Y)||^2.$$

La condition $\widehat{\theta_A}(Y) = \theta'_A$ est donc équivalente à :

$$\widehat{\beta}_A(Y) = \beta'_A \ , \ \widehat{\gamma}(Y)^2 + ||r(Y)||^2 = n{\sigma'_A}^2. \tag{3.9}$$

Comme $\widehat{\beta}_A(Y)$, $\widehat{\gamma}(Y)$ et $r(Y)$ sont indépendantes, seule la deuxième égalité a un effet sur les lois de $\widehat{\gamma}(Y)$ et de $r(Y)$. Elle va permettre de passer de la dominance stochastique entre les $\widehat{\gamma}(Y)$ obtenue dans la section précédente à une forme de dominance stochastique entre les $r(Y)$. Pour cela, nous avons besoin d'un résultat de conservation de la dominance par passage à une loi conditionnelle. Il porte sur la forme plus forte de dominance, le rapport de vraisemblance monotone, et non sur la dominance stochastique. Son équivalent pour la dominance stochastique n'est pas vrai.

Lemme 13. *Soient U_0 et U_1 deux variables aléatoires réelles de même support telles que la loi de U_1 vérifie la propriété du rapport de vraisemblance monotone sur celle de U_0, et V une variable aléatoire réelle indépendante de U_0 et de U_1 et suivant une loi à densité.*

Alors les lois de $U_0 + V$ et de $U_1 + V$ possèdent le même support et pour tout a

appartenant à celui-ci, la loi de U_1 conditionnellement à $U_1+V=a$ vérifie la propriété du rapport de vraisemblance monotone sur celle de U_0 conditionnellement à $U_0+V=a$.

Démonstration. Soient f_0 et f_1 les densité de probabilités des lois de U_0 et U_1, S leur support et g la densité de probabilité de la loi de V. Le support de U_0+V ainsi que de U_1+V est $S'=\{a : \int_S f_h(u)g(a-u)du>0\}$.

Soit $a\in S'$. Pour $h=0$ ou 1, soit $\tilde{f}_{a,h}$ la densité de la loi de U_h conditionnellement à $U_h+V=a$. Elle s'écrit :

$$\forall t\in\mathbb{R}, \tilde{f}_{a,h}(t)=\frac{f_h(t)g(a-t)}{\int_S f_h(u)g(a-u)du}.$$

Donc les lois de U_0 conditionnellement à $U_0+V=a$ et de U_1 conditionnellement à $U_1+V=a$ ont le même support, $S''(a)=\{t\in S : g(a-t)>0\}$, et :

$$\forall t\in S''(a),\ \frac{\tilde{f}_{a,1}}{\tilde{f}_{a,0}}(t)=\frac{f_1(t)}{f_0(t)}\frac{\int_S f_0(u)g(a-u)du}{\int_S f_1(u)g(a-u)du}.$$

a étant fixé, le rapport des deux intégrales est une constante. Le rapport f_1/f_0 est croissant d'après la propriété du rapport de vraisemblance monotone de la loi de U_1 sur celle de U_0. Donc la propriété est vérifiée par les lois conditionnelles. □

En appliquant ce lemme à la deuxième partie du conditionnement 3.9, comme la loi de $\widehat{\gamma}(Y_1)^2$ vérifie la propriété du rapport de vraisemblance monotone sur celle de $\widehat{\gamma}(Y_0)^2$ (conséquence du lemme 12), la loi de $\widehat{\gamma}(Y_1)^2$ conditionnellement à $\widehat{\theta_A}(Y_1)=\theta'_A$ vérifie la propriété du rapport de vraisemblance monotone sur celle de $\widehat{\gamma}(Y_0)^2$ conditionnellement à $\widehat{\theta_A}(Y_0)=\theta'_A$. Donc elle la domine stochastiquement.

Nous pouvons maintenant démontrer le résultat suivant, qui est une forme forte et multivariée de dominance stochastique de la loi conditionnelle de $r(Y_0)$ sur celle de $r(Y_1)$.

Lemme 14. *Il existe deux vecteurs aléatoires sur V_r, R_0 et R_1, telles que :*

- *Pour $h = 0$ ou $h = 1$, la loi de R_h est identique à la loi de $r(Y_h)$ conditionnellement à $\widehat{\theta_A}(Y_h) = \theta'_A$.*
- *Presque sûrement (p. s.), R_0 et R_1 sont positivement colinéaires et $||R_0|| \geq ||R_1||$.*

Cette caractérisation est un analogue multivarié du critère de dominance entre variables univariées introduit par le lemme 9.

Démonstration. Comme la loi de $\widehat{\gamma}(Y_0)^2$ conditionnellement à $\widehat{\theta_A}(Y_0) = \theta'_A$ est stochastiquement dominée par celle de $\widehat{\gamma}(Y_1)^2$ conditionnellement à $\widehat{\theta_A}(Y_1) = \theta'_A$, d'après le le lemme 9 il existe deux variables aléatoires G_0 et G_1, de même loi que ces deux lois conditionnelles respectivement, et telles que $G_0 \leq G_1$ p. s.

Pour tout $d \geq 0$, on note $S_{V_r,d^2} = \left\{u \in V_r : ||u||^2 = d^2\right\}$ la sphère centrée de rayon d dans l'espace des résidus V_r. La densité de probabilité de $r(Y)$ est proportionnelle à $u \to e^{-\frac{1}{2||u||^2}}$ donc elle est constante sur chacune des S_{V_r,d^2}.

Soit U un vecteur aléatoire de loi uniforme sur la sphère unité $S_{V_r,1}$. On définit R_0 et R_1 par :

$$\forall h \in \{0,1\},\ R_h = \sqrt{n{\sigma'_A}^2 - G_h}\, U.$$

R_0 et R_1 sont positivement colinéaires et puisque $G_0 \leq G_1$ p. s., $||R_0|| \geq ||R_1||$ p. s. Il reste à prouver que R_0 et R_1 suivent les lois désirées.

Soit $\gamma' \in \mathbb{R}$. La double condition $\left(\widehat{\theta_A}, \widehat{\gamma}^2\right)(Y) = \left(\theta'_A, {\gamma'}^2\right)$ est équivalente

à :

$$\widehat{\beta}_A(Y) = \beta'_A \ , \ \widehat{\gamma}(Y)^2 = \gamma'^2 \ , \ ||r(Y)||^2 = n{\sigma'_A}^2 - \gamma'^2$$

Comme $\widehat{\beta}_A(Y)$, $\widehat{\gamma}(Y)$ et $r(Y)$ sont indépendantes, seule la troisième égalité a un effet sur la loi de $r(Y)$. La loi de $r(Y)$ sous la double condition est donc la loi de $r(Y)$ conditionnellement à $r(Y) \in S_{V_r, n{\sigma'_A}^2 - \gamma'^2}$, c'est-à-dire, puisque la densité de $r(Y)$ est constante sur cet ensemble, une loi uniforme :

$$r(Y) | \left[\left(\widehat{\theta_A}, \widehat{\gamma}^2 \right)(Y) = \left(\theta'_A, \gamma'^2 \right) \right] \sim \mathcal{U}\left(S_{V_r, n{\sigma'_A}^2 - \gamma'^2} \right).$$

De plus, en $h = 0$ ou $h = 1$, d'après sa définition R_h suit la loi conditionnelle :

$$R_h | \left[\widehat{\gamma}^2(Y) = \gamma'^2 \right] \sim \mathcal{U}\left(S_{V_r, n{\sigma'_A}^2 - \gamma'^2} \right)$$

et la loi de G_h est identique à la loi de $\widehat{\gamma}(Y_h)^2$ conditionnellement à $\widehat{\theta_A}(Y_h) = \theta'_A$. La loi de R_h sans conditionnement et la loi de $r(Y_h)$ sous le seul conditionnement $\widehat{\theta_A}(Y_h) = \theta'_A$ s'obtiennent donc de la même façon, en intégrant la loi $\mathcal{U}\left(S_{V_r, n{\sigma'_A}^2 - \gamma'^2} \right)$ en γ'^2 selon la loi de G_h. Donc elles sont égales. □

Conclusion sur la loi conditionnelle de λ_{A-C}

Nous reprenons l'équivalence (3.6), qui caractérise la sélection d'un faux positif, en y injectant la décomposition de Y en projetés sur les différents sous-espaces (3.8). Les deux premiers termes de celle-ci sont des vecteurs orthogonaux aux $X_j, j \in C$ donc $\forall j \in C,\ \mathrm{Cov}(Y, X_j) = \mathrm{Cov}(r(Y), X_j)$. Donc pour tout $\lambda > 0$,

$$\lambda_{A-C}(Y) \geq \lambda \quad \Leftrightarrow \quad \left(\max_{j \in C} |\mathrm{Cov}(r(Y), X_j)| \geq \lambda \right)$$

$$\mathrm{P}\left(\lambda_{A-C}(Y) \geq \lambda \mid \widehat{\theta_A}(Y) = \theta'_A \right) \quad = \quad \mathrm{P}\left(\max_{j \in C} |\mathrm{Cov}(r(Y), X_j)| \geq \lambda \mid \widehat{\theta_A}(Y) = \theta'_A \right).$$

La loi de $\lambda_{A-C}(Y)$ sous le conditionnement par $\widehat{\theta_A}(Y)$ ne dépend donc que de la loi de $r(Y)$ sous ce même conditionnement. Celle-ci est décrite par le lemme 14, qui permet de distinguer selon que l'hypothèse nulle est vérifiée ou non. Pour $h = 0$ ou 1,

$$\mathrm{P}\left(\lambda_{A-C}(Y_h) \geq \lambda \mid \widehat{\theta_A}(Y) = \theta'_A\right) = \mathrm{P}\left(\max_{j \in C} |\mathrm{Cov}(R_h, X_j)| \geq \lambda \mid \widehat{\theta_A}(Y) = \theta'_A\right).$$

Or R_0 et R_1 sont positivement colinéaires avec $||R_0|| \geq ||R_1||$ p. s. Donc :

$$\begin{array}{rcl} \forall j \in C,\ |\mathrm{Cov}(R_0, X_j)| & \geq & |\mathrm{Cov}(R_1, X_j)| \text{ p.s.} \\ \left(\max_{j \in C} |\mathrm{Cov}(R_1, X_j)| \geq \lambda\right) & \Longrightarrow & \left(\max_{j \in C} |\mathrm{Cov}(R_0, X_j)| \geq \lambda\right) \text{ p.s.} \\ \mathrm{P}\left(\lambda_{A-C}(Y_1) \leq \lambda \mid \widehat{\theta_A}(Y_1) = \theta'_A\right) & \geq & \mathrm{P}\left(\lambda_{A-C}(Y_0) \geq \lambda \mid \widehat{\theta_A}(Y_0) = \theta'_A\right). \end{array}$$

Ceci étant vrai pour tout λ, la loi de $\lambda_{A-C}(Y_0)$ conditionnellement à $\widehat{\theta_A}(Y_0) = \theta'_A$ domine stochastiquement la loi de $\lambda_{A-C}(Y_1)$ conditionnellement à $\widehat{\theta_A}(Y_1) = \theta'_A$.

Conclusion de la preuve du théorème

Le test produit un faux positif si et seulement si $\lambda_{A-C}(Y) \geq \lambda_{A-B}(Y)$ et $\widehat{p_A}(Y) \leq \alpha$, $\widehat{p_A}(Y)$ étant elle-même un estimateur de la p-value conditionnelle $p_A(Y)$. On agrège cette double condition en une seule en définissant les variables aléatoires $\tilde{\lambda}_{A-C}(Y)$, $\tilde{p_A}(Y)$ et $\tilde{p_A}(Y)$ par :

$$\begin{array}{rcl} \text{si } \lambda_{A-C}(Y) \geq \lambda_{A-B}(Y) & : & \tilde{\lambda}_{A-C}(Y) = \lambda_{A-C}(Y) = \lambda_A(Y) \\ & : & \tilde{p}_A(Y) = p_A(Y) \\ & : & \widehat{\tilde{p}_A} = \widehat{p_A}(Y) \\ \text{sinon} & : & \tilde{\lambda}_{A-C}(Y) = 0 \\ & : & \tilde{p}_A(Y) = \widehat{\tilde{p}_A}(Y) = 1. \end{array}$$

Alors le test produit un faux positif si et seulement si $\widehat{\tilde{p}_A}(Y) \leq \alpha$. Nous allons établir des dominances stochastiques entre les lois de ces variables à tilde, avec ou sans conditionnement, en $Y = Y_0$ ou $Y = Y_1$.

Sous la condition $\widehat{\theta_A}(Y) = \theta'_A$, par reformulation de la définition 3, $p_A(Y)$ s'obtient en appliquant à $\lambda_A(Y)$ une fonction décroissante :

$$\begin{aligned} p_A(Y) &= \phi_{\theta'_A}(\lambda_A(Y)) \quad \text{où :} \\ \forall \lambda \geq 0,\ \phi_{\theta'_A}(\lambda) &= \mathrm{P}\left(\lambda_A(Y_0) \geq \lambda \mid \widehat{\theta_A}(Y_0) = \theta'_A\right). \end{aligned}$$

De plus, puisque $\phi_{\theta'_A}(0) = 1$, on a $\tilde{p}_A(Y) = \phi_{\theta'_A}\left(\tilde{\lambda}_{A-C}(Y)\right)$.

$\lambda_A(Y) \geq \lambda_{A-C}(Y) \geq \tilde{\lambda}_{A-C}(Y)$ donc la loi de $\lambda_A(Y_0)$ conditionnellement à $\widehat{\theta_A}(Y_0) = \theta'_A$ domine stochastiquement la loi de $\tilde{\lambda}_{A-C}(Y_0)$ sous le même conditionnement. Celle-ci domine elle-même la loi de $\tilde{\lambda}_{A-C}(Y_1)$ conditionnellement à $\widehat{\theta_A}(Y_1) = \theta'_A$ (sous-section 3.10.2).

Donc par décroissance de $\phi_{\theta'_A}$, la loi de $\tilde{p}_A(Y_1)$ conditionnellement à $\widehat{\theta_A}(Y_1) = \theta'_A$ domine stochastiquement la loi de $p_A(Y_0)$ conditionnellement à $\widehat{\theta_A}(Y_0) = \theta'_A$. Or, par construction, cette dernière domine elle-même la loi uniforme sur $[0, 1]$ (équation (3.4) de la preuve du lemme 3).

La loi conditionnelle de $\tilde{p}_A(Y_1)$ domine la loi uniforme pour toute valeur de θ'_A donc la loi non-conditionnelle de $\tilde{p}_A(Y_1)$ domine elle-même la loi uniforme.

Enfin, puisque $N\widehat{p_A}(Y) \mid p_A(Y) \sim \mathrm{Bin}(N, p_A(Y))$ où N est le nombre de simulations réalisées (lemme 7), et que la loi $\mathrm{Bin}(N, 1)$ est celle d'une constante p.s. égale à N, on a aussi $N\widehat{\tilde{p}_A}(Y) \mid p_A(Y) \sim \mathrm{Bin}(N, \tilde{p}_A(Y))$. Donc, comme dans le lemme 10, la dominance stochastique de la loi uniforme sur $[0, 1]$ par la loi de $\tilde{p}_A(Y_1)$ entraîne la dominance stochastique de la loi uniforme discrète sur $\{0, 1/N, .., 1\}$ par la loi de $\widehat{\tilde{p}_A}(Y_1)$.

Donc, comme sous l'hypothèse nulle, la probabilité de $\widehat{\tilde{p}_A}(Y_1) \leq \alpha$ (c'est-à-dire de sélection d'un faux positif) est majorée par $\alpha + \frac{1-\alpha}{N+1}$.

□

3.11 .Études de simulations

Pour mesurer les performances du test par simulation-calibration, nous avons réalisé une double étude de simulations.

D'une part, nous avons mesuré la distribution de la p-value générée sous $H_0(A)$ dans un grand nombre de scénarios différents. Par construction du test, cette distribution est censée être une loi uniforme sur $[0, 1]$. Plus précisément, nous avons prouvé au lemme 8 que dans le modèle linéaire, sous une hypothèse technique dite « de continuité », $\widehat{p_A}(Y)$ suit une loi uniforme sur un analogue discret de $[0, 1]$ (l'ensemble $\{0, 1/N, .., 1 - 1/N, 1\}$, N étant le nombre de simulations réalisées par l'algorithme), et au lemme 10 qu'en relâchant cette hypothèse, dans le modèle linéaire $\widehat{p_A}(Y)$ domine stochastiquement la loi uniforme, c'est-à-dire prend des valeurs systématiquement plus élevées. En revanche, nous n'avons pas de résultat équivalent dans les modèles linéaires généralisés, même si les propriétés de la calibration non linéaire démontrées dans la partie 3.8.2 permettent de s'attendre à une distribution proche de la loi uniforme aussi dans ces modèles.

D'autre part, nous avons mesuré les performances des procédures de sélection de variables — à la fois avec le critère d'arrêt par seuillage et le critère d'arrêt ForwardStop — selon trois métriques usuelles de la sélection de variables : le family-wise error rate (FWER) le taux de fausses découvertes (FDR) et la sensibilité. Les propriétés du seuillage et du ForwardStop permettent en effet de s'attendre à un contrôle du FWER par le premier et du FDR par le second, au moins dans les conditions où le théorème étendu s'applique : modèle linéaire et absence de corrélation entre covariables. De plus nous avons comparé ces performances à celles d'une procédure équivalente fondée sur

le CovTest de Lockhart *et al.* (2014).

3.11.1 . Plan de simulation

Le plan de simulation suivant est commun à l'étude de la distribution de la p-value sous l'hypothèse nulle et à celle de la procédure de sélection de variables. Nous simulons $n_{\text{sim}} = 500$ jeux de données pour chacun des 252 jeux de paramètres (ou scénarios). Dans tous les cas, le nombre d'observations est $n = 1000$ et le nombre de covariables est $p = 500$. Les paramètres variant selon le scénario sont :

- Le type de modèle : linéaire, binaire à données denses, binaire à données creuses, ou de Poisson. Les modèles binaires à données denses sont définis par $\mathrm{E}[Y|X = 0] = 0.5$ et ceux à données creuses par $\mathrm{E}[Y|X = 0] = 0.1$, la valeur de l'intercept β_0 permettant de distinguer ces deux cas.
- la matrice de corrélation utilisée pour simuler les régresseurs : une matrice de Toeplitz de coefficients $\rho_{(i,j)} = \rho^{|i-j|}$, avec $\rho = 0$, $\rho = 0.9$ ou $\rho = 0.99$. Ces valeurs élevées sont celles utilisées par Sabourin *et al.* (2015) dans leur étude de simulation.
- Le nombre de variables actives : $0, 1, 2, 5$ ou 10. Elles sont tirées uniformément parmi les 500 covariables.
- Pour les scénarios ayant au moins une variable active, le rapport signal-bruit (défini en 2.4.2) :

$$\begin{aligned} \mathrm{SNR}(X, \beta) &= \frac{\frac{1}{n-1}\Sigma_{i=1}^{n}\left(E_\beta[Y_i \mid X] - \frac{1}{n}\Sigma_{i=1}^{n}E_\beta[Y_i \mid X]\right)^2}{\frac{1}{n}\Sigma_{i=1}^{n}\mathrm{Var}_\beta(Y_i \mid X)} \\ &= 1, 0.3, 0.1, 0.03, \text{ ou } 0.01. \end{aligned}$$

Il mesure la force du signal porté par par les variables actives. Les scénarios à fort rapport signal-bruit sont donc les plus faciles du point de

vue de la sélection de variables. Les scénarios à 0 variable actives on nécessairement un rapport signal-bruit nul.

