JMS Journées de méthodologie statistique de l'Insee 2018

GETS modelling ou LASSO ? Les différentes méthodes de sélection de variables avec des séries temporelles.

Alizé PAPP(*), Clément ROUSSET(**)

(*) ENSAE

(**) INSEE, Division Synthèse conjoncturelle

alize.papp@ensae.fr clement.rousset@insee.fr

Mots-clés. GETS modelling, LASSO, Sélection de variables, Enquêtes de conjoncture, Étalonnages, Séries temporelles.

Résumé

Les enquêtes de conjoncture fournissent des renseignements précieux pour le conjoncturiste. Néanmoins, un modèle incluant toutes les séries issues des enquêtes de conjoncture serait très peu prédictif en raison de l'arbitrage biais-variance : il est nécessaire de sélectionner les séries qui ensemble seront les plus prédictives. La littérature fournit de nombreuses méthodes pour cela (Lasso, Ridge, Elastic Net, Gets modelling, sélection specific-to-general...) mais peu d'articles comparent ces différentes techniques.

À l'aide de simulations sur données réelles issues d'enquêtes de conjoncture, cet article commence par vérifier l'existence de l'arbitrage biais-variance dans un cadre de prédiction (et non de modélisation comme cela est souvent le cas) et compare ensuite deux méthodes très populaires de sélection de variables, à savoir le Lasso et le Gets modelling.

On commence par introduire à la littérature existant sur le sujet avant d'expliquer la démarche utilisée pour les simulations et d'en analyser les résultats. Dans le cadre de données hautement corrélées et contenant peu d'observation, le Gets modelling l'emporte largement, tant pour sélectionner les bonnes variables que pour prévoir la variable régressée. Ce résultat mérite toutefois d'être nuancé : d'une part, le Lasso est d'un point de vue computationnel bien plus rapide que le Gets modelling, d'autre part lorsque le vrai modèle n'est pas inclus dans l'ensemble des modèles candidats, le Lasso présente des performances souvent comparables au Gets modelling.

Abstract

Business Outlook Surveys are very useful for economic forecasting. Using all available surveys to build a predictive model would however probably lead to some very poor forecasts : because of the bias-variance trade-off, economic forecasters have to first select the variables for their models. Many methods have been developped to select features (Lasso, Ridge, Elastic Net, Gets modelling, specific-to-general selection...), but there is little litterature which compares these different techniques.

This article addresses this issue with simulations on real data. A first round of simulations shows that bias-variance trade-offs does exist for forecasting models (and not only in the case of explanatory models as is often seen). Further simulations compare Lasso and Gets modelling.

The article is organised as follows. A litterature review first introduces the different techniques and their conceptual frames.The method for simulations is then explained and simulation results are analysed. The main result of the simulations is that with this kind of highly correlated data, Gets modelling is far better both at selecting variables and at predicting economic agregates. However Lasso is faster than Gets modelling, and when the true model isn't included in the candidates models, as is the case in reality, Lasso and Gets modelling's efficiencies are very similar.

Introduction

Les enquêtes de conjoncture fournissent des renseignements précieux pour le conjoncturiste : nombreuses, variées et rapidement disponibles, elles couvrent un large pan des économies nationales. Toutefois, face à la multiplicité des séries ainsi disponibles, allant, par exemple, de la confiance des ménages à l'anticipation de l'évolution des prix dans la construction, comment faire le tri et savoir quelles variables sont prédictives vis-à-vis d'agrégats de la comptabilité nationale et donc peuvent améliorer la prévision, et lesquelles leur sont orthogonales ?

En 2006, Éric Dubois et Emmanuel Michaux publiaient un article dans lequel ils présentaient et évaluaient deux méthodes complémentaires, à savoir la trimestrialisation des données d'enquêtes de conjoncture d'une part, et l'automatisation de la sélection des variables issues d'enquêtes de conjoncture d'autre part. Si le premier volet de leur article est très largement appliqué aujourd'hui, la sélection de variables est à la Synthèse conjoncturelle davantage artisanale. Si l'idée d'effectuer une sélection automatique des séries à utiliser est séduisante, toute la difficulté est de choisir une méthode de sélection, et ses paramètres. La recherche sur la sélection de variables s'est en effet largement développée et de nombreuses techniques existent aujourd'hui (General-to-specific, specific-to-general, Lasso, Ridge, Elastic Net...) mais peu d'articles comparent ces différentes méthodes. Dès lors, comment choisir une méthode de sélection de variables ? Dans quelles situations est-il préférable d'utiliser chacune de ces méthodes ?

Cet article se propose de réévaluer l'opportunité d'introduire une sélection de variables automatique, et compare la méthode mise en avant par Dubois et Michaux, et la famille de méthodes qui y est liée, à savoir le General-to-Specific modelling (GETS) et les méthodes de Lasso (Least Absolute Shrinkage and Selection Operator).

On commencera par présenter ces deux approches relativement différentes de la sélection de variables, à savoir d'une part le GETS modelling, issu des économètres de la LSE, et d'autre part le LASSO, en soulignant leurs hypothèses et paramétrages respectifs. Après cela, on présentera les résultats des simulations, au regard des trois critères sus-mentionnés. Enfin, l'on exposera brièvement quelques conseils d'usage, pour les personnes envisageant d'y avoir recours. Pardelà l'usage des enquêtes de conjoncture, la question fondamentale qui se pose est alors celle de l'équilibre possible entre automatisation aveugle et «agnostique» et méthode artisanale mais interprétative.

1 Les critères d'un étalonnage réussi

1.1 Présentation du cadre général du problème

Dans l'ensemble de la discussion, on recherche pour une variable macroéconomique trimestrielle notée Y_t (PIB, salaires, IPC...) à déterminer de "bons" étalonnages à partir de variables explicatives X_{it} (enquêtes, températures...). On suppose réalisés les transformations nécessaires (différenciation avec ou sans log) pour que Y_t et les X_{it} soient considérées stationnaires. Au final on cherche un étalonnage de la forme :$Y_t = \sum_i \beta_i X_{it}$. Il doit vérifier les propriétés suivantes :

1. le modèle doit être "bon", le simulé doit être proche de l'observé
2. le modèle doit permettre de faire de la prévision, c.à.d. que le simulé $\sum_i \hat{\beta}_i X_{it}$ doit aller dans le temps plus loin que l'observé, au moins un trimestre de plus
3. le modèle doit être interprétable facilement par le conjoncturiste (signes dans le bon sens, économe)

1.2 Qu'est-ce qu'un bon étalonnage ?

Un bon étalonnage doit apporter une bonne prévision et donc il faut que le simulé soit proche de l'observé dans le passé et espérer que la relation soit encore vrai dans le futur proche. Plusieurs

possibilités :

1. On peut regarder le R^2 et le RMSE de la régression. C'est le plus naturel et le plus simple. Le problème si l'on veut être pointilleux c'est que ce n'est pas en temps réel : les coefficients β estimés le sont sur une période (le *in sample*) et ces mêmes points contribuent au calcul du RMSE. Ainsi l'erreur à un temps *in sample* est calculée avec un modèle qui connaît déjà le futur de ce temps-là.
2. Le RMFSE est plus rigoureux mais plus gourmand en calcul. Il consiste pour un temps T à estimer le modèle sur les données disponibles à l'instant T uniquement et à déterminer l'erreur de prévision. Ainsi les coefficients β sont estimés à chaque temps, en temps réel. Son calcul est décrit ci-dessous.
3. A partir du moment où on a l'erreur de prévision pour chaque temps en temps réel, on peut pondérer plus fortement les erreurs dans le passé récent, ou aussi chercher le modèle garantissant une erreur maximale de prévision la plus faible

Pour la période d'estimation, on ne prend pas généralement l'ensemble des données disponibles mais on laisse un ou deux ans en fin de période considérant que les comptes ne sont alors que provisoires. Ce genre d'argument est moins pertinent pour les séries de Première Estimation, par exemple la croissance du PIB trimestrielle non pas telle qu'elle est disponible à un instant t avec une partie des données définitives et une fin de série contenant des données provisoires, mais la première croissance publiée pour chaque trimestre, donc uniquement des grandeurs provisoires. Pour les données présentant des cycles, il est préférable que la période d'estimation contienne un nombre entier de cycles.