3.11.2 . p-value sous l'hypothèse nulle

Pour vérifier que le test suit son comportement attendu sous l'hypothèse nulle, nous supposons connu l'ensemble A des $0, 1, 2, 5$ ou 10 régresseurs actifs et nous testons $H_0(A)$, qui énonce qu'il n'existe pas d'autre variable active. Pour chaque scénario, dans chacun des $n_{\text{sim}} = 500$ jeux de données s caractérisé par A_s, X_s et Y_s, nous produisons un $\hat{p}_s = \widehat{p_{A_s}}(Y_s, X_s)$. Il est calculé par l'algorithme de simulation-calibration fondé sur $N = 100$ simulations de vecteurs réponse calibrés.

Dans un des scénarios (modèle linéaire, $\rho = 0.99$, 1 régresseur actif, $\text{SNR} = 1$), à titre d'exemple, nous avons également produit pour chacun des 500 jeux de données une estimation naïve de la p-value non conditionnelle $p^0_A(y)$ (voir la section 3.2). Elle est obtenue par Monte-Carlo sans l'étape de calibration, c'est-à-dire en appliquant l'algorithme 1 sans l'étape de calibration (2.3.) en prenant $\theta^{sim}_A = \widehat{\theta_A}(y)$. Cela a pour but d'illustrer l'impact de la calibration sur la distribution des p-values produites.

Pour chaque scénario, la population des $(\hat{p}_s)_{1 \leq s \leq n_{\text{sim}}}$ obtenues par simulation-calibration ne doit idéalement pas pouvoir être distinguée d'un échantillon simulé selon la loi uniforme sur $[0, 1]$ (à l'approximation près entre $[0, 1]$ et l'ensemble discret $\{0, 1/N, .., 1 - 1/N, 1\}$ auquel appartiennent les $\hat{p}_s$). Graphiquement, les adéquations ou non avec la loi uniforme s'observent par des diagrammes quantile-quantile où pour tout $s = 1, .., n_{\text{sim}}$, la s-ième plus petite valeur $\hat{p}_{(s)}$ est représentée au point de coordonnées $\left(\frac{s}{n_{\text{sim}}}, \hat{p}_{(s)}\right)$.

Les diagrammes de la figure 3.1 illustrent la nécessité de l'étape de calibration pour que les p-values empiriques produites soient valides. Dans l'exemple choisi, on observe l'adéquation des p-values produites par simulation-calibration

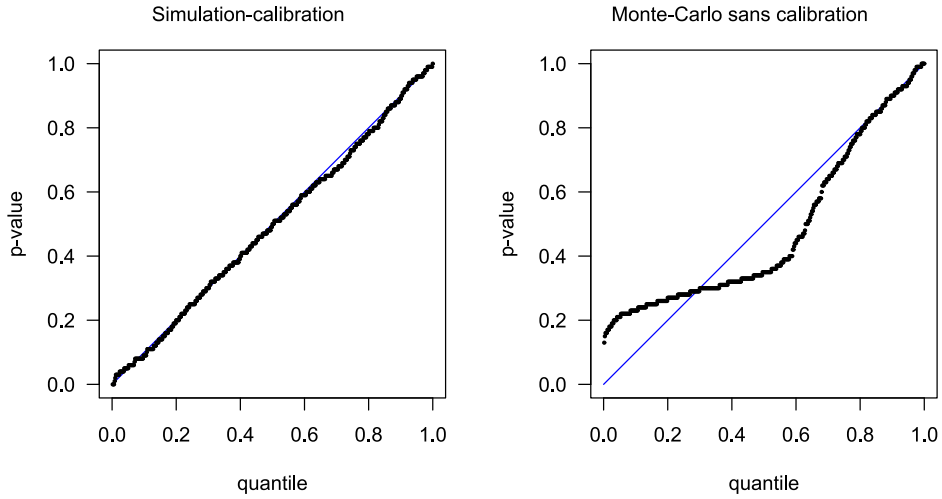

Figure 3.1 – Impact de la calibration dans un scénario : modèle linéaire, $\rho = 0.99$, 1 variable active, $\mathrm{SNR} = 0.1$.

à la loi uniforme. Au contraire, les p-values produites sans calibration s'en écartent fortement, avec une plus petite p-value parmi 500 égale à 0.13. Aux niveaux usuels de tests, les faux positifs sont donc pratiquement impossibles au lieu d'être possibles avec une probabilité contrôlée, ce qui suggère une très faible puissance du test. À des niveaux plus élevés du test, l'erreur de première espèce n'est au contraire plus contrôlée.

Pour évaluer systématiquement l'adéquation des $\hat{p}_s$ à la loi uniforme sur l'ensemble des scénarios, nous avons appliqué le test de Kolmogorov-Smirnov (test K-S) à la population des $(\hat{p}_s)_{1 \leq s \leq n_{\mathrm{sim}}}$ de chaque scénario. Nous avons réalisé le test dans sa version bilatérale, dont l'hypothèse nulle est que la population considérée est un échantillon de la loi cible (ici la loi uniforme sur $[0, 1]$), ainsi que dans l'une de ses versions unilatérales dont l'hypothèse nulle, moins stricte, est que la population est un échantillon d'une loi dominant stochastiquement (voir la définition au lemme 2) la loi cible. En effet, comme on l'a vu en conclusion de la section 3.7.2, fonder un test sur une p-value empirique dont la loi domine la loi uniforme sur $[0, 1]$ permet de contrôler son erreur de

première espèce.

Les résultats dépendent du type de modèle : linéaire, binaire équilibré, binaire déséquilibré ou modèle de Poisson. Pour chacun des quatre types et chacune des deux variantes du test de Kolmogorov-Smirnov, on applique la correction de Bonferroni à l'ensemble des p-values des 63 tests K-S appliqués aux scénarios de ce type de modèle.

Dans le modèle linéaire et le modèle binaire à données équilibrées, les tests K-S bilatéraux ne permettent pas de rejeter l'hypothèse que chaque population de p-values suit une loi uniforme sur $[0, 1]$. La plus petite p-value est de 0.0112 parmi les scénarios linéaires et de 0.0546 parmi les scénarios binaires équilibrés, ce qui, combiné à la correction de Bonferroni pour 63 tests, n'entraîne pas de rejet aux niveaux de test usuels ($63 \times 0.0112 = 0.706$). La figure 3.2 illustre cette adéquation généralisée à la loi uniforme sur huit exemples de scénarios.

Dans le modèle binaire à données déséquilibrées, il existe des scénarios dans lesquels la répartition des p-values générées s'éloigne significativement de la loi uniforme, et ne la domine pas. Au niveau 0.1 et avec la correction de Bonferroni, l'hypothèse nulle du test K-S (unilatéral comme bilatéral) est rejetée dans deux scénarios sur 63 : ceux où la corrélation entre covariables ($\rho = 0.99$) et le nombre de régresseurs actifs ($|A| = 10$) sont maximaux, et où le rapport signal-bruit est élevé ($\mathrm{SNR} = 0.3$ ou $\mathrm{SNR} = 1$). Ces deux scénarios ont des p-values du test K-S bilatéral respectivement égales à $p = 2.86 \times 10^{-6} = 1.81 \times 10^{-4}/63$ et $p = 2.55 \times 10^{-7} = 1.61 \times 10^{-5}/63$. Malgré cette non adéquation, on observe sur les diagrammes quantile-quantile (figure 3.3, qui inclut le scénario $\mathrm{SNR} = 1$ où l'écartement à la loi uniforme est le plus marqué) qu'on a bien $P_{H_0(A)}(\hat{p}_A \leq \alpha) \leq \alpha$ pour les petites valeurs α, le contraire ne se produisant que quand α dépasse 0.4 environ. On a donc en

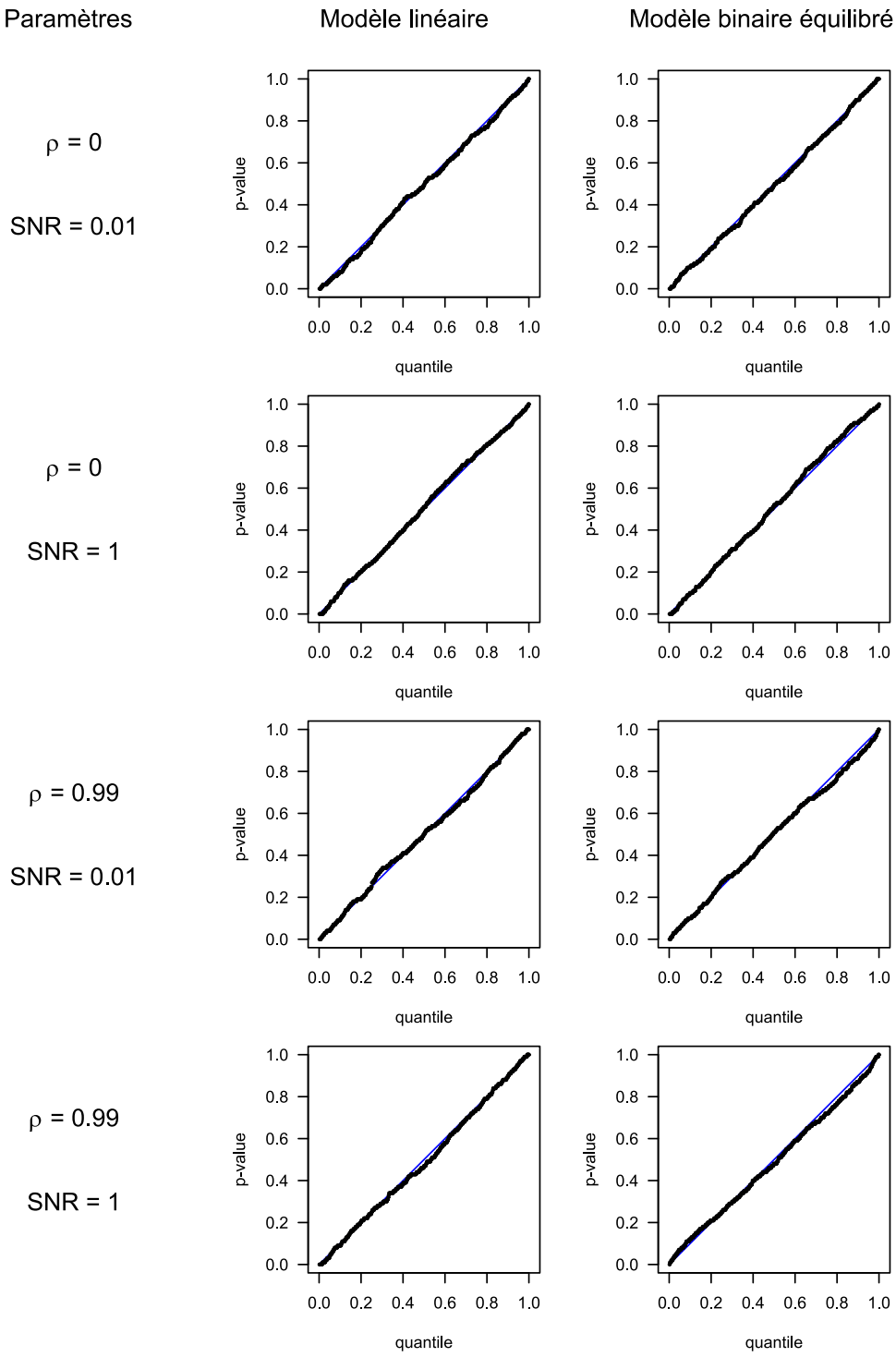

Figure 3.2 – Diagrammes quantile-quantile des 500 p-values empiriques générées par simulation-calibration dans 8 scénarios de modèle linéaire ou binaire équilibré ayant 10 variables actives.

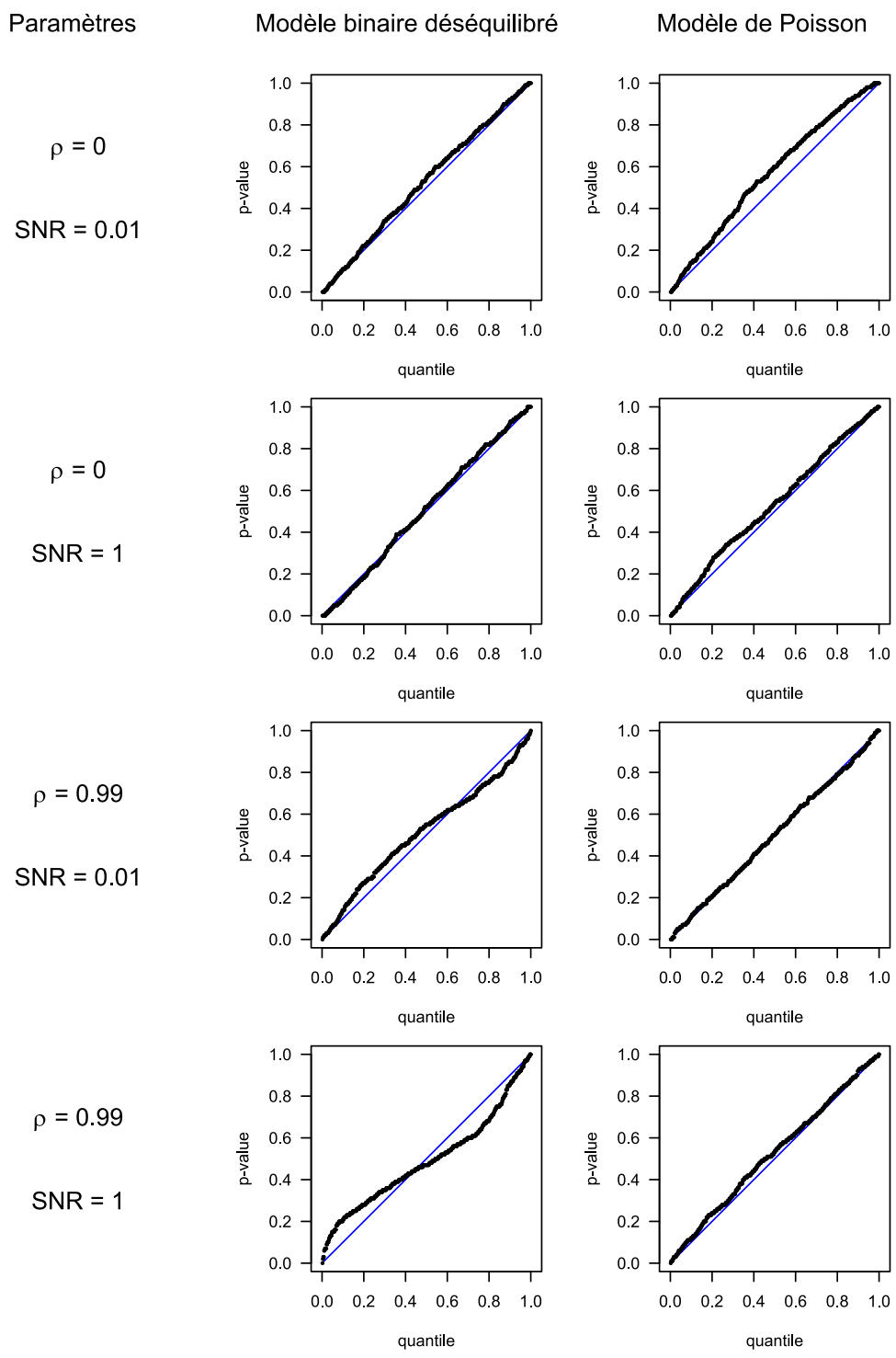

Figure 3.3 – Diagrammes quantile-quantile des 500 p-values empiriques générées par simulation-calibration dans 8 scénarios de modèle binaire déséquilibré ou de Poisson ayant 10 variables actives.

pratique un contrôle de l'erreur de première espèce plus conservatif que le niveau nominal aux niveaux de test usuels. Dans le scénario le plus divergent, le diagramme montre même une erreur de première espèce nettement plus faible que son niveau attendu ce qui signale une perte de puissance du test. Cet effet est cependant nettement plus faible que dans l'exemple de simulation non calibrée.

Dans le modèle de Poisson, les tests K-S unilatéraux ne permettent pas de rejeter l'hypothèse que chaque population de p-values domine la loi uniforme sur $[0, 1]$ (p-value minimale des test unilatéraux égale à $0.0214 = 1.35/63$). On observe donc un contrôle du FWER au moins aussi conservatif que le niveau nominal à tous les niveaux de test. En revanche, au niveau 0.1 et avec la correction de Bonferroni, les tests K-S bilatéraux permettent de rejeter l'hypothèse d'une distribution uniforme des p-values dans 5 scénarios sur 63. Ces scénarios présentent un profil différent de ceux où on observait un écartement à la loi uniforme dans le cas binaire déséquilibré : 4 sur 5 sont des scénarios à corrélation nulle, tous ont un SNR inférieur où égal à 0.1, et la p-value minimale ($p = 1.79 \times 10^{-6} = 1.13 \times 10^{-4}/63$) est atteinte à corrélation nulle, rapport signal-bruit minimal ($\mathrm{SNR} = 0.01$), et $|A| = 10$ régresseurs actifs connus. Par ailleurs, on observe sur la figure 3.3, où apparaît le scénario de p-value du test K-S minimale, que même dans celui-ci l'écartement à la loi uniforme est assez faible, en particulier aux petits quantiles. Le niveau réel de l'erreur de première espèce est donc très proche de son niveau nominal aux valeurs usuelles de celui-ci.

3.11.3 . Procédure de sélection de variables

Dans cette étude de simulation, contrairement à la section 3.11.2, nous n'avons pas supposé connu l'ensemble des régresseurs actifs. Cet ensemble est estimé par la procédure séquentielle décrite par l'algorithme 5. Le nombre

de simulations dans le calcul des p-values empirique est supérieur, $N = 500$, ce qui permet une estimation plus précise des p-values.

L'algorithme est d'abord appliqué avec le critère d'arrêt par seuillage à $\alpha = 0.95$, un niveau élevé qui permet d'obtenir sur chaque jeu de données une suite relativement complète de variables susceptibles d'être sélectionnées, ainsi que leurs p-values associées. Dans un second temps, la procédure de sélection de variables est appliquée à ces suites de p-values pour chaque α appartenant à un maillage relativement dense de valeurs (de 0.01 à 0.5, avec un pas de 0.01), avec le critère d'arrêt par seuillage ainsi que par ForwardStop. Cette seconde étape permet d'observer les performances de la procédure sur un grand nombre de valeurs de α de façon computationnellement peu intensive car elle ne demande pas de recalculer les p-values empiriques par simulation-calibration, cela ayant été fait lors de la première étape.

Contrôle du FWER et du FDR

Les figures 3.4 et 3.5 présentent l'évolution du FWER des procédures avec seuillage ou ForwardStop en fonction de α dans les mêmes 16 exemples de scénarios que les figures 3.2 et 3.3 : 10 variables actives, une corrélation entre covariables nulle ou maximale, un SNR minimal ou maximal, et la présence de chacun des quatre types de modèle. Les figures 3.6 et 3.7 présentent l'évolution du FDR en fonction de α dans ces 16 scénarios.

Bien que les résultats théoriques de contrôle du FWER ou du FDR ne soient valables que dans le modèle linéaire, l'allure des ces courbes varie peu en fonction du type de modèle. Dans les scénarios où il est le plus facile de capter un signal, c'est-à-dire ceux à SNR élevé et parmi ceux-ci surtout les scénarios non corrélés, on observe que le critère ForwardStop est nettement moins conservatif que le seuillage à α égal, avec des FWER et FDR plus élevés. Cela s'explique par un nombre substantiel de petites p-values parmi les premières

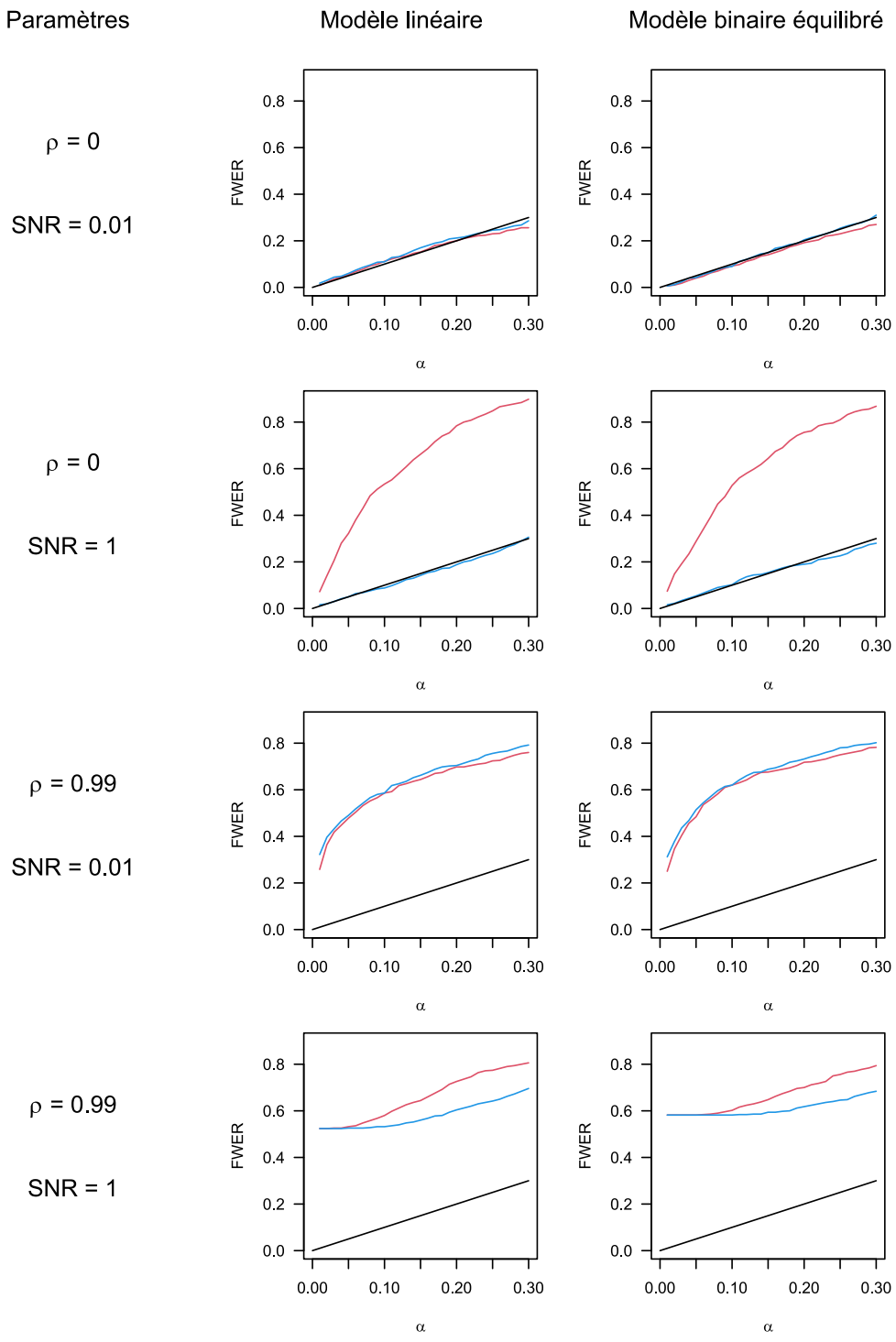

Figure 3.4 – FWER de la procédure de sélection de variables avec seuillage (bleu) ou *ForwardStop* (rouge) en fonction de α dans 8 scénarios de modèle linéaire ou binaire équilibré ayant 10 variables actives.

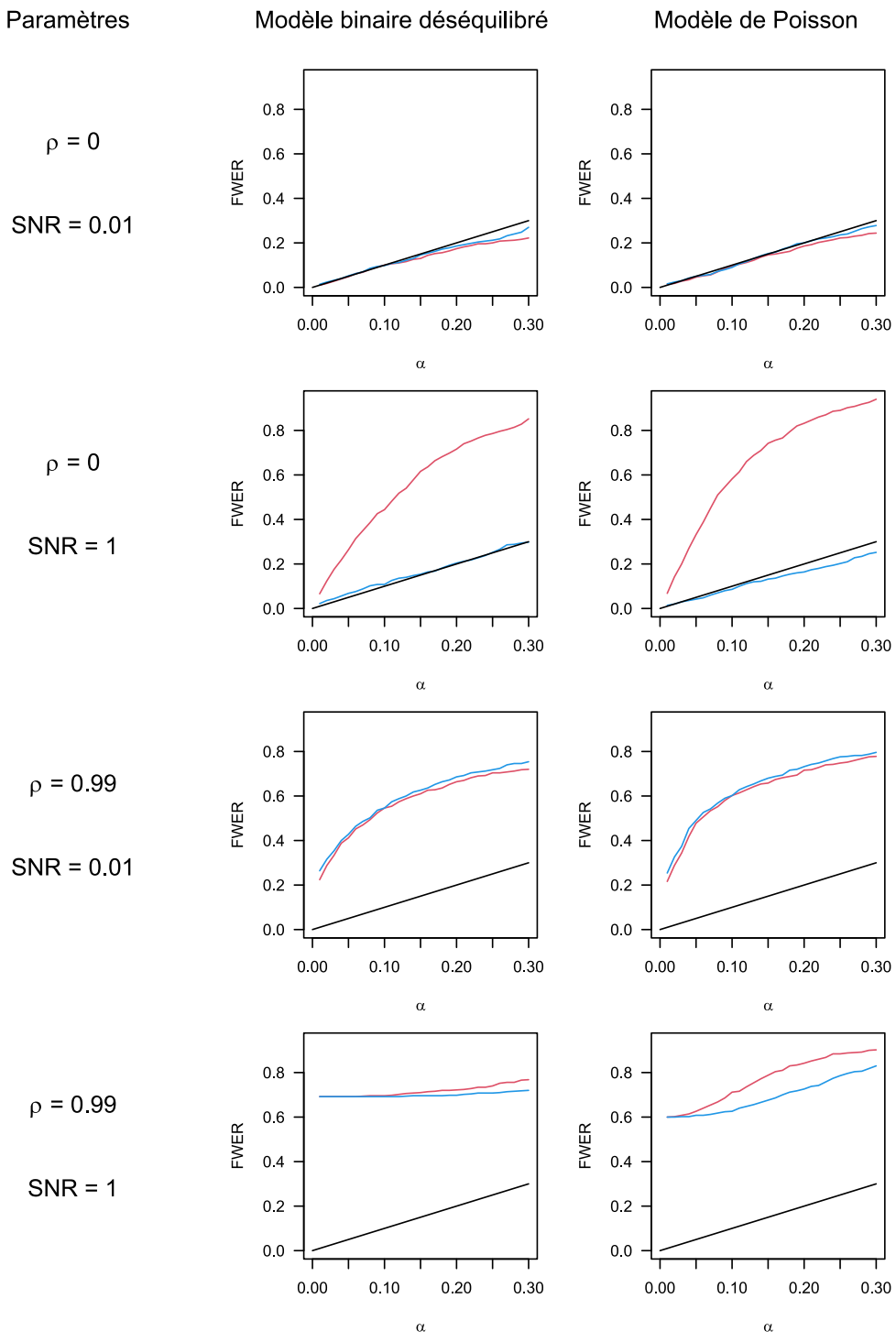

Figure 3.5 – FWER de la procédure de sélection de variables avec seuillage (bleu) ou *ForwardStop* (rouge) en fonction de α dans 8 scénarios de modèle binaire déséquilibré ou de Poisson ayant 10 variables actives.

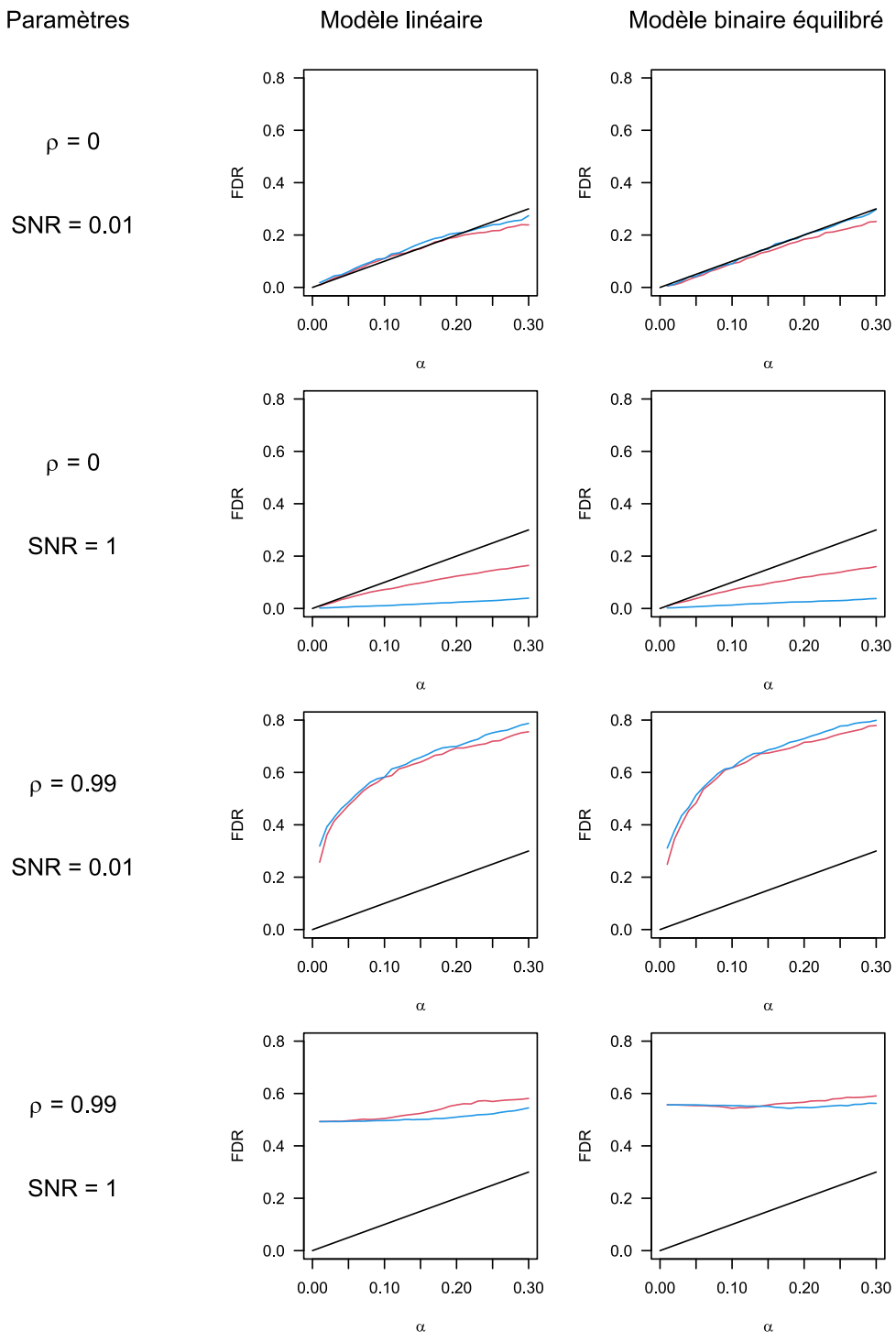

Figure 3.6 – FDR de la procédure de sélection de variables avec seuillage (bleu) ou *ForwardStop* (rouge) en fonction de α dans 8 scénarios de modèle linéaire ou binaire équilibré ayant 10 variables actives.

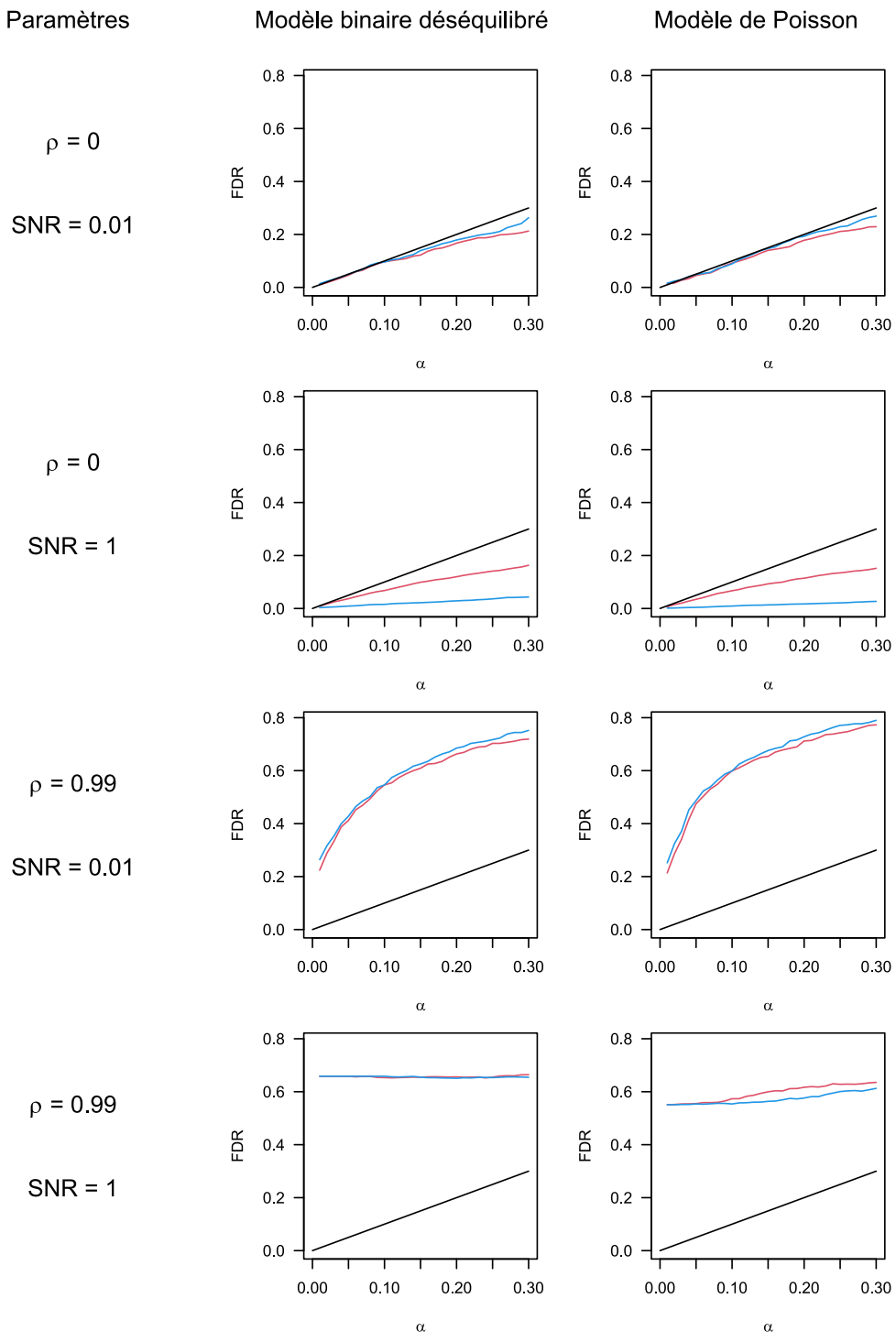

Figure 3.7 – FDR de la procédure de sélection de variables avec seuillage (bleu) ou *ForwardStop* (rouge) en fonction de α dans 8 scénarios de modèle binaire déséquilibré ou de Poisson ayant 10 variables actives.

valeurs de la suite (p_k), correspondant aux variables facilement détectées, qui, par moyennage, tirent vers le bas la quantité p_k^{FS} aux k plus élevés, la rendant nettement inférieure à p_k.

Dans tous les scénarios non corrélés, on observe comme attendu le contrôle du FWER à un niveau très proche de α avec le critère d'arrêt par seuillage (courbes bleues en figures 3.4 et 3.5 proches ou en-dessous des bissectrices). De plus, dans ces scénarios, on observe comme attendu le contrôle du FDR à un niveau proche ou inférieur à α avec le critère ForwardStop (courbes rouges en figures 3.6 et 3.7). En raison du caractère plus conservatif du seuillage par rapport au ForwardStop, le seuillage contrôle aussi le FDR à un niveau inférieur à α dans les scénarios non corrélés mais le ForwardStop ne contrôle pas le FWER. Le gain de sensibilité du ForwardStop par rapport au seuillage est modeste, atteignant au maximum 0.127 sur l'ensemble des 252 scénarios (voir en annexe les courbes de sensibilité sur les 16 scénarios exemples, figures B.1 et B.2).