Le RMSFE se définit comme suit :

$$RMSFE = \sqrt{E[(Y_{t+1} - \hat{Y}_{t+1|t})^2]} \tag{1}$$

Le RMSFE, défini en 1, est la racine carrée de l'erreur de prévision moyenne au carré, ou *Root mean squared forecast error*. Il représente l'erreur "typique" de prévision d'un point. Le RMSFE est une mesure hors échantillon de l'erreur : on compare la prévision réelle à celle que l'estimation à la période précédente aurait annoncé. Pour l'estimer, il faut calculer l'erreur de prévision sur toutes les fenêtres temporelles passées. Pour cela, on effectue les opérations suivantes, pour chaque période, de $t = t_1 - 1$ à $T - 1$

1. On estime le modèle,
2. On effectue une prévision de Y en $t + 1$, on obtient donc $\hat{Y}_{t+1|t}$,
3. On calcule l'erreur de prévision au carré : $(Y_{t+1} - \hat{Y}_{t+1|t})^2$

On dispose alors d'un vecteur d'erreurs de prévision hors échantillon, sa moyenne est le RMSFE estimé, noté RMSFE par la suite pour plus de simplicité.

1.3 Qu'est-ce qu'un étalonnage interprétable ?

Le conjoncturiste doit pouvoir intrepréter le modèle, c'est-à-dire traduire en une phrase la prévision du modèle du type "les enquêtes en baisse conduisent à un ralentissement", "la tendance haussière des permis induiraient une accélération de" ... Le problème se pose surtout pour les séries mensuelles.

1.3.1 Cas des séries trimestrielles

Pour une série trimestrielle comme par exemple le TUC dans l'industrie, il suffit a priori d'avoir une bonne performance du modèle et le bon signe. Un modèle du type

$$\Delta log(FBCF_{equipement}) = \beta TUC,$$

avec un $\beta > 0$ est interprétable : "Le TUC est élevé et donc la FBCF serait dynamique". On pourrait imaginer aussi

$$\Delta log(FBCF_{equipement}) = \beta \Delta TUC,$$

avec un $\beta > 0$, "Le TUC est en hausse et donc la FBCF serait dynamique".

1.3.2 Cas des séries mensuelles

Pour une série mensuelle on peut avoir trop de choix. A partir d'une série mensuelles X_m, on peut extraire 3 séries trimestrielles sans perte d'information X_{m1},X_{m2} et X_{m3}. Un étalonnage du type

$$\Delta log(PIB) = \beta X_{m1} + \gamma X_{m3},$$

sera difficile à interpréter si $\beta > 0$ et $\gamma < 0$ "le bon mois 1 donne une activité en hausse mais le bon mois 3 le fait baisser". Plusieurs solutions :

1. on prend la moyenne trimestrielle, en prenant bien la moyenne partielle si l'intégralité du dernier trimestre n'est pas tombée sinon on n'en prend pas en compte l'information. Si on n'a que deux mois du dernier trimestre, soit on prend la moyenne des deux premiers mois dans le passé aussi, soit on la prend uniquement pour le dernier trimestre en restant sur la moyenne des trois trimestres avant
2. $Y_t = \beta X_{\text{dernier mois}} + \gamma(X_{\text{dernier mois}} - X_{\text{dernier mois}-1}) + ...$, le premier terme prend le niveau de l'enquête, les autres termes sont des termes de variations, tous les signes doivent être positifs.

1.4 La recherche du bon modèle en pratique

En 1995, dans un Insee Méthode [1], Grégoir et Le Rey décrivent la méthode employée à l'INSEE :

> Une investigation systématique est effectuée ; parmi les équations ayant passé avec succès l'ensemble des tests, celle qui explique la plus grande variabilité de la variable étalonnée pour un nombre raisonnable de variables explicatives est sélectionnée.

En d'autres termes, à raison de p variables candidates pour un étalonnage, 2^p étalonnages sont testés [2], ce qui représente un coût de calcul exponentiel.

En 2006, Dubois et Michaux introduisent une approche qu'ils désignent eux-mêmes d' « automatique » [3], à savoir l'utilisation de l'algorithme de Krolzig et Hendry [4] pour sélectionner les soldes les plus pertinents. Ils la mettent à l'épreuve et obtiennent des performances légèrement meilleures que celles obtenues avec des modèles VAR, et très proches de celles des prévisions conjoncturelles de l'INSEE. Ce travail s'insère dans une démarche de réintroduction à la synthèse conjoncturelle de l'algorithme de Krolzig et Hendry. Nous avons utilisé un développement de l'algorithme de Krolzig et Hendry, pour actualiser ces résultats et comparer la performance de cet algorithme à d'autres techniques de sélection de variables.

1. E. Le Rey S. Grégoir. « La pratique des étalonnages dans l'analyse conjoncturelle ». In : *INSEE Méthodes* 33 (1995), p. 52–98.

2. S'il y a p variables envisagées, on teste tous les modèles possibles en incluant ou non chaque variable. Il y a donc deux choix par variable (l'inclure ou non), soit 2^p modèles en tout.

3. E. Michaux E. Dubois. « Etalonnages à l'aide d'enquêtes de conjoncture : de nouveaux résultats ». In : *Economie et prévision* 172 (2006), p. 11–28.

4. D. Hendry H. Krolzig. « Computer Automation of General-to-Specific Model Selection Procedures ». In : *Journal of Economic Dynamics and Control* (2001).

2 Comment choisir un bon étalonnage ? Revue de la littérature existante

2.1 Backward et forward selection

La méthode la plus intuitive consiste à commencer par un modèle avec une unique constante, puis ajouter progressivement des variables si elles augmentent la qualité de l'étalonnage au regard d'un certain critère (son R^2 ou son RMSFE par exemple), éventuellement en ajoutant à chaque étape la variable qui améliore le plus l'étalonnage. Il s'agit de la méthode dite de *forward selection*. Cette méthode a pour elle sa simplicité. Elle a néanmoins de nombreux défauts, le plus important d'entre eux étant sa dépendance au sentier (ou *path dependency* en anglais). Les variables déjà présentes à une étape t influencent le choix de la variable qui sera la plus utile en t et sera donc introduite en $t+1$. Ainsi, l'étape t est conditionnée par toutes les étapes précédentes, de 0 à $t-1$. Selon les chemins pris, les modèles finaux ne sont donc pas les mêmes. De plus, comme les variables y sont ajoutées individuellement, une combinaison linéaire prédictive risque d'être ignorée, à moins que les variables de la combinaison aient seules déjà un pouvoir prédictif. Par exemple, si la balance commerciale est prédictive mais ni les exportations seules, ni les importations, cette méthode de sélection risque d'occulter ce fait et de ne pas retenir la balance commerciale.