En revanche, dans les scénarios à nombreuses variables actives et forte corrélation entre covariables, on n'observe de contrôle du FWER ni du FDR par aucun des deux critères d'arrêt. En effet, la forte corrélation entre variables actives et inactives permet souvent à une variable nominalement inactive d'être sélectionnée par le Lasso avant la variable active correspondante. L'association entre la variable sélectionnée et la réponse peut alors être statistiquement significative, car elle est une manifestation de l'association réelle entre la variable active et la réponse, et il n'est pas possible de l'attribuer à la variable nominalement active, qui n'est pas connue et dans ce cas pas détectée par le Lasso. Le statut de ces « faux positifs » porteurs d'un signal statistique a été discuté dans la littérature (G'Sell *et al.*, 2016).

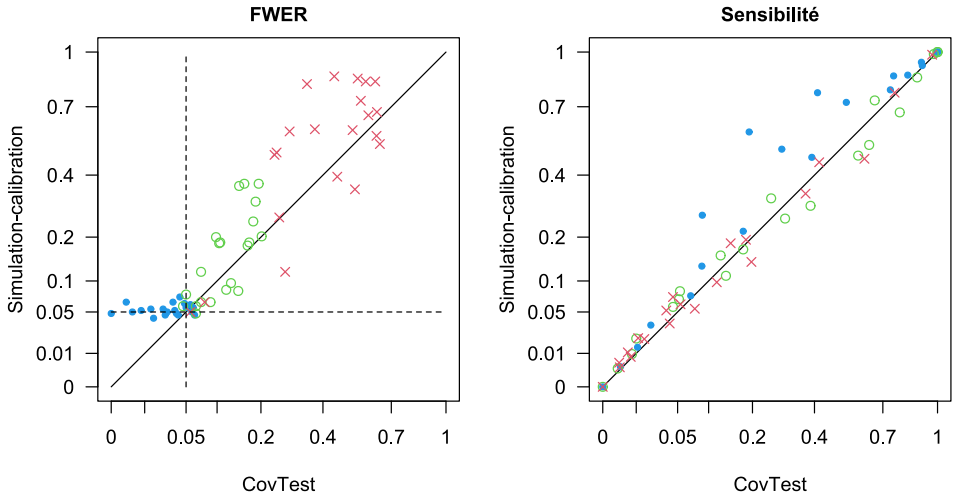

Figure 3.8 – Comparaison des performances des procédures de sélection avec seuillage du CovTest et du test par simulation-calibration au niveau $\alpha = 0.05$ sur les 63 modèles linéaires (• $\rho = 0$, ○ $\rho = 0.9$, × $\rho = 0.99$). La position d'un point indique la performance des deux méthodes sur un même scénario, les bissectrices signifiant une performance égale. L'échelle est quadratique.

Comparaison avec le test de covariance de Lockhart et al.

Pour pouvoir juger des résultats de sensibilité de la procédure de sélection de variables par simulation-calibration, il est utile de de la comparer à une autre méthode de sélection visant les mêmes objectifs. Ce comparateur est le test de covariance (CovTest) de Lockhart *et al.* (2014), qui comme le test par simulation-calibration mesure la significativité des variables apparaissant sur le chemin du Lasso, affectant à chacune d'entre elles une p-value.

Pour des raisons de faisabilité du CovTest nous nous sommes concentrés sur les 63 scénarios du modèle linéaire. Nous avons appliqué au CovTest l'équivalent de l'algorithme 5, avec le critère d'arrêt par seuillage au niveau $\alpha = 0.05$, ce qui est censé entraîner le contrôle du FWER à ce niveau. La figure 3.8 présente le FWER et la sensibilité des deux méthodes sur chacun des scénarios linéaires.

On observe l'échec du contrôle du FWER dans les scénarios à corrélation

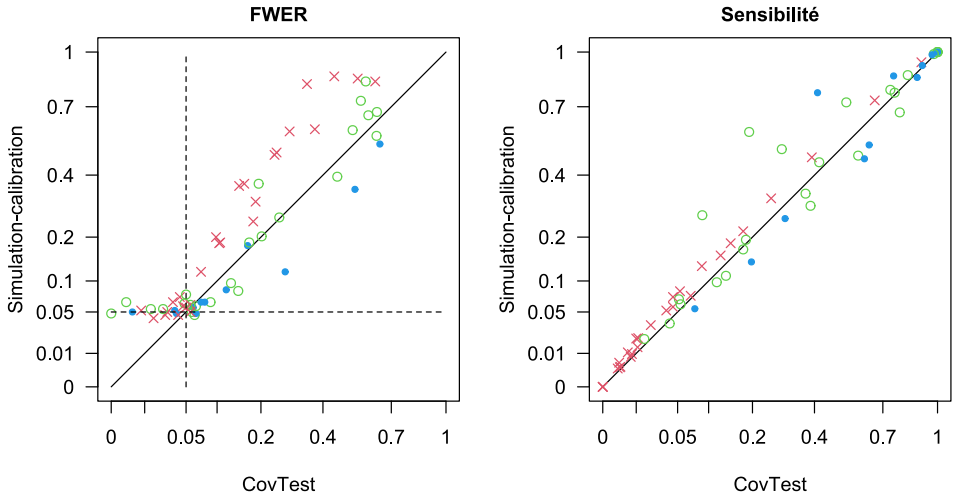

Figure 3.9 – Comparaison des performances des procédures de sélection avec seuillage du CovTest et du test par simulation-calibration au niveau $\alpha = 0.05$ sur les 63 modèles linéaires (• $\mathrm{SNR} = 1$, ○ $\mathrm{SNR} = 0.3$ ou 0.1, × $\mathrm{SNR} = 0.03$, 0.01 ou 0). La position d'un point indique la performance des deux méthodes sur un même scénario, les bissectrices signifiant une performance égale. L'échelle est quadratique.

entre covariables, que ce soit par le CovTest ou par la simulation-calibration (points ○ et × sur la figure 3.8). Dans ces scénarios, la simulation-calibration a tendance a produire des faux positifs plus fréquemment que le CovTest lorsque le rapport signal-bruit est faible, mais moins fréquemment lorsqu'il est élevé (voir le rôle du SNR sur la figure 3.9). En revanche, il y a bien contrôle du FWER à 5% ou moins sur les 21 scénarios linéaires sans corrélation à la fois par le CovTest et par la simulation-calibration (points •) mais avec un nette différence de régime entre les deux méthodes.

Le FWER observé avec la procédure de sélection par CovTest descend bien en-dessous de son niveau nominal dans certains scénarios, atteignant même 0 — c'est-à-dire aucun faux positif observé en 500 simulations — dans le scénario à 10 régresseurs, $\mathrm{SNR} = 0.3$.

Par contraste, les FWER observés avec la procédure par simulation-calibration sur les 21 scénarios non corrélés sont tous proches de leur niveau nominal.

Ils sont compris entre 0.042 et 0.072 ce qui signifie que le nombre de jeux de données simulées sur lesquels la procédure produit au moins un faux positif est réparti entre 21 et 36 parmi 500. Compte tenu de la correction de Bonferroni, cela est compatible avec l'hypothèse que le FWER réel est égal à α dans chacun de ces scénarios, c'est-à-dire que le nombre de jeux de données présentant au moins un faux positif qui suit dans chaque scénario une loi $\mathrm{Binomiale}(500, 0.05)$. En effet la p-value associée au plus grand des 21 FWER observés est de $0.0196 = 0.412/21$.

Ce plus fort conservatisme de la procédure par CovTest dans les scénarios non corrélés se traduit par des écarts notables de sensibilité à l'avantage de la procédure par simulation-calibration. La différence de sensibilité est positive ou nulle dans chacun de ces 21 scénarios, dépasse 0.1 dans 6 d'entre eux (qui ont tous au moins 5 régresseurs actifs, et un SNR d'au moins 0.1) et atteint un maximum de 0.388 dans le scénario à 10 régresseurs, $\mathrm{SNR} = 0.3$. Le test par simulation-calibration représente donc un progrès sur le CovTest dans la condition idéale qu'est l'absence de corrélation entre covariables, les gains substantiels de sensibilité étant permis par une hausse contrôlée du FWER qui ne dépasse pas significativement son niveau nominal.

3.12 .Application aux données de pharmacovigilance

Nous avons appliqué la procédure séquentielle de sélection de variables par simulation-calibration aux données de la base nationale de pharmacovigilance décrite en 2.5. Comme dans le cas des combinaisons du Lasso avec les critères d'information, nous nous sommes fondés sur la régression Lasso de l'évènement indésirable DILI sur les $p = 1692$ expositions médicamenteuses présentes dans la base, sur $n = 452\,914$ notifications. Ces données sont donc de plus grande taille que celles produites par l'étude de simulation. Pour ré-

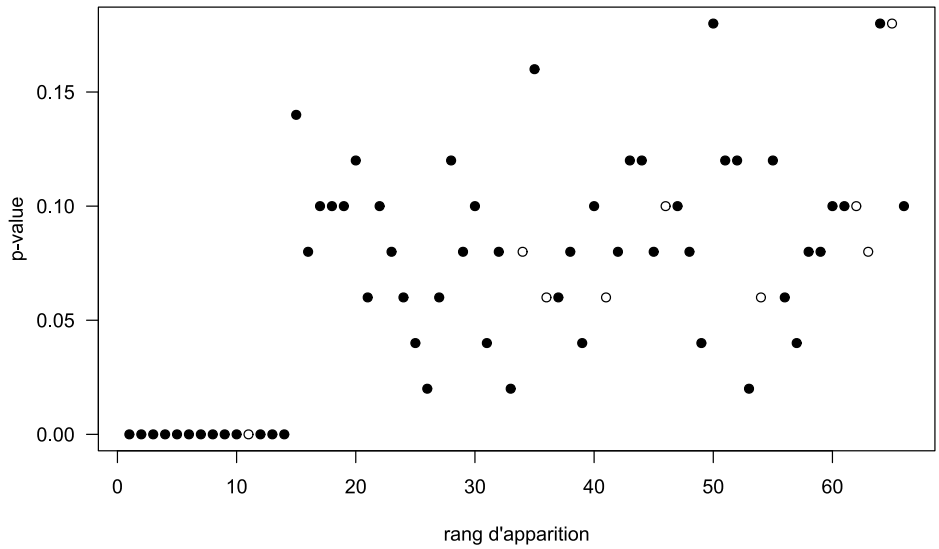

Figure 3.10 – p-values estimées par simulation-calibration sur le chemin du Lasso sans retirer les corrélations les plus fortes. ● : l'association trouvée est positive ($\widehat{\beta}_j^{Lasso} > 0$); ○ : négative.

duire le temps de calcul, nous nous sommes limités à $N = 50$ simulations dans le calcul des p-values empiriques. Même ainsi, le temps de calcul a été considérable : plusieurs jours en parallélisant les calculs sur 20 cœurs du serveur du CESP.

Au vu des premiers résultats, il est apparu qu'un traitement préalable des données avant de réaliser le Lasso était souhaitable. On observe en effet sur la figure 3.10 que les p-values estimées par simulation-calibration sont nulles pour les 14 premières variables du chemin du Lasso, puis réparties autour de 0.1. Ce changement de régime soudain coïncide avec une particularité de la structure de corrélation de la matrice des expositions X. La 14e exposition sélectionnée, le triméthoprime (de code ATC J01EA01) a la particularité d'être exceptionnellement corrélée à une autre exposition, le sulfaméthoxazole (J01EC01), un médicament avec lequel le triméthoprime est presque toujours coprescrit. Les deux variables ne diffèrent qu'en trois individus sur l'en-

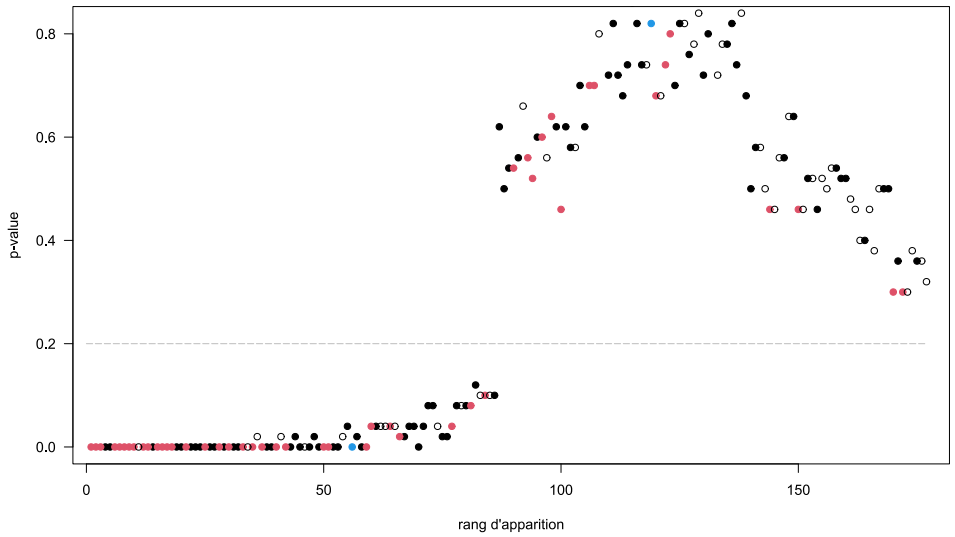

Figure 3.11 – p-values estimées par simulation-calibration sur le chemin du Lasso après retrait des corrélations les plus fortes, la couleur indiquant le statut des signaux d'après DILIrank. ● : association positive connue comme telle ; ● : association positive contredite par DILIrank ; ● : association positive non répertoriée ; ○ : association négative (pas de signal).

semble des notifications soit une corrélation de 0.9998.

Nous avons donc retiré de la matrice X les 6 covariables présentant une corrélation d'au moins 0.9 avec au moins une des variables apparaissant sur le chemin du Lasso (réalisé avant ce retrait), et appliqué le Lasso et la séquence de tests par simulation-calibration à ces données expurgées ($p = 1686$). Le résultat du Lasso est identique puisque les variables retirées n'étaient pas sélectionnées par celui-ci. En revanche les p-values estimées par simulation-calibration (figure 3.11) sont plus faibles à partir de la 15^{e} d'entre elles, et ne présentent plus de changement de régime à ce point. Elles présentent cependant un changement de régime comparable à un point plus avancé du chemin du Lasso, après la p-value associée à la 86^{e} variable sélectionnée (P01BDXX, diaminopyrimidines). Celle-ci est également corrélée à une autre variable présente dans la base (P01BD01, pyriméthamine), mais à un niveau plus faible :

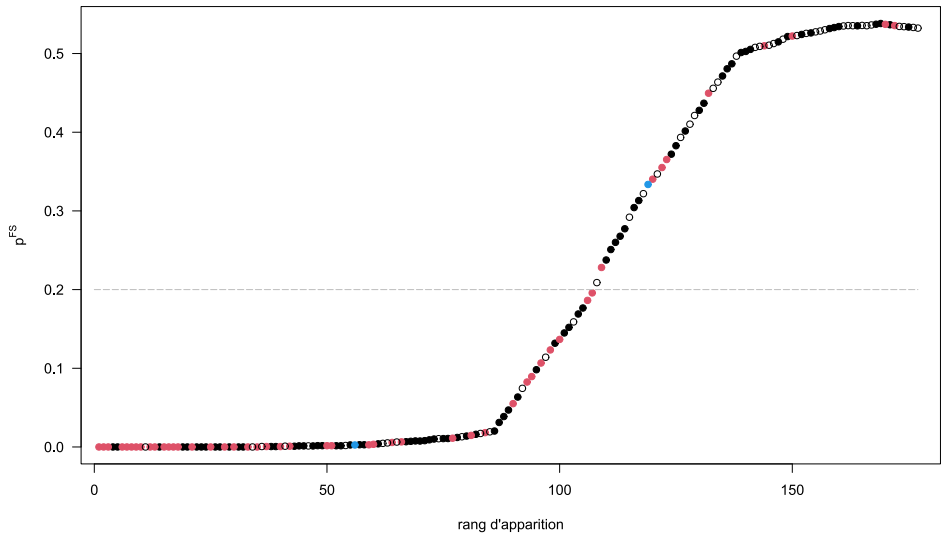

Figure 3.12 – quantités d'intérêt du critère *ForwardStop*, p^{FS} (voir la section 3.9.2), déduites des p-values estimées par simulation-calibration après retrait des corrélations les plus fortes. La couleur indique le statut des signaux comme sur la figure 3.11.

$\rho = 0.307$, du même ordre que les autres corrélations impliquant des vaiables sélectionnées.

On observe sur la figure 3.12 que l'utilisation du ForwardStop à la place du seuillage, en plus d'entraîner une sélection moins conservative à α égal, a un effet de stabilisation : tandis que les p-values empiriques présentent des fluctuations dues au nombre limité de simulations à partir desquelles on les estime, la quantité d'intérêt p^{FS} sur laquelle se fonde le ForwardStop présente un profile beaucoup plus lisse, étant calculée par moyenne à partir des p-values empiriques.

La table 3.1 présente les performances de sélection de variables de la procédure — avec le critère d'arrêt seuillage ou ForwardStop, au même niveau $\alpha = 0.2$ — en les calculant à partir du statut des expositions médicamenteuses recensé dans l'ensemble de référence DILIrank (Chen *et al.*, 2016). Elle est l'équivalent de la table 2.2 où figurent les performances des critères

Table 3.1 – Performance de la sélection de variable par simulation-calibration sur les données de la BNPV.

Critère d'arrêt	Signaux	Signaux à statut connu	Faux positifs	FDP (%)	Spécificité (%)	Sensibilité (%)
Seuillage à $\alpha = 0.2$	73	33	1	3.0	99.5	24.1
ForwardStop à $\alpha = 0.2$	91	41	1	2.4	99.5	30.0

d'information, dont l'EAIC. Comme pour ces derniers, seuls les expositions où le Lasso estime une association positive avec les DILI sont considérées comme des signaux. Le statut des expositions est également indiqué sur les figures 3.11 et 3.12 ce qui permet de visualiser les performances à d'autres niveaux de α. Les deux approches sont nettement plus conservatives que les critères d'information (y compris ForwardStop, lui-même moins conservatif que le seuillage). Cela leur permet d'obtenir un plus faible taux de faux positifs au prix d'une plus faible sensitivité. La daunorubicine (L01DB02) est l'unique faux positif d'après DILIrank.

3.13 . Discussion

Nous avons proposé un test de la significativité des variables apparaissant sur le chemin du Lasso, qui est utilisable de façon séquentielle pour sélectionner un modèle. Il teste l'hypothèse nulle $H_0(A)$ qui énonce qu'un ensemble connu de variables — A — comprend toutes les variables actives, et il s'intéresse à λ_A, la plus grande valeur de λ où apparaît une variable n'appartenant pas à A. Le rejet du test entraîne la sélection de cette variable.

Nous avons construit le test de manière à contourner la difficulté qu'il y a à utiliser directement λ_A comme statistique de test. De la même façon qu'une p-value est la probabilité sous l'hypothèse nulle que la statistique de test dé-

passe sa valeur observée, la statistique que nous considérons, p_A, est une probabilité conditionnelle que λ_A dépasse sa valeur observée : conditionnelle à la structure de corrélation reliant la réponse Y aux variables déjà sélectionnées X_A, donnée par $\widehat{\theta_A}$. C'est pourquoi nous l'appelons p-value conditionnelle. Nous l'estimons par la méthode de Monte-Carlo donnée par l'algorithme 1 : la simulation-calibration. Elle consiste à simuler des vecteurs réponse sous l'hypothèse nulle puis à les calibrer sur la condition portant sur $\widehat{\theta_A}$, ce qui permet de les simuler sous la conjonction de l'hypothèse nulle et de la condition recherchée, et donc d'estimer la p-value conditionnelle.

On peut y voir une généralisation de la sélection par permutations de Sabourin *et al.* (2015) (voir section 1.3.3) où se trouve l'idée de simuler une population de vecteurs réponses ayant la même distribution que le y observé, et de s'intéresser à la population des λ d'apparition d'une variable dans le Lasso appliqués à ces vecteurs simulés. Dans la sélection par permutations, on considère le λ d'apparition de la toute première variable, c'est-à-dire, avec nos notations, $\lambda_\emptyset$. La permutation est par ailleurs un moyen d'imposer la conservation de $\widehat{\theta_\emptyset} = \left(\widehat{\beta}_0, \hat{\sigma}\right)$, la moyenne et l'écart-type empiriques de y. La sélection par permutations ressemble par ces aspects au test de $H_0(\emptyset)$ par simulation-calibration. Puisque la sélection par permutations retient la médiane des $\lambda_\emptyset$, l'utiliser pour décider seulement de la sélection ou non de la première variable du chemin du Lasso reviendrait essentiellement à réaliser ce test au niveau $\alpha = 0.5$ (la seule différence étant dans la méthode de simulation calibrée, permutation ou calibration après-coup de vecteurs simulés quelconques).

Nous avons prouvé en sections 3.3 à 3.7 la validité de notre méthode dans le cas du modèle linéaire. La p-value conditionnelle $p_A(Y)$ suit sous l'hypothèse nulle une loi qui domine stochastiquement la loi uniforme sur $[0, 1]$ (lemme 3), et son estimateur $\widehat{p_A}(Y)$ est non biaisé, consistant en le nombre

N de simulations que l'on choisit de réaliser (lemme 7 et ses conséquences) et domine lui-même la loi uniforme sur l'ensemble discret des valeurs qu'il peut prendre (lemme 8). Cela signifie que le rejet de l'hypothèse nulle selon un seuil α sur $\widehat{p_A}(Y)$ entraîne le contrôle du taux d'erreur de première espèce à ce niveau α, à un petit terme résiduel près. D'un point de vue sélection de variables, cette erreur de première espèce signifie la sélection (nécessairement à tort) d'une variable supplémentaire alors que toutes les variables actives sont déjà sélectionnées, c'est-à-dire appartiennent à A. Nous avons de plus prouvé (section 3.10) que même lorsque certaines variables actives n'ont pas été sélectionnées, le risque de sélection d'une variable inactive par simulation-calibration est contrôlé à ce même niveau. Ce dernier résultat nécessite l'orthogonalité entre covariables actives et inactives. Le contrôle de l'erreur sous $H_0(A)$ ne dépend quand à lui d'aucune hypothèse sur la structure de corrélation.

Dans les modèles linéaires généralisés discrets (binaire et de Poisson), le cadre théorique n'est pas le même puisqu'il n'existe pas de p-value conditionnelle qui suive où domine la loi uniforme continue sur $[0, 1]$. Néanmoins, la production d'une p-value empirique par simulation-calibration s'adapte bien à ces modèles. En théorie, $\widehat{p_A}(Y)$ y est vue comme l'estimateur d'une probabilité dont la définition approche celle de la p-value conditionnelle (section 3.8.1). En pratique, la partie « calibration » de l'algorithme devient plus complexe : itérative et stochastique (algorithme 3) là où la calibration linéaire était une simple fonction affine. Bien que cela ne soit pas garanti par un théorème, l'étude de simulation (section 3.11.2) montre que la distribution des p-values empiriques n'est pas distinguable de la loi uniforme dans la majorité des scénarios de simulation non-linéaires et qu'elle s'en éloigne légèrement dans d'autres, sans qu'il y ait de scénario d'échec du contrôle de l'erreur de

première espèce aux niveaux usuels de α.

La simulation-calibration a l'inconvénient de demander un temps de calcul qui peut être important dans certaines circonstances. Elle implique, pour chacun des N vecteurs simulés, d'ajuster le modèle non pénalisé restreint à A afin de réaliser la calibration, puis d'appliquer le Lasso sur l'ensemble des covariables au vecteur calibré. Dans les modèles non linéaires, la calibration est elle-même itérative et demande de réajuster de façon répétée le modèle restreint non pénalisé. Dans l'application aux données de la BNPV, les conditions d'un temps de calcul très important étaient réunies : des données de grande taille ($452\,914$ par 1686), un modèle binaire, et surtout un ensemble A de variables sélectionnées qui atteint une taille importante (plus de 100). Il a donc été nécessaire d'ajuster de façon répétée des modèles logistiques non pénalisés de dimension relativement grande. Par contraste, l'étude de simulation a été réalisée avec au maximum 10 variables actives donc le nombre de variables sélectionnées ne dépassait pas cet ordre de grandeur, ce qui a limité le temps de calcul.

Notre test est comparable, par ses objectifs, au test de covariance (CovTest) de Lockhart *et al.* (2014) qui est également un test de significativité d'une variable préselectionnée par le Lasso. En nous concentrant sur la distribution du seul λ_A, et non comme Lockhart et al. sur l'évolution du Lasso entre deux λ d'apparition de variables consécutives, nous évitons le cas évoqué en section 1.3.4 où la sélection de deux variables à des valeurs de λ trop proches conduit artificiellement à n'en sélectionner aucune. Cela pourrait expliquer la meilleure sensibilité en l'absence de corrélation entre covariables que nous observons en section 3.11.3. En cas de corrélation, en revanche, les performances de la procédure de sélection par simulation-calibration ne sont pas toujours meilleures que celle de la sélection par CovTest, avec un FWER plus

faible à sensibilité égale dans les scénarios à fort SNR, mais plus élevé à sensibilité égale dans les scénarios à faible SNR (figures 3.8 et 3.9).

L'application aux données de la base nationale de pharmacovigilance a mis en évidence un phénomène indésirable qui pourrait contribuer à expliquer les moins bonnes performances dans certains cas de corrélation. On observe sur les figures 3.10 et 3.11 un changement de régime dans la suite des p-values estimées où, dès lors qu'une certaine variable j_0 appartient à A, $\widehat{p_A}(y)$ est toujours relativement élevé. Le détail des Lasso réalisés dans les simulations-calibrations fournit une explication de ce phénomène. Nous observons que lorsqu'il se produit, il existe une variable $j_1 \notin A$ corrélée à j_0 et éventuellement à d'autres variables dans A telle que, parmi la population des $\lambda_A(y^{(l)})$ obtenus par simulation-calibration (voir l'algorithme 4), une proportion importante sont des valeurs élevées associées à une sélection précoce de j_1 par le Lasso appliqué à $y^{(l)}$. En effet, la calibration des $y^{(l)}$ sur une ou des associations à des variables corrélées à j_1 peut entraîner une association avec j_1 elle-même, qui est capturée par le Lasso. Ces $\lambda_A(y^{(l)})$ élevés entraînent l'estimation d'un $\widehat{p_A}(y)$ élevé.

Ce phénomène témoigne de l'influence de j_A sur la loi de λ_A sous l'hypothèse nulle. Il est possible que le plus pertinent du point de vue de la puissance du test soit de comparer le $\lambda(y)$ observé, non pas comme nous l'avons fait à la loi de $\lambda_A(Y)$ conditionnellement à $\widehat{\theta_A}(Y)$, mais à sa loi sous un double conditionnement par $\widehat{\theta_A}(Y), j_A(Y)$, ou plus généralement tenant compte de $j_A(Y)$. En pratique, cela passerait par l'introduction d'une pondération dans le calcul par moyenne de $\widehat{p_A}(Y)$.

Ce phénomène produit, lorsque $j_0 \in A$, une perte de puissance dans la sélection de toutes les variables, y compris celles qui ne sont pas corrélées à j_0 ou à j_1. Cette perte de puissance dans certains cas de corrélation pourrait

compenser la tendance du test par simulation-calibration, observée sur les simulations sans corrélation, à être plus puissant mais moins conservatif que le CovTest ce qui expliquerait que dans certains scénarios corrélés, la sélection de variables par simulation-calibration soit moins conservative que celle par CovTest alors qu'elle a pratiquement la même sensibilité.

Le FWER de la procédure par simulation-calibration plus faible que celle par CovTest dans les scénarios corrélés à fort rapport signal-bruit pourrait s'expliquer par le fait qu'en raison du fort SNR, il est plus fréquent que toutes les variables actives soient sélectionnées avant les variables inactives sur le chemin du Lasso. Lorsque c'est le cas, $H_0(A)$ est vérifiée (donc le risque de faux positif est contrôlé) dans tous les tests de la procédure itérative qui sont suceptibles de produire un faux positif. Pour que le contrôle du risque échoue, il faut à la fois que les variables actives soient mélangées aux variables inactives sur le chemin du Lasso — pour qu'il existe une itération où l'on teste une $H_0(A)$ non vérifiée qui peut néanmoins entraîner la sélection d'une variable inactive —, et que les variables actives soient corrélées aux variables inactives — pour que le théorème étendu de contrôle de l'erreur de sélection (section 3.10) ne s'applique pas.

Il est par ailleurs possible que le contrôle de l'erreur de sélection par la simulation-calibration soit valide dans un cadre un peu plus large que celui des résultats théoriques que nous avons obtenus. Il nous semble en effet possible d'adapter la démonstration du théorème étendu pour relâcher l'hypothèse d'orthogonalité entre variables actives et inactives en la remplaçant par une orthogonalité entre les résidus des variables actives inconnues sur les variables actives connues, et les résidus des variables inactives sur les variables actives connues :

Conjecture 1 (Contrôle plus étendu de l'erreur de séléction)**.** *Soit* $\{1, .., p\} =$

$A \cup B \cup C$ avec $\forall j \in C, \beta_j = 0$. Supposons que les résidus des variables de B sur celles de A sont orthogonaux à ceux des variables de C sur celles de A, c'est-à-dire qu'il existe Π_{AB}, $\widetilde{X}_B$, Π_{AC}, $\widetilde{X}_C$ vérifiant :

$$\begin{aligned} X_B &= X_A \Pi_{AB} + \widetilde{X}_B, \;\; \widetilde{X}_B^T X_A = 0 \\ X_C &= X_A \Pi_{AC} + \widetilde{X}_C, \;\; \widetilde{X}_C^T X_A = 0 \\ &\widetilde{X}_B^T \widetilde{X}_C = 0. \end{aligned}$$

Alors, pour tout $\alpha \in [0, \frac{1}{2}]$, le test par simulation-calibration de $H_0(A)$ au niveau α a une probabilité inférieure à $\alpha + \frac{1-\alpha}{N+1}$ de sélectionner un faux positif.

L'idée de l'orthogonalité entre résidus sur A est que la corrélation entre deux variables d'indices $j_B \in B$ (active inconnue) et $j_C \in C$ (inactive), qui en temps normal empêche de contrôler la probabilité de sélectionner le faux positif j_C à la place de j_B, ne joue plus ce rôle si elle repose entièrement sur leur corrélations commune aux variables dans A. La perturbation (par rapport à l'hypothèse nulle) causée par la variable active inconnue j_B serait entièrement absorbée par la calibration sur A.

4 - Conclusion et perspectives

Nous avons proposé deux procédures de sélection de variables dans les modèles de régression linéaires ou linéaires généralisés de grande dimension. Elles visent toutes deux à contrôler le *Family-Wise Error Rate* (FWER) en s'adaptant à ce contexte de grande dimension. Par conséquent, ce sont des méthodes conservatives comparativement à d'autres méthodes qui n'ont pas pour objet d'éviter la sélection de faux positifs — comme la validation croisée, qui optimise les performances de prédiction — ou qui ne prennent pas en compte les effets de la grande dimension, comme le BIC.

Les deux procédures sont utilisées en post-traitement de la régression Lasso, pour choisir l'un des modèles que propose celle-ci sur son chemin de régularisation. Néanmoins l'AIC étendu est défini dans le contexte plus général de sélection d'un sous-modèle parcimonieux d'un modèle de grande taille. Nous l'associons au Lasso pour disposer d'une présélection de taille raisonnable de modèles candidats, mais cette présélection peut être faite par une autre méthode, appartenant elle aussi à la famille des régression pénalisées ou non. Le test par simulation-calibration est quant à lui, dans sa forme actuelle, spécifique au Lasso.

Le test par simulation-calibration a l'avantage de garantir mathématiquement le FWER dans deux situations : lorsque toutes les variables actives sont sélectionnées avant les variables inactives (son hypothèse nulle est alors vérifiée) et lorsqu'il n'y a pas de corrélation entre variables actives et inactives. L'EAIC ne présente quant à lui pas de garantie mathématique. Néanmoins, sa conception se fonde sur un rapprochement entre la sélection qu'il opère parmi des sous-modèles parcimonieux, et la réalisation de tests de rapport

de vraisemblance multiples. Les propriétés asymptotiques de ces tests multiples se traduisent en pratique par une bonne capacité de l'EAIC à contrôler le FWER. Compte tenu de sa simplicité et du faible temps de calcul qu'il demande, nous recommandons son utilisation plutôt que celle du BIC dès lors qu'un faible risque, ou même une faible proportion, de faux positifs est souhaitée dans une sélection de variables en grande dimension. Par ailleurs la notion de premier minimum local d'un critère d'information semble peut mériter plus de considération, par exemple en vue de la généraliser à d'autres contextes qu'une suite de modèles telle que le chemin du Lasso.

Nous avons mesuré les performances des deux procédures par des études de simulations très extensives. Il n'est cependant pas possible d'explorer exhaustivement toutes les combinaisons de paramètres qui caractérisent un problème de sélection de variables en grande dimension, et les deux études de simulations obéissent à des plans différents. Dans l'étude des performances de l'EAIC comparé aux autres IC, nous avons choisi en particulier de faire varier n et p, le nombre d'observations et la dimension de la régression, car l'EAIC est défini en fonction de p et nous le comparons à des critères définis en fonction de n : le BIC et les EBICs.

Les deux méthodes ont en revanche été appliquées à la même base de données réelles, les cas de pathologie hépatique d'origine médicamenteuse de la base nationale de pharmacovigilance. En plus de son rôle illustratif, cette application a permis de mettre en évidence des limites du test par simulation (temps de calcul, faible puissance dans certaines circonstances) qui n'étaient pas manifestes dans les simulations. Cette perte occasionnelle de puissance pourrait être corrigée par des modifications à apporter au test par simulation-calibration.

Le test par simulation-calibration, bien qu'il ne soit dans sa forme actuelle

applicable qu'à la sélection de variables fondée sur le Lasso, est porteur d'idées qui possèdent un potentiel plus large. Nous avons repris l'idée existante de la p-value empirique, qui permet d'associer une p-value à potentiellement toute statistique de test en estimant la p-value par Monte-Carlo via des simulations de données suivant l'hypothèse nulle. Il s'y est ajouté le problème de la définition de la p-value associée à une hypothèse nulle qui n'est à elle seule pas assez spécifiée pour définir la loi de la statistique de test souhaitée, ou simuler selon celle-ci. Nous l'avons résolu par l'utilisation de la p-value conditionnelle, nous ramenant à une loi conditionnelle qui, combinée à l'hypothèse nulle, est suffisamment spécifiée. On peut donner la définition générale suivante de la notion de p-value conditionnelle associée à des données observées x, réalisation de la variable aléatoire X, une hypothèse nulle H_0, une statistique de test T, et une statistique auxiliaire U :

- p-value : $p_t(x) = \mathrm{P}_{H_0}\left(T(X) \geq T(x)\right)$
- p-value conditionnelle : $p_{t|u}(x) = \mathrm{P}_{H_0}\left(T(X) \geq T(x) \mid u(X) = U(x)\right)$

Dans le test appliqué au Lasso, $x = y$, $T(x) = \lambda_A(y)$, et $U(x) = \widehat{\theta_A}(y)$. De même que l'utilisation de p-values empiriques permet une plus grande flexibilité dans le choix des statistiques de test dont elle dispense de connaître mathématiquement la loi sous l'hypothèse nulle à condition de savoir simuler selon celle-ci, l'utilisation de p-values conditionnelles doit permettre une plus grande flexibilité dans le choix des hypothèses nulles et de la loi à laquelle on compare la statistique observée.

Bibliographie

Christophe Giraud : *Introduction to High-Dimensional Statistics*. CRC Press, 2021. ISBN 9780367716226.

Robert Tibshirani : Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246. Publisher : [Royal Statistical Society, Wiley].

Trevor Hastie, Robert Tibshirani et Martin Wainwright : *Statistical Learning with Sparsity*. CRC Press, 2021. ISBN 1498712169.

Trevor Hastie, Junyang Qian et Kenneth Tay : An Introduction to 'glmnet', 2023. URL https://glmnet.stanford.edu/articles/glmnet.html.

Peter Bühlmann et Sara van de Geer : *Statistics for High-Dimensional Data*. Springer, 2011. ISBN 978-3-642-20191-2.

Trevor Hastie, Robert Tibshirani et Jerome Friedman : *The Elements of Statistical Learning*. Springer, 2017. ISBN 978–387-84857-0.

Jiahua Chen et Zehua Chen : Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, septembre 2008. ISSN 0006-3444.

Sadanori Konishi et Genshiro Kitagawa : Generalised Information Criteria in Model Selection. *Biometrika*, 83(4):875–890, 1996. ISSN 0006-3444. URL https://www.jstor.org/stable/2337290. Publisher : [Oxford University Press, Biometrika Trust].

Hirotugu Akaike : Information theory and an extention of the maximum likelihood principle. *2nd International Symposium on Information Theory, 1973*, pages 267–281, 1973. Publisher : Akademiai Kiado.

Gideon Schwarz : Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, mars 1978. ISSN 0090-5364, 2168-8966. Publisher : Institute of Mathematical Statistics.