Réciproquement, on définit aussi le *backward selection* comme la méthode consistant à partir d'un modèle avec toutes les variables candidates, et à les retirer petit à petit, si cela améliore la qualité du modèle. Là encore, il y a un risque de dépendance au sentier.

2.2 GETS modelling

Le GETS est une méthode de la famille des *backward selections* qui introduit une recherche à sentiers multiples pour répondre au risque de dépendance au sentier.

2.2.1 De la critique du data mining au general to specific modelling.

Présentée en 1999 par Hoover et Perez [5], l'approche *General-to-Specific*, ou GETS, se propose de reproduire la méthode économétrique dite de la LSE (selon les auteurs alors mise en place dans les laboratoires, et enseignée dans les salles de cours, de cette école). Dans les années 1980 et 1990, les méthodes de choix de modèle (et donc de sélection de variables) font l'objet d'un vif débat et la méthodologie de la LSE est critiquée, notamment par Pagan [6] qui montre que le chemin qui mène au choix d'un modèle est très important et insuffisamment explicité, ainsi que par Lovell [7] qui s'oppose à ce qui s'appelait alors le "*data mining*", à savoir la recherche systématique de liens au sein des données de sorte à y trouver des corrélations, et montre par une argumentation qualitative et par simulations que le risque de première espèce des tests de significativité d'un coefficient dans un modèle choisi à l'issue d'une sélection de modèle est bien plus élevé que les t-tests sur le modèle seul ne le laissent penser.

En 1999, Hover et Perez répondent à ces critiques en introduisant le principe d'un recherche multi-path, de sorte à éviter la dépendance au sentier et en formalisant la démarche de recherche general-to-specific. Ainsi, ils répondent d'une part à la critique de Pagan en formulant précisément le déroulement de la recherche de modèle et les règles de prise de décision. Ils répondent

5. S. Perez K. Hoover. « Data mining reconsidered : encompassing and the general-to-specific approach to specification search ». In : *Econometrics Journal* 2 (1999), p. 1–25.

6. A. Pagan. « Three Econometric Methodologies : A Critical Appraisal ». In : *Journal of Economic Surveys* 1.1 (1987), p. 3–24.

7. Michael C. Lovell. « Data Mining ». In : *The Review of Economics and Statistics* 65.1 (1983), p. 1–12.

d'autre part à Lovell grâce à cette formalisation, puisqu'ils effectuent des simulations en suivant le même procédé que lui mais en évaluant leur propre méthode de sélection de variables. Cette formalisation dont le but n'était au début que de répondre aux critiques du general-to-specific modelling donne naissance au premier algorithme de GETS modelling.

2.2.2 L'algorithme de Hoover et Perez

Hoover et Perez créent ainsi le premier algorithme de GETS. Il se déroule de la manière suivante :

1. Un modèle général est constitué à partir de toutes les variables candidates, on l'estime et on effectue des tests sur ce modèle [8]. Si un test échoue, on ne l'utilisera plus par la suite. Si plusieurs tests échouent, on cherche un autre modèle général.
2. On prend les dix variables aux t-statistiques les moins significatives, chacune définit un chemin. Pour chaque chemin, on estime le modèle en éliminant la variable peu significative qui le définit. On estime alors la régression et on supprime petit à petit les variables les moins significatives (à la plus petite p-value), tant que les modèles passent les tests. Les tests effectués sont ceux en 1 ainsi qu'un F-test.
3. Le processus pour chaque chemin s'achève lorsque, soit lorsque toutes les variables sont significatives, soit lorsque l'on ne peut enlever une variable supplémentaire sans que le modèle échoue aux tests.
4. On obtient donc 10 modèles finaux, et on retient celui qui a le plus faible RMSE.

2.2.3 Postérité et développements de l'algorithme

L'algorithme de Hoover et Perez a été repris et modifié de nombreuses fois ensuite et a donné naissance à un vaste champ théorique visant à améliorer les techniques de GETS. Parmi les principales évolutions, on notera PcGets, développé à partir de l'article de Hendry et Krolzig de 1999 [9], que Dubois et Michaux utilisent [10], qui apporte plusieurs améliorations comme l'usage des critères d'information (AIC, BIC...) pour le choix du modèle final en étape 4, l'ajout d'une étape préliminaire de suppression des variables qui ne sont pas significatives dans le modèle général à un seuil moins exigeant que dans les tests de significativité suivants et la modification du seuil de significativité d'un test plutôt que sa suppression lors de la première étape. Un autre article qui a fait histoire dans ce domaine est celui de Sucarrat et Genaro (2009) [11]. Les auteurs y généralisent la méthode de Hoover et Perez en rendant possible l'existence d'un terme autorégressif en plus des variables explicatives, ainsi qu'un résidu hétéroscédastique avec un modèle log- GARCH(pour log Generalised AutoRegressive Conditional Heteroscedasticity). Ils enrichissent alors l'algorithme de Hoover et Perez en modifiant les tests ainsi qu'en ajoutant systématiquement un modèle vide dans la liste finale, du moment qu'il passe les tests. De plus, contrairement à Hoover et Perez, ils examinent autant de sentiers qu'il y a de variables non significatives dans le modèle général. D'après leurs propres simulations, leur algorithme est plus performant que celui de Hoover et Perez et celui de Hendry et Krolzig, et notamment en ce qu'il a moins tendance à retenir des variables à tort. Sucarrat et Genaro ont développé une librairie

8. Hoover et Perez proposent un test de normalité des résidus, un d'autocorrélation des résidus jusqu'au second ordre, un test d'autocorrélation conditionnelle des résidus ARCH jusqu'au second ordre, un test de stabilité dans l'échantillon et un test hors échantillon, cf p. 9

9. H. Krolzig D. Hendry. « Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez ». In : *Econometrics Journal* 2.2 (1999), p. 202–219.

10. E. Dubois, op. cit.

11. A. Escribano G. Sucarrat. « Automated Financial Model Selection : General-to- Specific Modelling of the Mean and Volatility Specifications ». In : *Oxford Bulletin of Economics and Statistics* (2012), p. 716–735.

R mettant en place leur algorithme, appelée *gets*, et cet algorithme que nous avons évalué par la suite.

2.3 LASSO

Le LASSO est une méthode de sélection de variable qui vient d'un domaine relativement distinct, à savoir les méthodes de régularisation. De ce fait, l'approche de la sélection de variables est quelque peu différente de celle du GETS modelling, qui vient des économétriciens, et principalement de ceux formés à la LSE. Les méthodes de régularisation partent d'un problème d'optimisation dans lequel on cherche à minimiser la somme d'un terme d'erreur entre le réalité et le simulé, comme le RMSE, et d'une pénalisation. Dans le cadre du LASSO, présenté en 1996 par Robert Tibshirani [12], cette pénalisation est l'ampleur des coefficients de la régression OLS, mesurée comme somme des normes L1 des coefficients, ie somme des valeurs absolues des coefficients. Ainsi, l'estimation LASSO des coefficients est définie par 2 :

$$\arg \min_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \text{ sous la contrainte } \sum_{j=1}^{p} |\beta_j| \leq t. \tag{2}$$

$$\arg \min_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

où λ est inversement proportionnel à t

En intégrant la contrainte sur la taille des coefficients par le multiplicateur de Lagrange, on obtient 3, dans lequel λ est le paramètre de régularisation. Ainsi, si t vaut 0 (réciproquement λ tend vers $+\infty$), tous les coefficients sont contraints à 0, à l'inverse si t tend vers $+\infty$ (réciproquement λ vaut 0), on retombe dans le cas OLS.

Pour comprendre l'idée du Lasso, on peut se placer dans le cas où les régresseurs sont orthonormaux, on obtient alors :

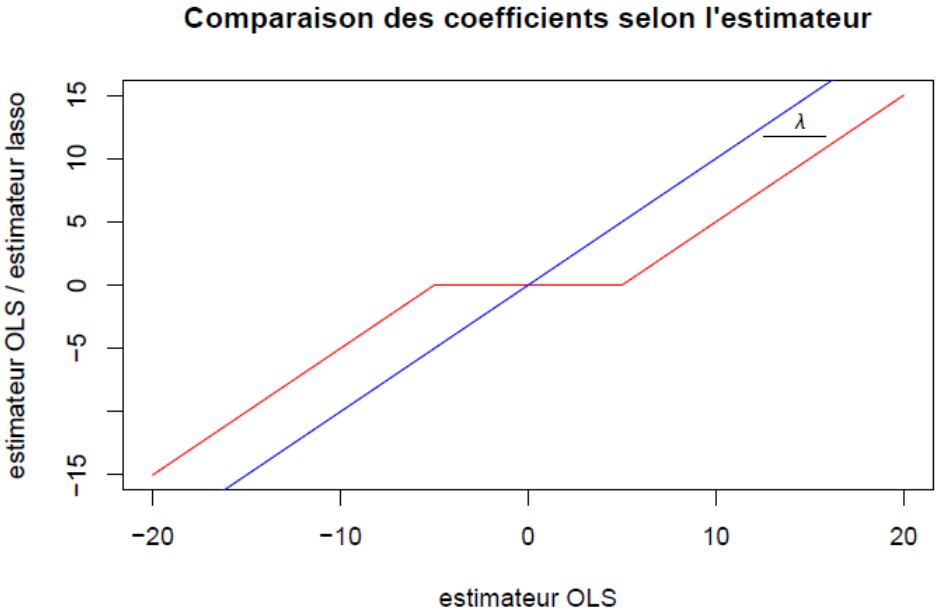
$$\hat{\beta}_j^L = sign(\hat{\beta}_j^{LS}) \cdot \max(0, |\hat{\beta}_j^{LS}| - \lambda) \tag{4}$$

Ainsi, jusqu'à un certain seuil, les coefficients sont annulés, et au-delà de ce seuil, ils sont retranchés d'une constante. Le graphique 1 illustre ce phénomène : la courbe bleue est l'estimation par OLS, il s'agit donc de la première bissectrice du plan, tandis que la seconde, qui décrit l'évolution de l'estimation lasso lui est parallèle sur la majorité du plan, à l'exception des valeurs pour lesquelles l'estimation OLS est inférieure au terme de pénalisation.

Toute la question est alors de trouver le juste λ, suffisamment grand pour que le lasso diminue les coefficients vers 0 en valeur absolue, mais pas trop pour que les coefficients des régresseurs prédictifs restent distincts de 0. La solution la plus fréquente est de l'obtenir à l'aide d'une cross-validation, décrite par Tibshirani dans son article fondateur de 1996 [13]. Le principe d'ensemble de cette technique est d'estimer l'erreur de prévision en fonction de λ et de choisir le λ qui la minimise. Pour cela, on pourrait estimer le Lasso pour différents λ, néanmoins on risquerait alors d'avoir un paramètre de régularisation dépendant des données : il y a un risque d'*overfitting*. Plutôt, on divise l'échantillon en k sous-échantillons et on estime l'erreur en fonction de λ pour

12. R. Tibshirani. « Regression Shrinkage and Selection via the Lasso ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), p. 267–288.

13. Ibid., p. 275.



[H]

FIGURE 1 – Effet du Lasso sur la valeur d'un coefficient

chaque échantillon. On en déduit alors l'erreur de prévision en fonction de λ, par exemple comme moyenne des erreurs pour un λ donné entre les différents échantillons. On choisit alors le λ qui minimise l'erreur de prévision.

La technique du Lasso a connu un large succès dans le domaine de la sélection de variables notamment en raison de son faible coût computationnel. En utilisant l'algorithme du *Least angle regression*, le Lasso a la même complexité computationnelle qu'une régression OLS [14]. Le principe du Lasso a été adapté à de nombreux types de régressions, parmi lesquels on pourra citer les modèles VAR [15]. D'autres développements se sont penchés sur les conditions de validité des estimations par LASSO. Zhao et Yu [16] ont notamment montré que le Lasso sélectionne le bon modèle asymptotiquement lorsque les variables candidates qui ne doivent pas être incluses ne sont pas "représentables" [17] par les variables candidates qui doivent être incluses. En particulier, cette condition de représentation est violée lorsque les variables candidates sont hautement corrélées. Le Lasso ne sélectionne alors pas le bon modèle, y compris asymptotiquement.

2.4 GETS ou LASSO ? Les comparaisons des deux méthodes dans la littérature

Le GETS et le LASSO viennent de familles théoriques bien distinctes, d'un côté de la réflexion d'économètres sur la sélection de modèle, de l'autre de statisticiens sur la régularisation de régressions. Par conséquent, peu d'articles les comparent directement, Epprecht et ses coau-

14. B. Efron T. Hastie I. Johnstone R. TIBSHIRANI. « Least angle regression ». In : *Ann. Statist.* 32.2 (avr. 2004), p. 407–499.

15. C. Ya-Mei N. HSU H. Hung. « Subset Selection for Vector Autoregressive Processes Using Lasso ». In : 52 (mar. 2008), p. 3645–3657.

16. B. Yu P. ZHAO. « On Model Selection Consistency of Lasso ». In : *Journal of Machine Learning Research* 7 (2006), p. 2541–2563.

17. Les auteurs précisent le sens de cette appellation dans leur article, auquel on se reportera pour une description précise de ces conditions.

teurs [18] reportent même qu'il n'y aurait aucun article antérieur au leur sur ce sujet en 2013. Epprecht et ses coauteurs comparent les performances du logiciel Autometrics, qui met en place une variante du *GETS modelling*, et du Lasso et AdaLasso tels qu'implémentés sous Matlab dans le package *glmnet*. Les conclusions d'Epprecht et ses coauteurs sont les suivantes :

1. Autometrics estime mieux les paramètres (en termes de biais et de variance), ce qui était attendu dans la mesure où le Lasso restreint les paramètres vers 0 quand bien même il les sélectionne (voir ci-dessus pour une explication).
2. AdaLasso est la méthode la plus performante en termes de sélection de variables, sauf pour le plus petit échantillon, de taille 50.
3. Autometrics a la meilleure capacité prédictive (en termes de RMSFE), y compris quand AdaLasso sélectionne mieux les variables.