Jeremy A. Sabourin, William Valdar et Andrew B. Nobel : A permutation approach for selecting the penalty parameter in penalized model selection. *Biometrics*, 71(4):1185–1194, 2015. ISSN 1541-0420.

Emeline Courtois, Pascale Tubert-Bitter et Ismaïl Ahmed : New adaptive lasso approaches for variable selection in automated pharmacovigilance signal detection. *BMC Medical Research Methodology*, 21(1):271, décembre 2021. ISSN 1471-2288.

Vivian Viallon : *Approches pénalisées et autres développements statistiques pour l'épidémiologie*. thesis, Université Claude Bernard Lyon 1, mai 2016. URL https://hal.science/tel-01366359.

Nicolai Meinshausen et Peter Bühlmann : Stability selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 72 (4):417–473, 2010. ISSN 1467-9868. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00740.x. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2010.00740.x.

Francis Bach : Bolasso : Model Consistent Lasso Estimation through the Bootstrap. *Proceedings of the 25th international conference on Machine learning*, 33-40, mai 2008.

Ismaïl Ahmed, Antoine Pariente et Pascale Tubert-Bitter : Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Statistical Methods in Medical Research*, 27(3):785–797, mars 2018. ISSN 1477-0334.

Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani et Robert Tibshirani : A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, avril 2014. ISSN 0090-5364.

Minjun Chen, Ayako Suzuki, Shraddha Thakkar, Ke Yu, Chuchu Hu et Weida Tong : DILIrank : the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discovery Today*, 21(4):648–653, avril 2016. ISSN 1878-5832.

S. S. Wilks : The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, mars 1938. ISSN 0003-4851, 2168-8990. Publisher : Institute of Mathematical Statistics.

Karline Soetaert et Peter M.J. Herman : *A Practical Guide to Ecological Modelling. Using R as a Simulation Platform*. Springer, 2009. ISBN 978-1-4020-8623-6.

Jerome Friedman, Trevor Hastie et Robert Tibshirani : Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Jiahua Chen et Zehua Chen : Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 22(2):555–574, 2012. ISSN 1017-0405. Publisher : Institute of Statistical Science, Academia Sinica.

Daunorubicin. *In LiverTox : Clinical and Research Information on Drug-Induced Liver Injury*. National Institute of Diabetes and Digestive and Kidney Di-

seases, Bethesda (MD), 2012. URL http://www.ncbi.nlm.nih.gov/books/NBK548259/.

Jianqing Fan et Runze Li : Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, décembre 2001. ISSN 0162-1459. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1198/016214501753382273.

Hui Zou et Trevor Hastie : Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 1467-9868.

Nicholas W. Galwey : A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology*, 33(7):559–568, novembre 2009. ISSN 1098-2272.

Miao-Xin Li, Juilian M. Y. Yeung, Stacey S. Cherny et Pak C. Sham : Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics*, 131(5):747, 2012. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3325408/. Publisher : Springer.

Yoav Benjamini et Yosef Hochberg : Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 0035-9246. URL http://www.jstor.org/stable/2346101. Publisher : [Royal Statistical Society, Wiley].

Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova et Robert Tibshirani : Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(2):423–

444, 2016. ISSN 1369-7412. URL https://www.jstor.org/stable/24775345. Publisher : [Royal Statistical Society, Wiley].

Ruth Marcus, Eric Peritz et K. R. Gabriel : On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. *Biometrika*, 63(3):655–660, 1976. ISSN 0006-3444. URL https://www.jstor.org/stable/2335748. Publisher : [Oxford University Press, Biometrika Trust].

Jutta Roosen et David A. Hennessy : Testing for the Monotone Likelihood Ratio Assumption. *Journal of Business & Economic Statistics*, 22(3):358–366, 2004. ISSN 0735-0015. URL https://www.jstor.org/stable/1392602. Publisher : [American Statistical Association, Taylor & Francis, Ltd.].

A - Résultats de simulations complémentaires : EAIC et autres critères d'information

A.1 .Sensibilité

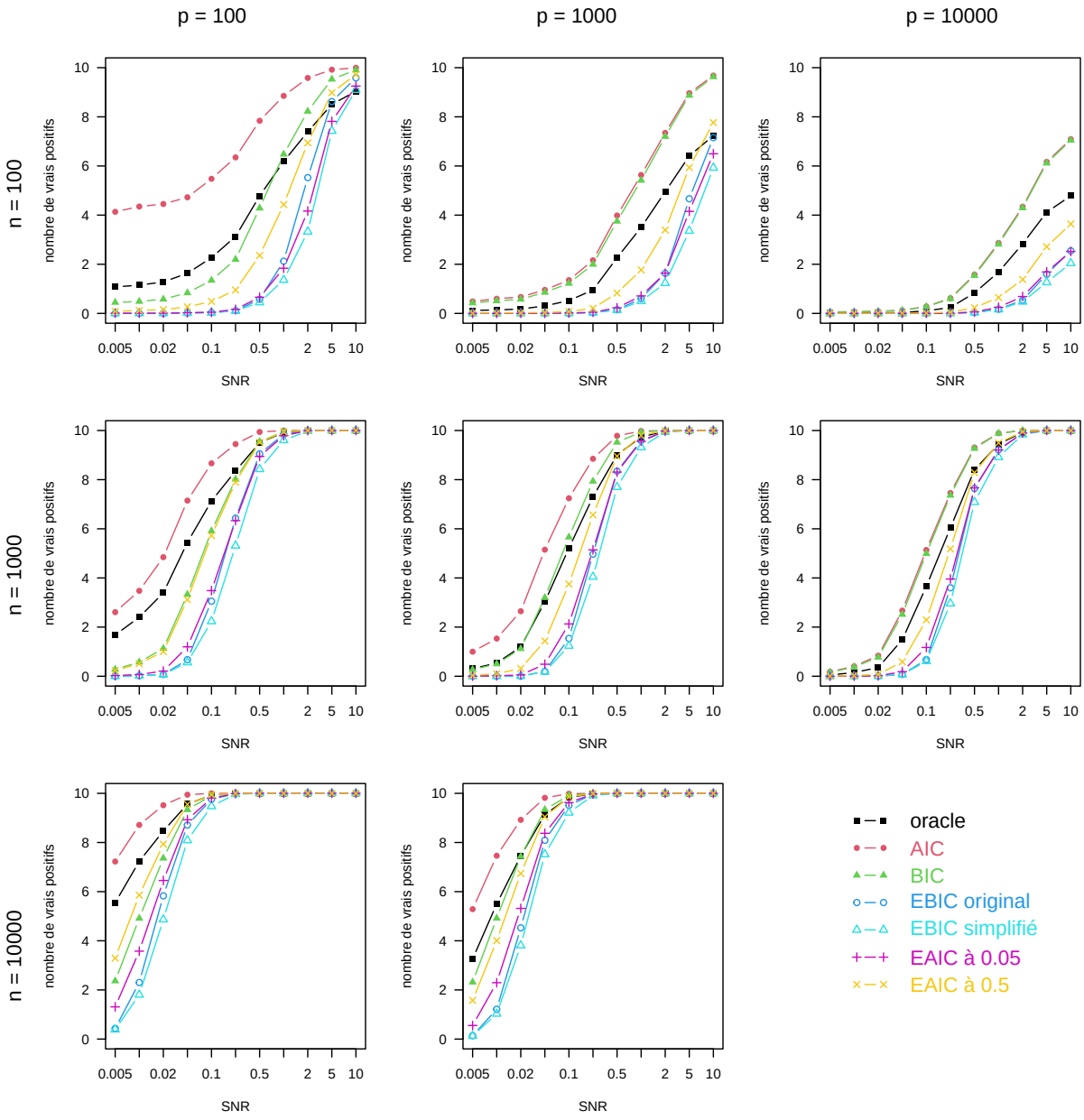

Figure A.1 – Étude de simulation de la procédure complète : nombre de vrais positifs par paramètre, moyenné sur 1000 simulations. Modèle linéaire, $\rho = 0$.

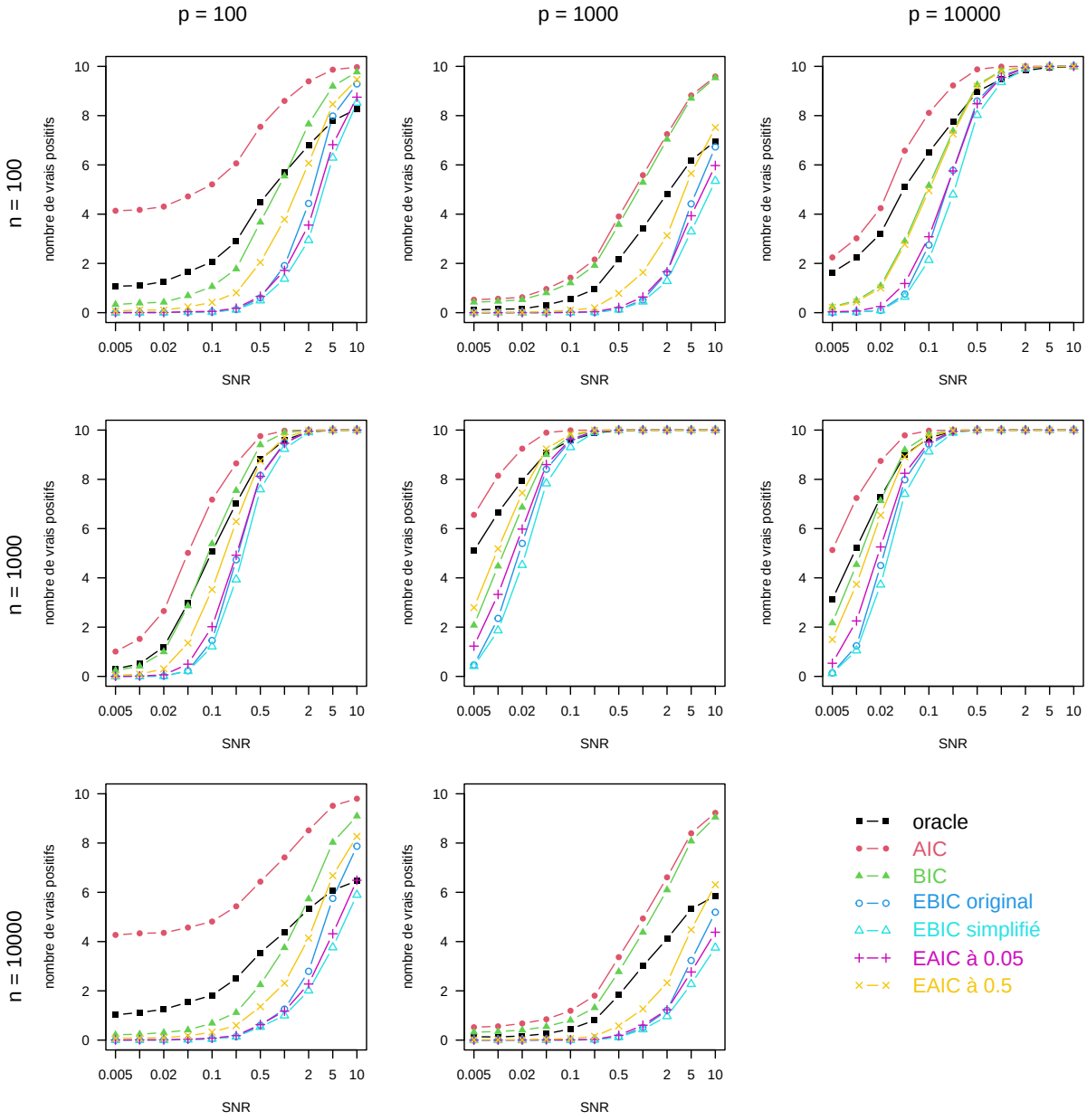

Figure A.2 – Étude de simulation de la procédure complète de sélection de modèle par IC : nombre de vrais positifs par paramètre, moyenné sur 1000 simulations. Modèle logistique, $\rho = 0$.

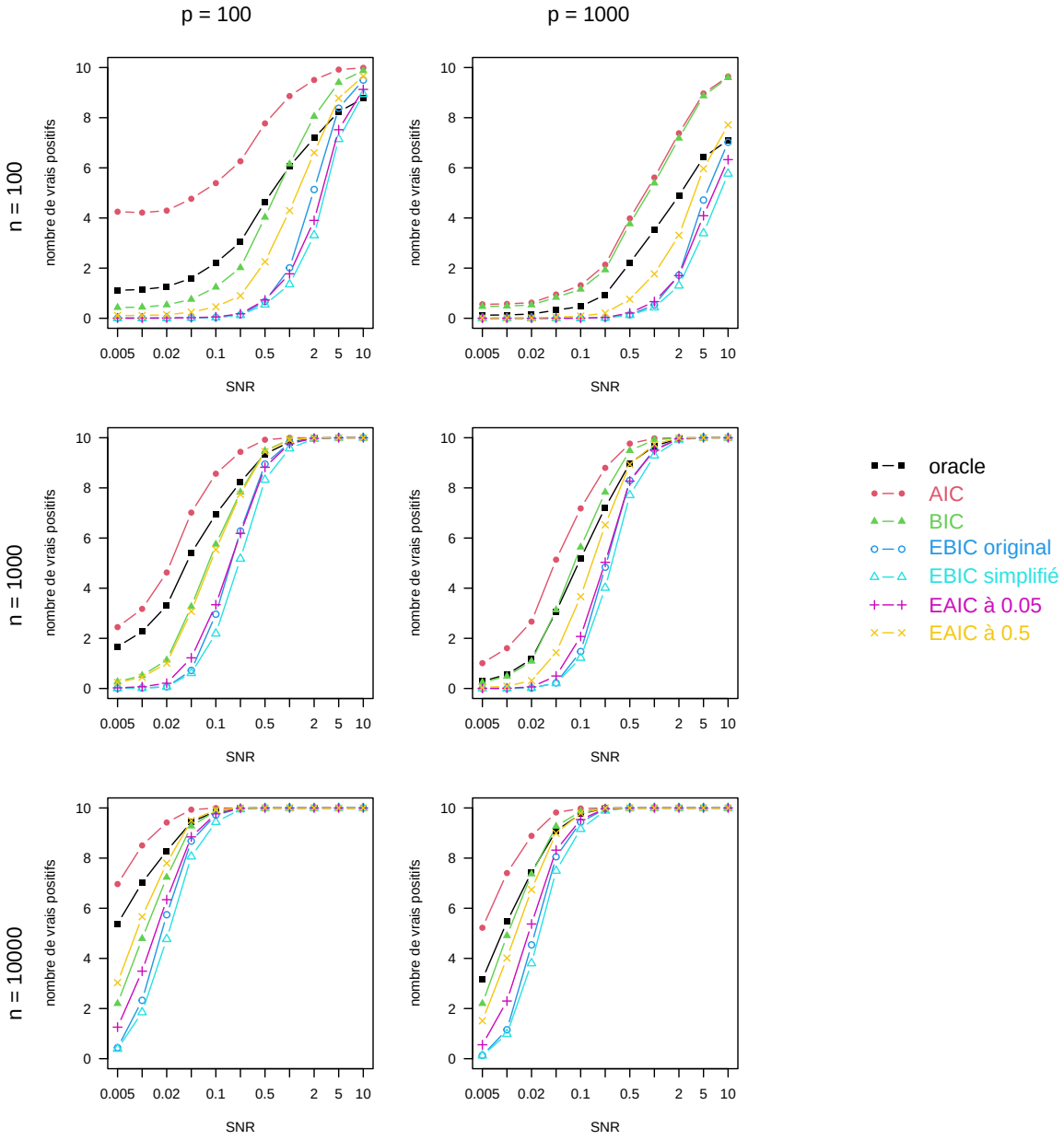

Figure A.3 – Étude de simulation de la procédure complète de sélection de modèle par IC : nombre de vrais positifs par paramètre, moyenné sur 1000 simulations. Modèle linéaire, $\rho = 0.5$.

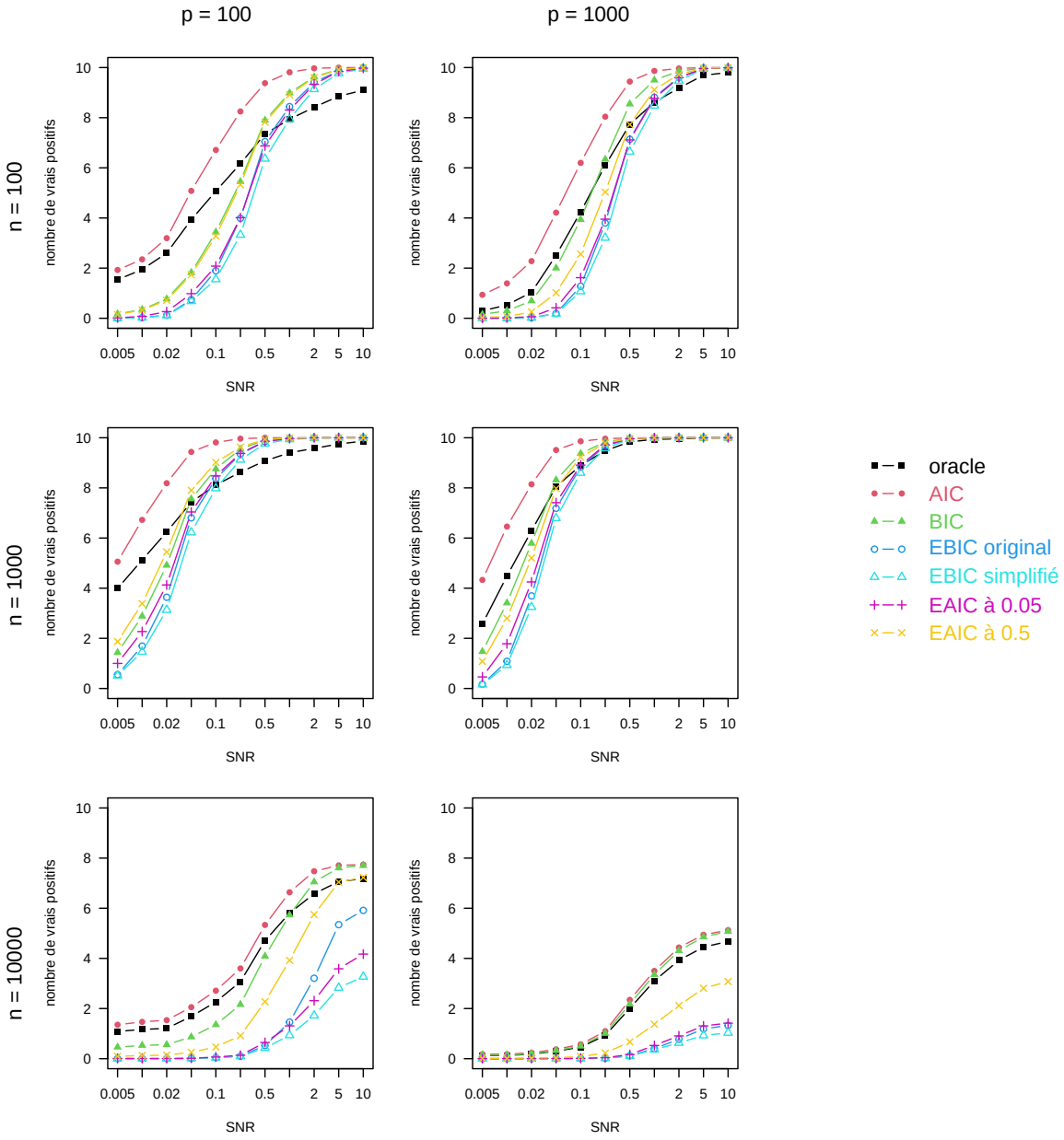

Figure A.4 – Étude de simulation de la procédure complète de sélection de modèle par IC : nombre de vrais positifs par paramètre, moyenné sur 1000 simulations. Modèle logistique, $\rho = 0.5$.