Le travail d'Epprecht et ses coauteurs, bien que courageux, souffre d'une limite majeure : ses conclusions ne s'appliquent que dans le cas orthogonal. Or, à moins d'effectuer des prétraitements sur les variables, cela est rarement le cas dans la pratique.

D'autres travaux évaluent le Lasso pour la sélection de variables face à des techniques plus classiques. En particulier, dans son article fondateur, Tibshirani [19] effectue des simulations pour comparer le Lasso à une simple régression OLS, à deux méthodes de régularisation (le Garotte et le Ridge) et à une méthode de *backward selection*, le subset selection avec *cross-validation*. Tibshirani simule différents échantillons selon plusieurs modalités, mais toujours avec une corrélation entre les régresseurs, et il obtient que chaque méthode de sélection de variables a ses aires d'expertise (et de défaillance). En particulier, le Lasso serait la technique la plus efficace dans le cadre d'un nombre relativement limité de vrais régresseurs aux effets modérés.

Plus récemment, I. Savin [20] a effectué des simulations pour comparer les performances du Lasso et de méthodes de sélection de variable dites heuristiques [21]. Savin tient compte de l'article de Zou et Yu [22] décrit plus haut, qui explique les pauvres performances du Lasso lorsque les régresseurs sont corrélés entre eux. Savin effectue donc des simulations en générant des régresseurs plus ou moins corrélés et il retrouve empiriquement les défauts du Lasso dans un cadre de forte corrélation, ce qui induit une forte erreur de prévision. Savin observe de plus que dans toutes les configurations, le Lasso tend à retenir moins de variables, ce qui augmente son erreur de prévision mais lui donne un fort pouvoir de discrimination des variables peu pertinentes. Savin en conclut que le Lasso est particulièrement efficace dans le cas où la plupart des variables sont informatives et sont peu corrélées entre elles. Savin ajoute à cela une estimation du temps moyen de computation, et souligne l'efficacité computationnelle du lasso.

Dans ce cadre, il semble utile d'effectuer nos propres simulations, de sorte à :

1. évaluer getsm (librairie R implémentant l'algorithme de GETS modelling de Sucarrat et Genaro (2009) [23]) et non Autometrics (logiciel implémentant l'algorithme de Doornik), contrairement à l'article d'Epprecht,
2. évaluer le Lasso dans le cadre de variables hautement corrélées, comme Tibshirani et Savin, en le mettant en regard avec du GETS modelling,

18. Camila Epprecht et al. *Variable selection and forecasting via automated methods for linear models : LASSO/adaLASSO and Autometrics*. Documents de travail du Centre d'Economie de la Sorbonne. Université Panthéon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne, nov. 2013.

19. Tibshirani, op. cit.

20. I. Savin. *A Comparative Study of the Lasso-Type and Heuristic Model Selection Methods*. Working Paper 04. COMISEF, 2010.

21. L'idée d'une méthode heuristique est de sonder l'espace des 2^p modèles possibles sans exhaustivité mais avec une part d'aléatoire pour approximer le point de cet espace qui minimise l'erreur de prédiction.

22. P. Zhao, op. cit.

23. G. Sucarrat, op. cit.

3. effectuer l'évaluation avec des données réelles et susceptibles d'être utilisées pour la prévision conjoncturelle, et non des données simulées, contrairement à tous les articles susmentionnés.

3 Les données

3.1 De nombreuses séries temporelles

Les données sont issues des enquêtes de conjoncture des *Business and Consumer Surveys* (BCS). En France, ces enquêtes sont menées par la division des Enquêtes conjoncturelles de l'Insee. Afin que toutes les données présentent la même fréquence, les données utilisées dans les simulations sont uniquement les données mensuelles du BCS, à savoir celle sur l'industrie manufacturière, la construction, les consommateurs, le commerce de détail, les services et les services financiers. En se cantonant aux données mensuelles issues du BCS, le nombre de modèles possibles est déjà impressionnant (à raison de 46 variables mensuelles, il y a 2^46 modèles possibles). Les séries représentent des soldes d'opinion après pondération par le chiffre d'affaire. Bien que les questions portent généralement sur des variations explicitement hors saisonnalité, les réponses tendent à présenter une saisonnalité, elles sont donc corrigées des variations saisonnières et des jours ouvrés.

Le nombre d'enquêtés est disponible en annexe A1. La documentation du BCS [24] fait état d'un taux de non-réponse aux alentours de 20 à 35 % selon les pays. Le site de la Commission européenne donne le contenu des questionnaires exhaustivement [25].

Les données sont des séries temporelles, ce qui conduit en pratique à un certain nombre de contraintes. Premièrement, la fréquence n'est pas nécessairement la même pour toutes les séries, il faut donc procéder à une première étape d'harmonisation des fréquences des séries avant de les sélectionner, ce qui peut augmenter le nombre de séries candidates, et donc augmenter le coût computationnel de la sélection [26]. Deuxièmement, la date de début de la série la plus tardive limitera le volume de donnée disponible pour la sélection parmi toutes les variables : ajouter une série candidate n'a donc pas toujours un effet positif sur la qualité de la sélection de variables, quand bien même la série serait très prédictive de l'aggrégat à prédire. Ce problème est amplifié par le fait qu'en pratique, les séries disponibles remontent dans le meilleur des cas aux années 1990, ce qui ne représente malheureusement pas assez d'information pour que les estimations soient véritablement convergentes. Enfin, le caractère temporel des données impose d'effectuer des tests de non-stationnarité.

3.2 Des séries fortement corrélées

Les réponses aux enquêtes sont fortement corrélées entre elles comme on peut le voir dans la matrice des corrélations représentée sur le graphique 2. La matrice représente les corrélations entre chaque paire de variables, pour des raisons de lisibilité, on n'a représenté que sa partie supérieure, puisque cette matrice est symétrique. Les corrélations positives y sont colorées en rouge, et les négatives en bleu, il apparaît assez clairement que les faibles corrélations (peu colorées) sont assez peu nombreuses. Plus précisément, lorsque l'on étudie cette matrice, on

24. *The joint harmonised EU programme of business and consumer surveys*. User Guide. European Commission, 2016. URL : `https://ec.europa.eu/info/files/user-guide-joint-harmonised-eu-programme-business-and-consumer-surveys_en`.

25. https://ec.europa.eu/info/business-economy-euro/indicators-statistics/economic-databases/business-and-consumer-surveys/methodology-business-and-consumer-surveys/national-questionnaires_en

26. Par exemple, si parmi les séries candidates une est trimestrielle, il faudra trimestrialiser toutes les séries mensuelles, ce qui multiplie par quatre leur nombre

obtient que 75% des corrélations en valeur absolue sont supérieures à 0,2, et que la corrélation (en valeur absolue) moyenne est de 0,5. On en déduit que, d'après Zhao et Yu [27], le Lasso risque de ne pas avoir de très bonne performances.