A.2 .Taux de fausses découvertes

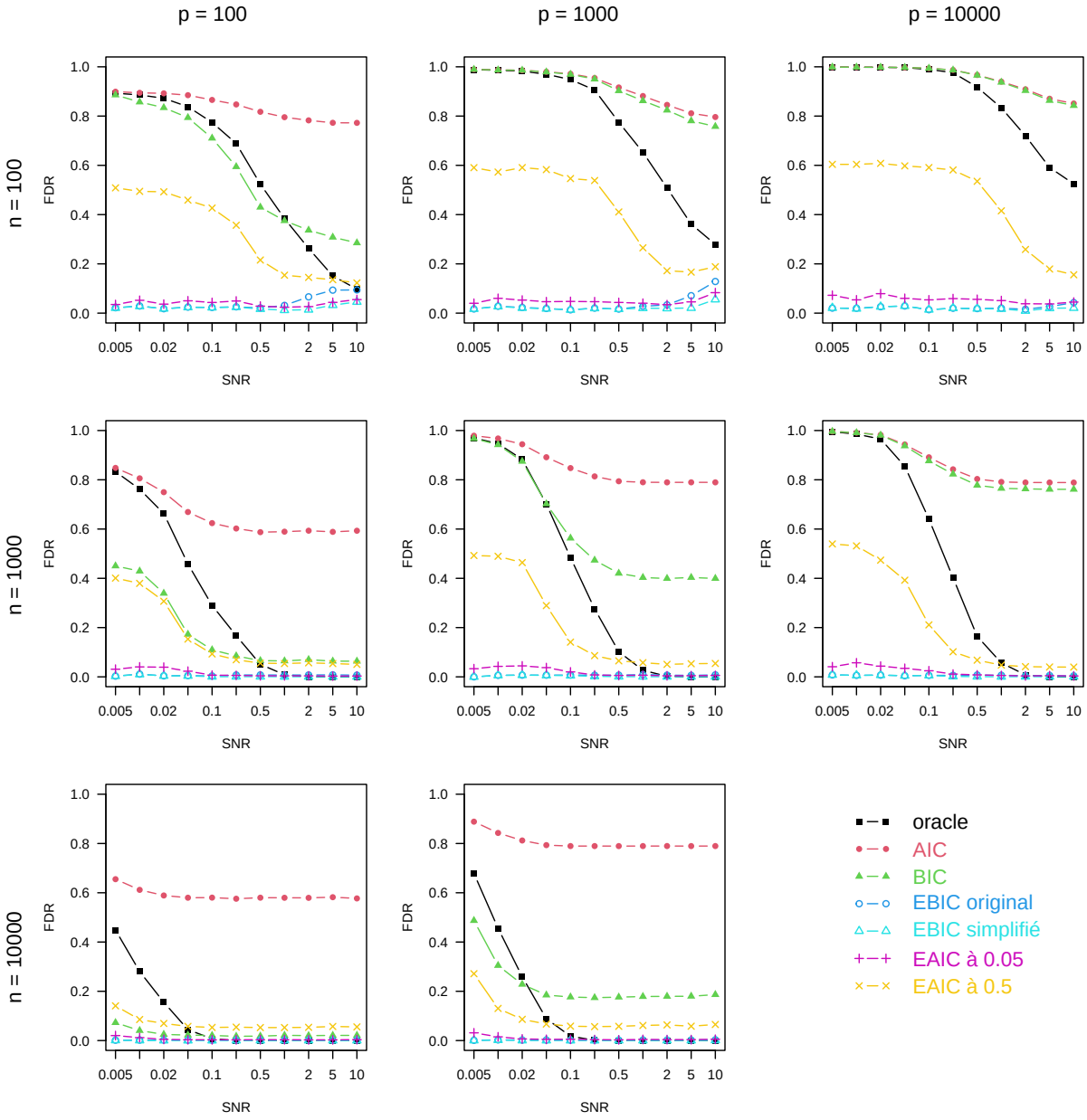

Figure A.5 – Étude de simulation de la procédure complète de sélection de modèle par IC : FDR par paramètre, moyenné sur 1000 simulations. Modèle linéaire, $\rho = 0$.

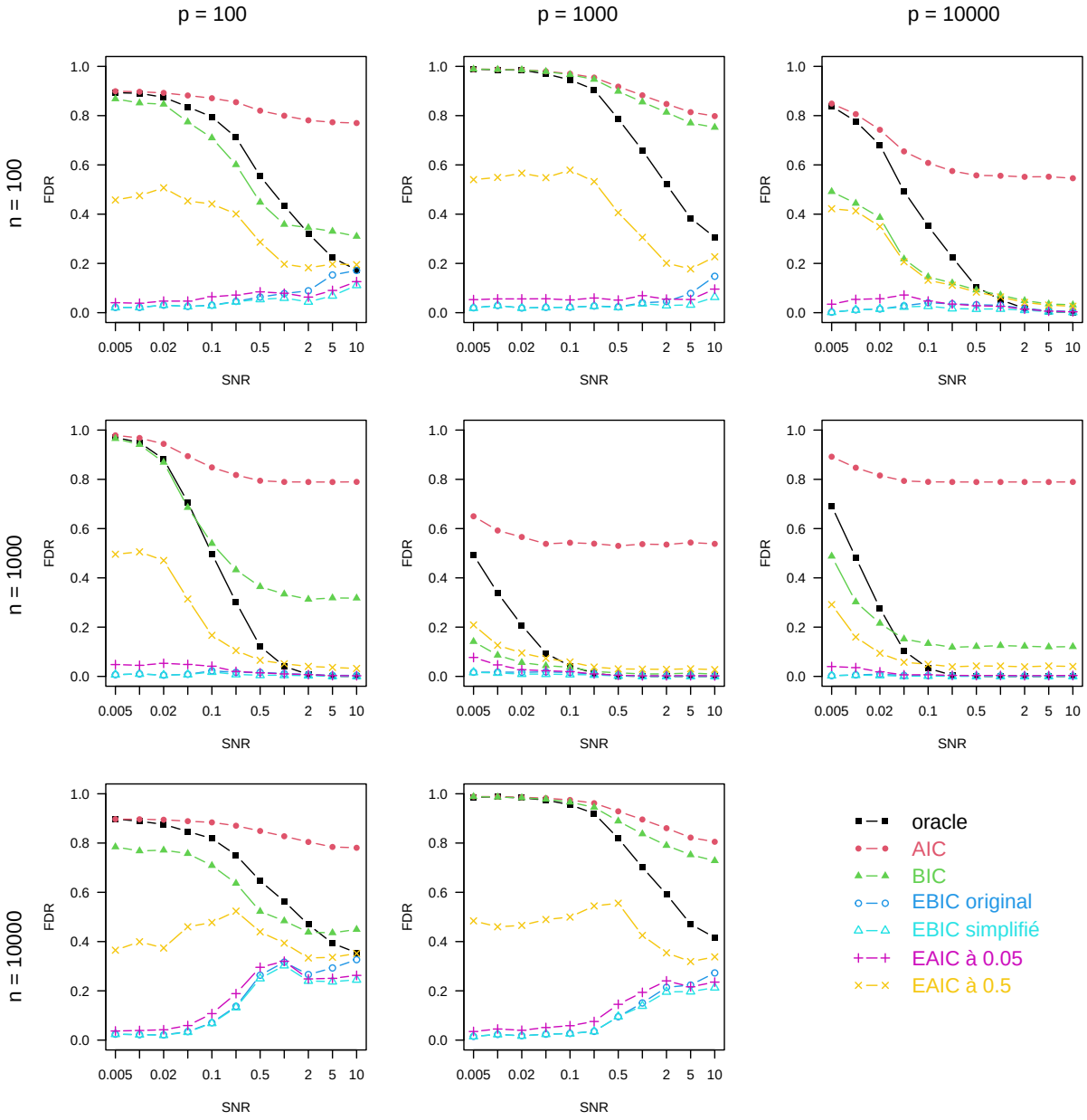

Figure A.6 – Étude de simulation de la procédure complète de sélection de modèle par IC : FDR par paramètre, moyenné sur 1000 simulations. Modèle logistique, $\rho = 0$.

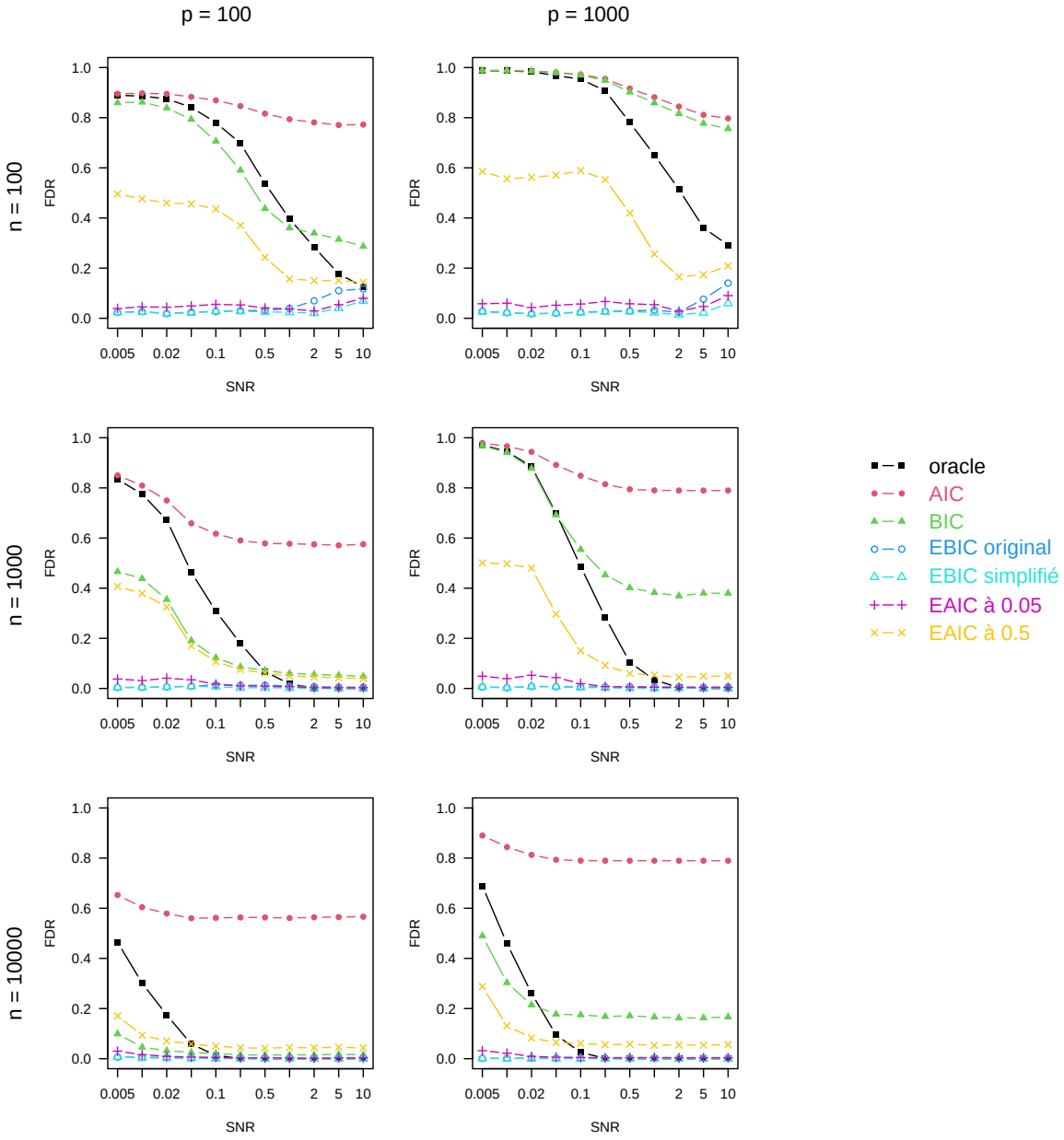

Figure A.7 – Étude de simulation de la procédure complète de sélection de modèle par IC : FDR par paramètre, moyenné sur 1000 simulations. Modèle linéaire, $\rho = 0.5$.

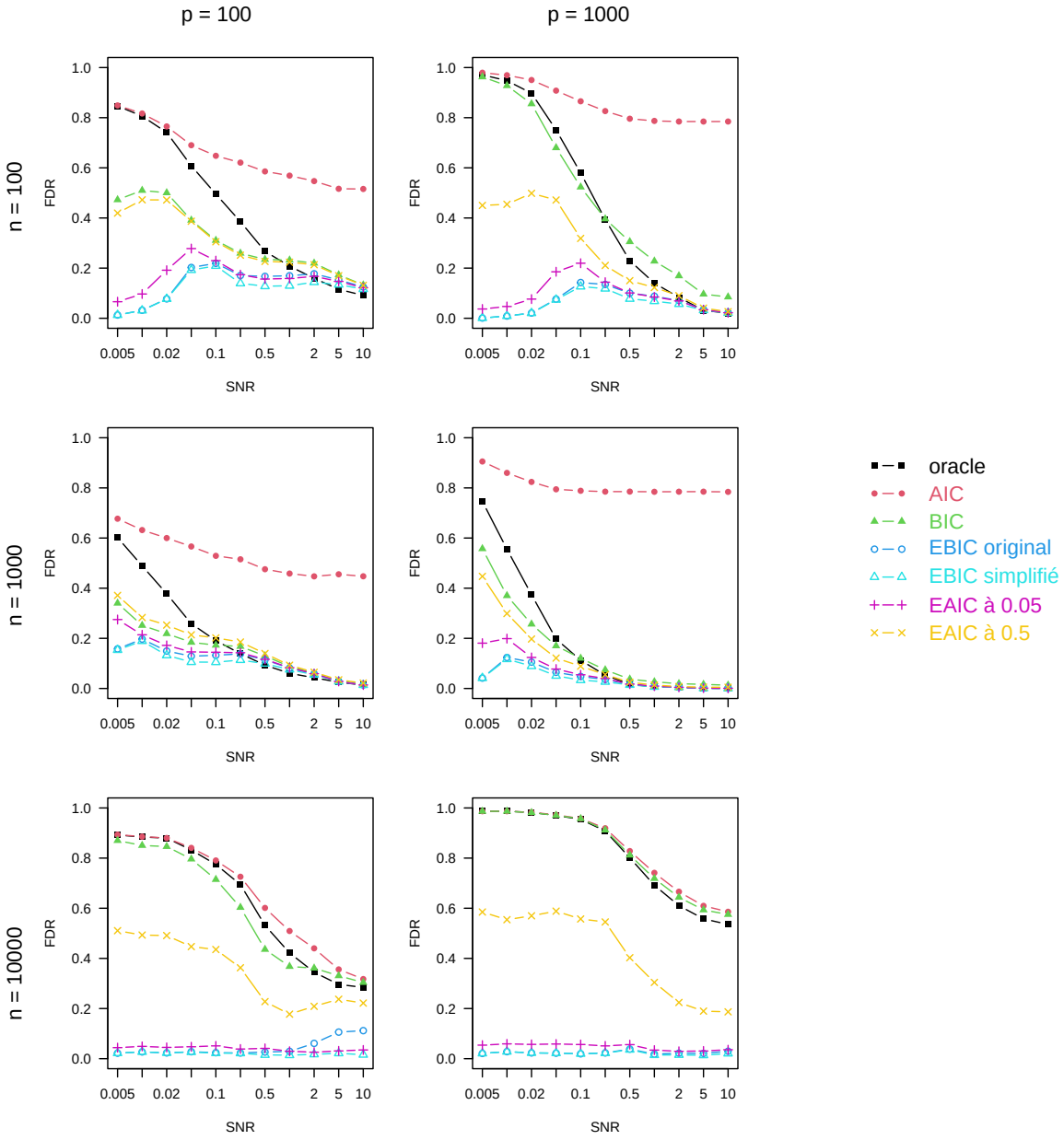

Figure A.8 – Étude de simulation de la procédure complète de sélection de modèle par IC : FDR par paramètre, moyenné sur 1000 simulations. Modèle logistique, $\rho = 0.5$.