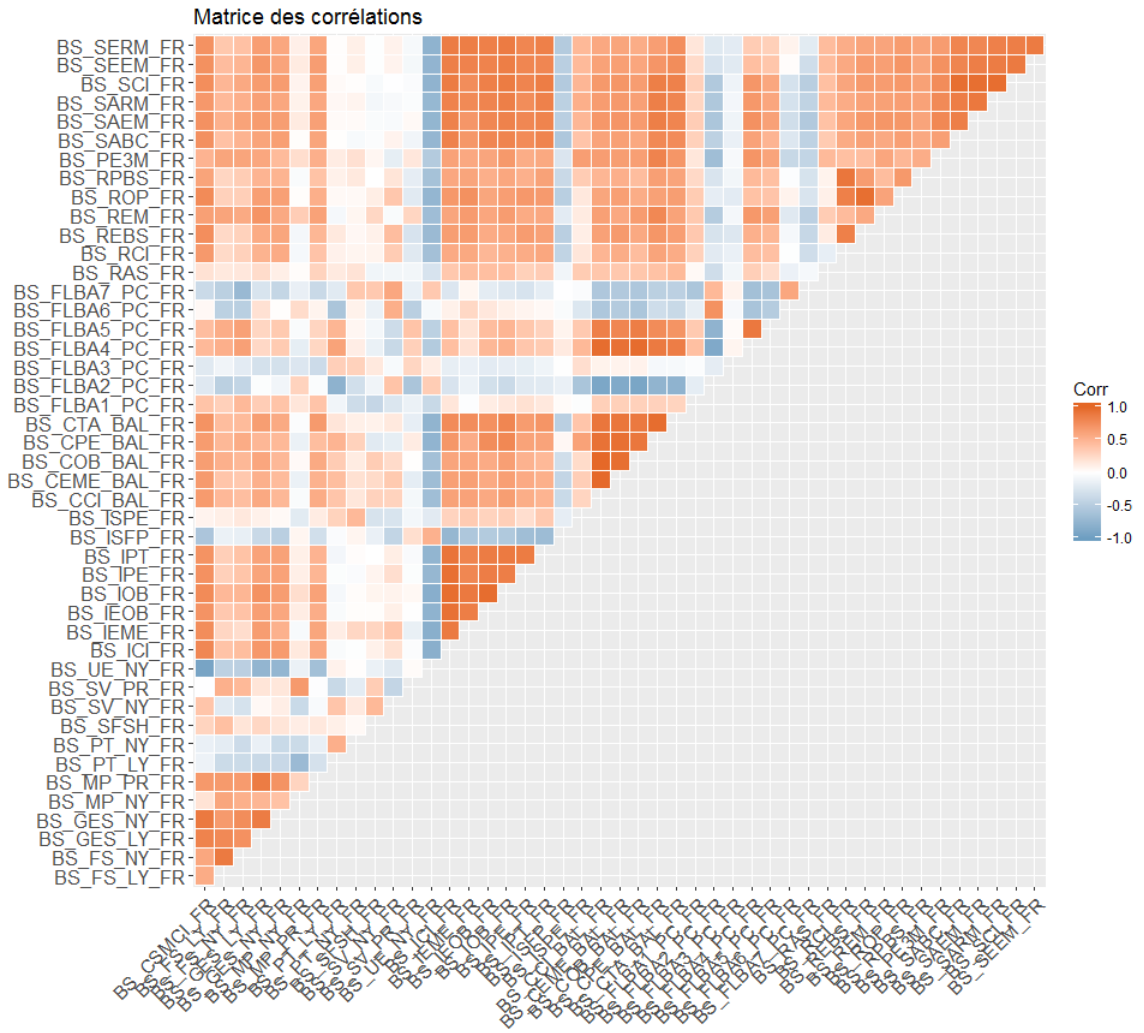


FIGURE 2 – Matrice des corrélations des variables du BCS dans le cadre français
Note de lecture : la matrice des corrélations est ici représentée. Plus une case a une teinte rouge (respectivement bleue), plus forte (respectivement faible) la corrélation entre les deux variables. Les cases blanches représentent à l'inverse les corrélations faibles.
Champ : ensemble des séries mensuelles françaises du BCS

4 Comparaisons par simulations du GETS modelling et du LASSO

4.1 Pourquoi sélectionner des variables ? *L'arbitrage biais-variance reproduit*

À première vue, on pourrait penser que plus on a de variables, plus on a d'informations et meilleures sont les prévisions. Cela remettrait en cause toute utilité à la sélection de variables.

27. P. ZHAO, op. cit.

C'est sans compter l'arbitrage biais-variance. L'intuition décrite correspond à l'évolution du biais : plus on complexifie un modèle, plus il vise juste, diminue le biais. Néanmoins, cet ajout de complexité rend le modèle plus difficile à estimer précisément : en termes statistiques, cela augmente sa variance. Ainsi donc, le biais est décroissant de la complexité, et la variance croissante. Lorsqu'ils se croisent, l'erreur quadratique moyenne atteint son minimum : elle décroit dans un premier tend sous l'effet de la décroissance du biais, avant de remonter à cause de la variance qui devient trop importante au regard du biais. Nous avons voulu vérifier que cet arbitrage se pose bien lorsque l'on mesure l'erreur de prévision, capturée par le RMSFE, plutôt que l'erreur quadratique moyenne, et avec les séries issues des enquêtes de conjoncture.

4.1.1 Méthode opératoire

Pour ce faire, au sein de chaque simulation, nous avons généré une variable à prédire Y à partir des variables du BCS. Nous avons limité le nombre de données en n'étudiant les enquêtes BCS qu'à partir de janvier 2015, ce qui fournissait 36 observations par variable, de sorte à rendre plus difficile la convergence des régressions. Le Y est une combinaison linéaire de 10 variables, tirées au hasard parmi le BCS et d'un bruit normal centré et de variance 20. L'équation 5 résume la construction de Y.

$$Y = X^T \beta + \epsilon \tag{5}$$

où :

- les β sont tirés selon une probabilité discrète uniforme entre -5 et 5, 0 exclu,
- les X_i sont tirés selon un tirage aléatoire sans remise parmi les variables mensuelles du BCS français
- les ϵ_i sont tirés par une gaussienne centrée de variance 20 pour créer suffisamment de bruit.

Ensuite, on tire une variable parmi les variables ayant construit Y, on régresse Y sur cette variable et on calcule le RMSFE associé, on fait de même avec deux variables, puis trois et ainsi de suite, jusqu'à avoir calculé le RMSFE avec les 10 variables. La simulation est ensuite répétée 100 fois (100 Y sont donc générés et 1000 RMSFE calculés, puisque 10 RMSFE sont calculés pour chaque Y).

4.1.2 Résultats

Le graphique 3 représente les RMSFE moyens obtenus sur les 100 simulations. Le RMSFE diminue progressivement jusque 5 à 6 variables, avant de remonter, le modèle devenant trop complexe, ce qui conduit à une instabilité de l'étalonnage hors échantillon, et donc à un RMSFE plus élevé.