A.3 .Comparaison entre minimum global et premier minimum local

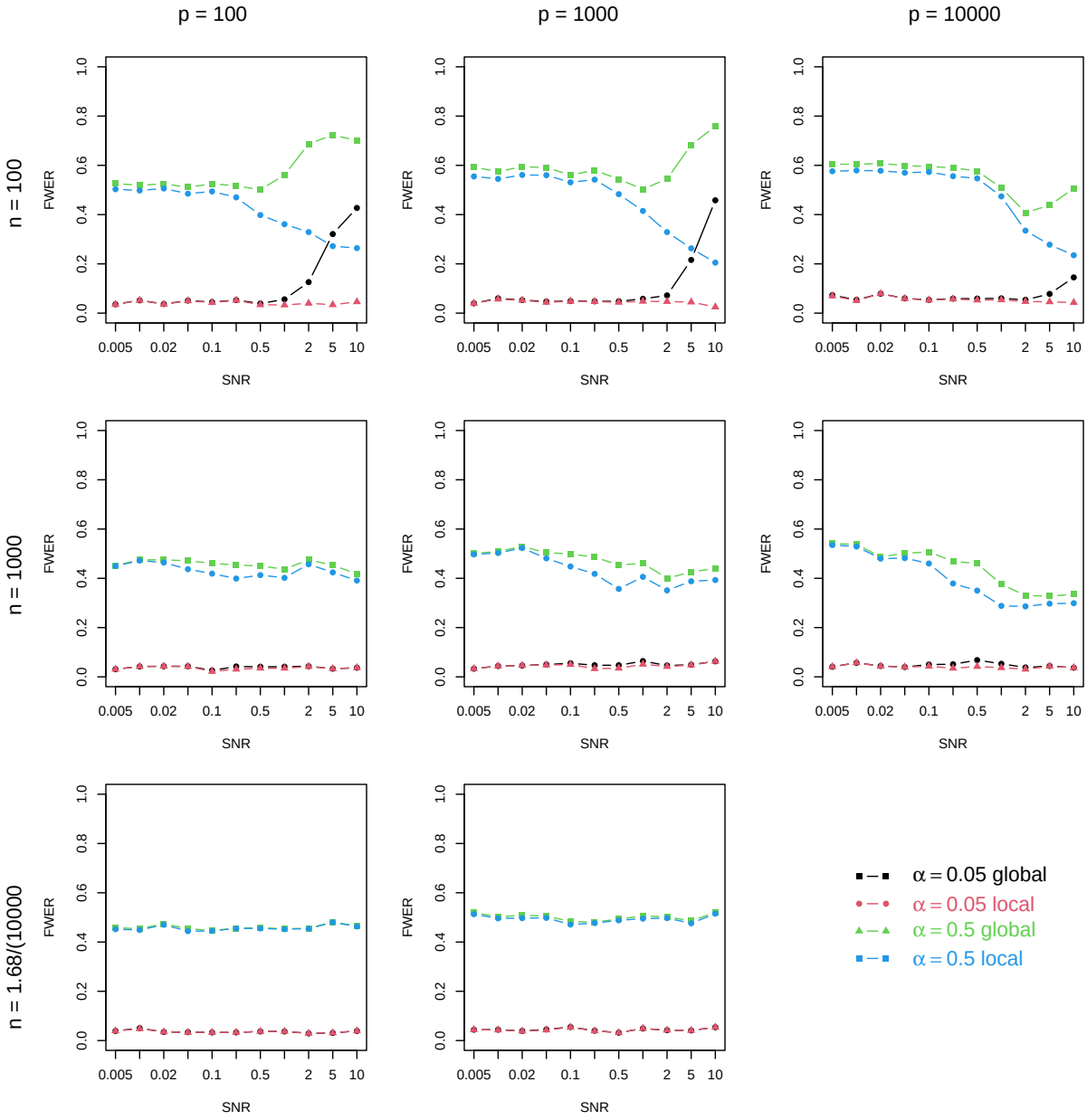

Figure A.9 – Étude de simulation de la procédure complète avec les deux types de minimum de l'EAIC : FWER par paramètre, moyenné sur 1000 simulations. Modèle linéaire, $\rho = 0$.

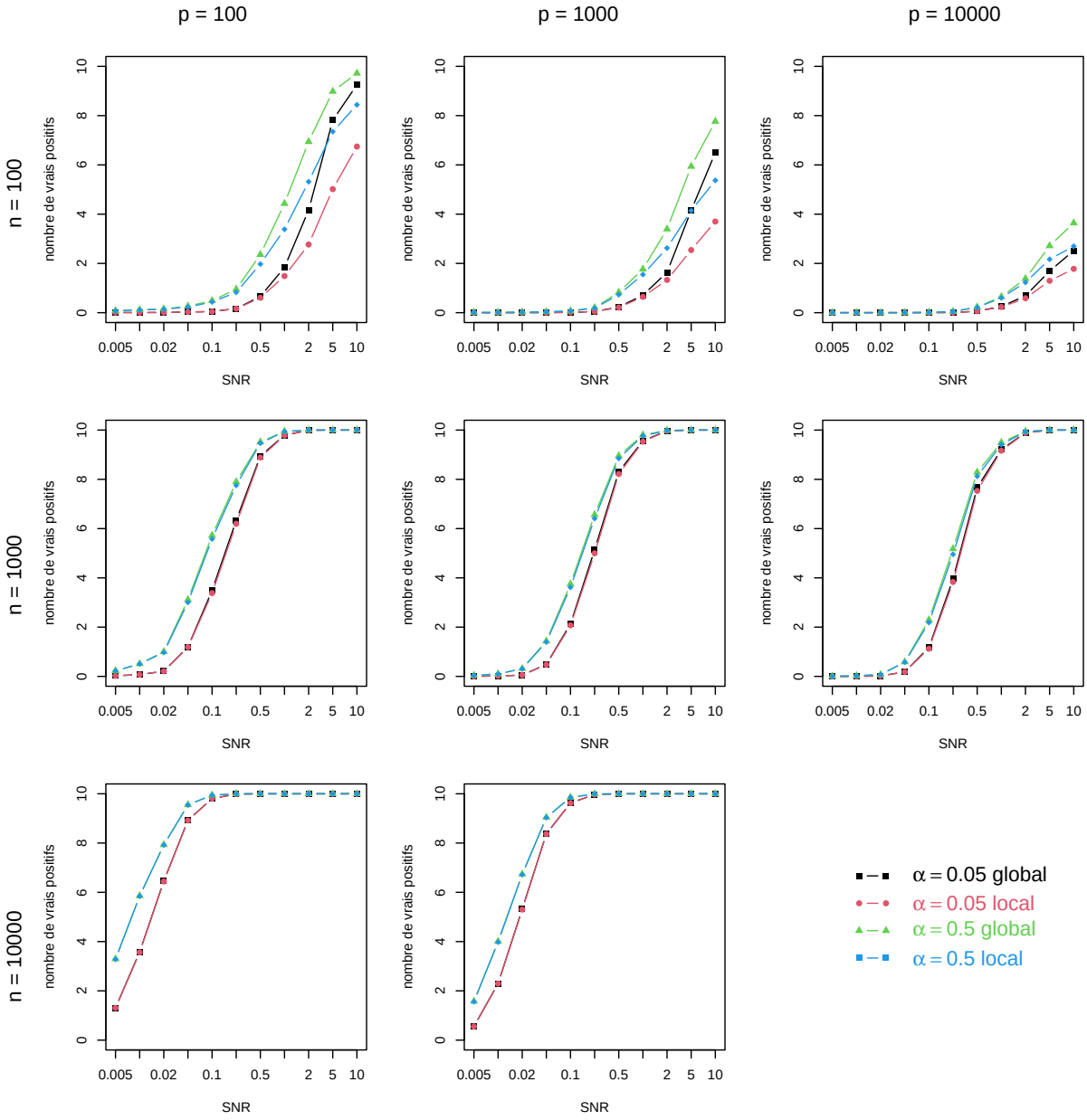

Figure A.10 – Étude de simulation de la procédure complète avec les deux types de minimum de l'EAIC : nombre de vrais positifs par paramètre, moyenné sur 1000 simulations. Modèle linéaire, $\rho = 0$.

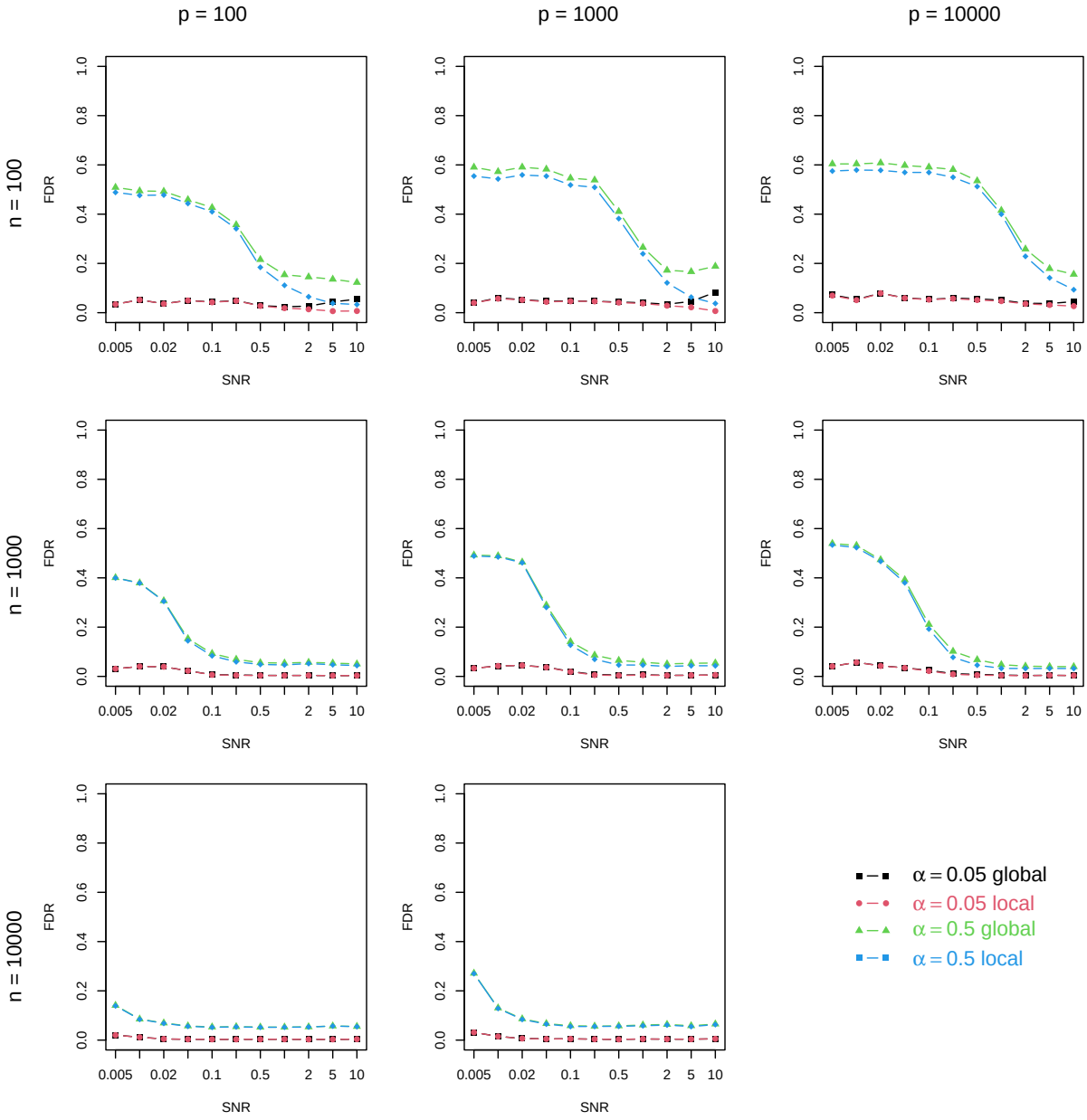

Figure A.11 – Étude de simulation de la procédure complète avec les deux types de minimum de l'EAIC : FDR par paramètre, moyenné sur 1000 simulations. Modèle linéaire, $\rho = 0$.

B - Résultats de simulations complémentaires : sensibilité de la procédure de sélection de variable simulation-calibration

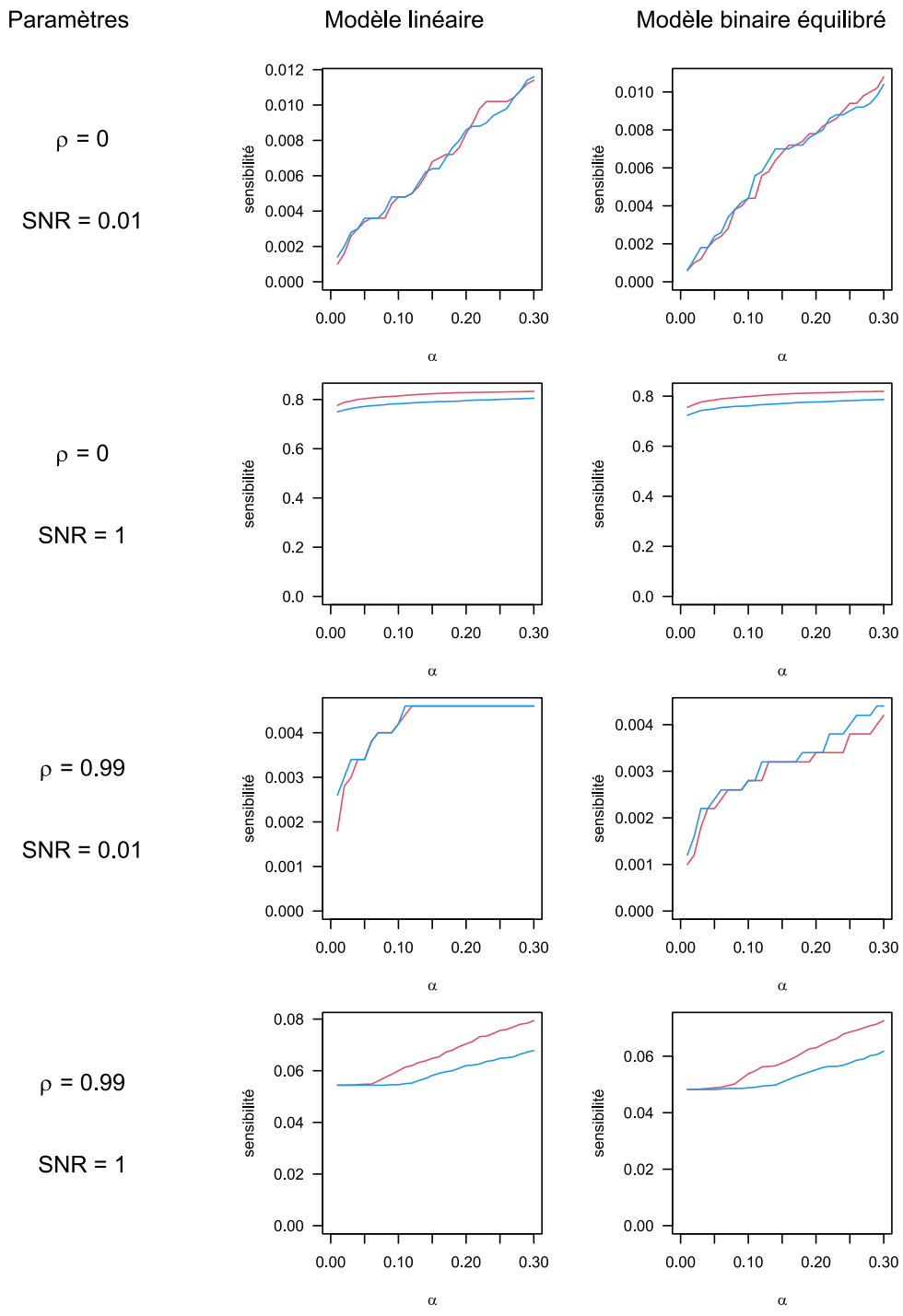

Figure B.1 – Sensibilité de la procédure avec seuillage (bleu) ou *ForwardStop* (rouge) en fonction de α dans 8 scénarios de modèle linéaire ou binaire équilibré ayant 10 variables actives.

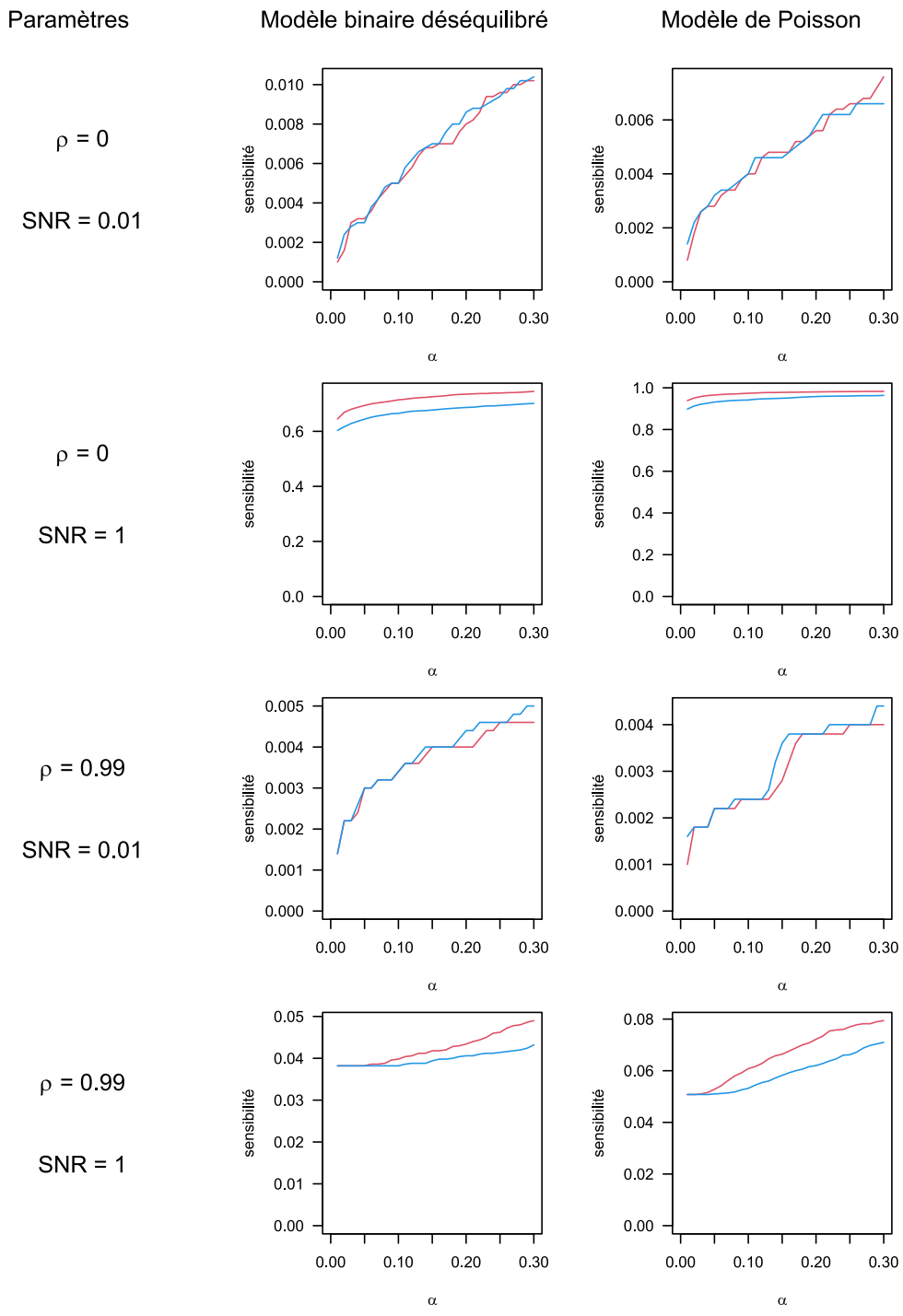

Figure B.2 – Sensibilité de la procédure avec seuillage (bleu) ou ForwardStop (rouge) en fonction de α dans 8 scénarios de modèle binaire déséquilibré ou de Poisson ayant 10 variables actives.