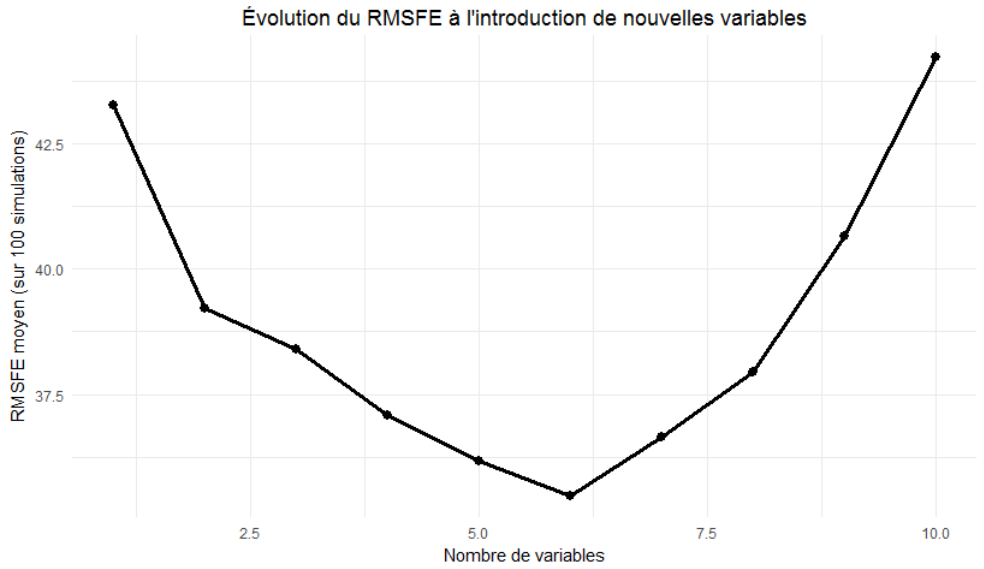


FIGURE 3 – Résultats des simulations de l'arbitrage biais-variance.
Note de lecture : en moyenne sur 100 simulations, le RMSFE des étalonnages à une variable est de 43. L'introduction de nouvelles variables qui ont participé à simuler la variable à régresser diminue d'abord le RMSFE avant de l'augmenter en moyenne.

L'arbitrage biais-variance est donc reproduit sur le RMSFE : pour obtenir de bonnes capacités prédictives, il vaut mieux sélectionner les variables que d'en ajouter, quand bien même on ne rajouterait que des variables qui ont servi à construire la variable à prévoir.

4.1.3 Pertinence de l'arbitrage pour la prévision conjoncturelle.

Il faut toutefois souligner en toute rigueur scientifique que l'évolution du RMSFE à l'introduction de nouvelles variables n'a pas toujours cette forme : nous avons restreint le volume de données à 36 observations pour accélérer le phénomène de remontée du RMSFE, mais lorsque le nombre d'observations est plus élevé, la courbe a tendance à être strictement décroissante, à un rythme de plus en plus ralenti. Néanmoins, les données que nous avons utilisées sont mensuelles, elles sont donc relativement abondantes, alors que dans le cadre de la prévision conjoncturelle, les données sont trimestrielles, ce qui divise le volume de données par 3. Le restriction à 36 points nous a donc semblé plus représentative des conditions concrètes de production des prévisions conjoncturelles, et donc de notre problématique. De plus, dans un cadre concret, les données ajoutées peuvent ne pas être prédictives, ce qui joue à la hausse sur le RMSFE. L'effet de limitation du volume de données et celui de l'ajout potentiel de variables qui ne sont pas prédictives jouant en sens opposés, il semble raisonnable de juger que l'arbitrage biais-variance affecte la prévision conjoncturelle, et donc que la sélection de variables y est nécessaire.

4.2 Description de la simulation

4.2.1 Modélisation du problème

On dispose de $X_1, X_2, ..., X_p$, p séries, dont on juge qu'elles peuvent prévoir une variable Y. Parmi ces p variables, q sont pertinentes et ont effectivement permis de construire Y. On postule sans perte de généralité que ces q variables sont les premières : ainsi, $X_1, X_2, ..., X_q$ sont pertinentes et $X_{(q+1)}, ..., X_p$ ne le sont pas. On sait de plus dans notre cas que ces variables sont fortement corrélées. Dès lors, on a deux objectifs, potentiellement antagonistes :

1. Sélectionner les bonnes variables. Pour cela, il faut d'une part retrouver les q variables pertinentes et d'autre part ne pas sélectionner de variables qui ne le sont pas.
2. Prévoir au mieux Y [28].

On pose P de cardinal p, l'ensemble des variables candidates, S de cardinal s, l'ensemble des variables sélectionnées et Q de cardinal q l'ensemble des variables pertinentes. On introduit ici trois métriques, qui permettent de mesurer le respect des deux conditions posées ci-dessus.

$$TPR = \frac{card(Q \cap S)}{s} \tag{6a}$$

$$FNR = \frac{card(Q \cap \bar{S})}{q} \tag{6b}$$

$$RMSFE = \sqrt{E[(Y_{t+1} - \hat{Y}_{t+1|t})^2]} \tag{6c}$$

où Y_{t+1} est la valeur de Y en t+1

et $\hat{Y}_{t+1|t}$ est la valeur de Y en t+1 prédite avec l'information disponible en t.

Le TPR, pour *True positive Rate*, mesure la proportion de variables pertinentes parmi les sélectionnées et le FNR, pour *False negative Rate*, mesure la proportion de variables pertinentes qui n'ont pas été sélectionnées. Idéalement, on voudrait un TPR à 1 et un FNR à 0. Ces deux mesures sont complémentaires et elles assurent une juste sélection des variables. L'idée d'utiliser ces deux métriques vient de Savin [29]. Le RMSFE quant à lui assure de la qualité prédictive de la sélection.

Différentes méthodes s'offrent à nous pour choisir ces variables. Laquelle choisir ? Quels sont les avantages et inconvénients de chaque méthode ? Pour le savoir, dans la lignée des travaux d'Epprecht, de Tibshirani et de Savin, nous allons procéder à des simulations.

4.2.2 Méthode de la simulation

On construit des variables Y à prévoir selon une méthode très proche de celle de la simulation de l'arbitrage biais-variance (cf. 4.1.1). La seule différence tient au fait que dans l'arbitrage biais-variance, on avait fixé $p = q$, alors qu'ici, puisque l'objectif est de trouver les variables pertinentes, on choisit a priori p et q (où $q < p$). On tire donc p variables parmi les variables du BCS, et parmi ces p variables, on en tire q pour créer Y. On a choisi 10 configurations, en faisant varier :

— p, le nombre de variables candidates,
— q, le nombre de variables pertinentes,
— la variance des ϵ,
— le volume de données disponibles [30].

On trouvera en annexe A2 la liste de ces configurations. Pour chaque configuration, on effectue 100 simulations. Les résultats sont ensuite des moyennes et écart interquartiles par configuration.

4.2.3 Description des 6 méthodes comparées

L'objectif de la simulation étant de comparer le Gets modelling et le Lasso, on met en place ces deux méthodes de sélection et on y ajoute deux méthodes qui servent de points de repère.

28. Cet objectif est potentiellement concurrent du premier dans la mesure où, comme on l'a montré plus haut, l'erreur de prévision peut être minimisée en limitant le nombre de variables sélectionnées en dessous de q. Ceci est confirmé par nos résultats, voir 4.3.1 en page 16

29. Savin, op. cit.

30. Pour diminuer le volume de données, on les rend trimestrielles, en calculant leur moyenne trimestrielle

Gets modelling. On utilise la fonction getsm du package gets notamment développé par Sucarrat à la suite de son article décrit plus haut [31]. On utilise alors deux paramétrages de la fonction, qui correspondent à deux méthodes de sélection :

1. Une méthode s'assurant de l'absence d'hétéroscédasticité des résidus (centrés et réduits). Pour cela, un modèle, pour être retenu doit passer un test Portmanteau modifié de Ljung-Box sur les autocorrélations des résidus jusqu'à l'ordre 4, et un autre sur l'autocorrélation des résidus au carré. On fixe la p-value à 0,05 pour chaque test [32].
2. Une méthode ne s'assurant pas de l'absence d'hétéroscédasticité. Cette méthode s'assure uniquement de la significativité des coefficients estimés à l'aide de tests de Student. Par la suite, cette méthode est abusivement désignée comme "getsm sans test", pour signaler l'absence de tests sur les résidus.

Lasso. On utilise la fonction glmnet du package éponyme, qui a notamment été développé notamment par Tibshirani, l'auteur de l'article fondateur sur le Lasso [33], pour le Lasso et la fonction cv.glmnet pour la validation croisée. Les données étant des séries temporelles, elles ne sont pas identiques et indépendantes, hypothèse utilisée pour la validation croisée standard. On a donc mis en place deux formes de validation croisée pour choisir le paramètre de pénalisation :

1. Une où la validation croisée est classique, désignée comme "lasso (cv idd)" pour lasso à cross-validation sous hypothèse d'expériences identiques et indépendantes. Voir la partie sur le Lasso 2.3 en page 8 pour une description plus précise de cette méthode.
2. L'autre où la validation croisée tient compte de la nature des données, qui sont des séries temporelles, issues de processus dont on ne connait qu'une réalisation, désignée comme "lasso (cv ts)", pour lasso à cross-validation adaptée aux séries temporelles. Pour cela, les observations sont réparties en groupes par fenêtres glissantes (les k premières observations sont dans le premier groupe, les k suivantes dans le deuxième etc.), et on effectue ensuite la cross-validation à partir de ces groupes.

Deux OLS comme méthodes de référence. Afin d'avoir des points de référence, nous avons ajouté deux autres méthodes de sélection de variables. La première est un OLS à partir du modèle d'ensemble, avec les p variables candidates. On ne retient alors que les variables significatives au seuil de 5 %. Ce modèle est le plus simple que l'on puisse imaginer, il tient lieux de référence de la "pire" méthode de sélection : si une méthode est moins performante que celle-ci, cela signifie qu'une simple régression OLS est plus efficace. On désigne cette technique comme OLS GUM, pour *General unrestricted model*, appellation dans la littérature sur l'approche *General to specific* du modèle général dont on part. La seconde représente l'autre côté du spectre, il s'agit d'un OLS sur les q variables pertinentes. De la même manière, les variables sélectionnées sont celles qui sont significatives au seuil de 5%. On appelle cette méthode "OLS vrai".

4.3 Résultats : un large succès du GETS modelling

Une fois une méthode fixée, des critères d'évaluation choisis et les éléments à évaluer choisis, il ne reste plus qu'à faire tourner l'algorithme et en analyser les résultats.

4.3.1 Un succès du GETS en moyenne.

Les TPR, FNR et RMSFE moyens sont retranscrits en annexe A3. Dans l'ensemble, le FNR est un critère relativement peu discriminant : il n'est jamais supérieur à 10 %, à l'exception de la

31. G. Sucarrat, op. cit.
32. Pour tenir compte du fait qu'il y ait deux tests, on utilise une correction de Bonferroni, et on divise la p-value d'ensemble, que l'on fixe à 0,1, par 2, pour une p-value par test de 0,05
33. Tibshirani, op. cit.

sélection OLS sur le GUM. Ceci signifie que dans l'ensemble, il est rare qu'une variable pertinente ne soit pas trouvée. On remarque néanmoins que les valeurs les plus élevées (entre 5 et 10 %) sont systématiquement obtenus par l'OLS sur GUM et le Lasso, ainsi ces deux méthodes ont davantage tendance à rejeter des variables à tort.

Le TPR discrimine bien mieux les méthodes de sélection. Ainsi, on trouve comme on s'y attendait que les dix premiers scores sont ceux des OLS sur le bon modèle, avec un TPR à 100 %. Ensuite, le Gets modelling sans test et l'OLS sur le GUM se partagent les résultats suivants. Les TPR inférieurs à 50 % sont tous obtenus soit par lasso (pour l'une ou l'autre méthode de cross-validation) soit par Gets modelling avec test. La méthode la plus apte à retenir les variables pertinentes est le Gets modelling sans test.

Enfin, concernant le RMSFE, les 14 meilleures performances (plus petits RMSFE) ont été obtenues par OLS sur le vrai modèle ou Gets modelling, principalement sans test : pour le choix d'un modèle prédictif, le Gets modelling sans test rivalise avec la sélection par OLS sur le vrai modèle. Les 8 pires prévisions ont été obtenues principalement par OLS sur le GUM, et 2 d'entre elles ont été obtenues par gets modelling avec test. Le décalage entre le TPR et le FNR d'une part et le RMSFE d'autre part s'explique par la forte corrélation entre les variables : ainsi, pour avoir un modèle prédictif, il n'est pas nécessaire de prendre exactement la bonne variable, une variable qui lui est très corrélée convient également.

Ces résultats sont cohérents dans la mesure où l'on retrouve bien que la sélection par OLS sur le véritable modèle est presque systématiquement la meilleure. Le Gets modelling (en particulier sans test) est, dans presque toutes les configurations et au regard des trois critères, meilleur que le Lasso, quel que soit son type de validation croisée. Enfin, au regard des trois critères, la validation croisée sur 10 groupes choisis de façon identique et indépendante et celle sur groupes de fenêtres glissantes semblent équivalentes.

Un résultat peut sembler contre-intuitif : les meilleurs scores en moyenne du Gets modelling sans test par rapport au Gets modelling avec test Portmanteau sur les résidus et les résidus au carré. Bien évidemment, dans l'absolu, l'hétéroscédasticité empêche une juste estimation des paramètres et donc un bon choix de modèle. Néanmoins, dans notre cadre, le nombre d'observations disponibles est assez réduit. On peut donc supposer que les tests Portmanteau ne disposent pas d'assez d'observations pour être fiables, et qu'ils écartent donc des modèles que la procédure getsm sans test accepte et qui sont plus proches du bon modèle. On observe de fait que lorsque le nombre d'observations est encore plus réduit (pour les configurations dans lesquelles la colonne Réduction vaut vrai, dans le tableau A3), l'écart se creuse entre le getsm sans et avec test : par exemple, pour la configuration avec 4 variables pertinentes et 40 candidates, getsm avec test n'a en moyenne que 24 % de variables justes parmi celles qu'il sélectionne, contre 58 % pour getsm sans test, ce qui conduit à une explosion du RMSFE moyen pour cette configuration, qui vaut 164,5 pour getsm avec test, contre 1,7 pour getsm sans test.

4.3.2 Stabilité des méthodes de sélection

L'observation des boîtes du Tuckey des résultats apporte quelque compléments à ces premières conclusions. Les graphiques 4 représentent les boîtes de Tuckey pour deux configurations dont les résultats nous ont semblé particulièrement saillants. Ces graphiques sont également disponibles pour les autres configurations en annexe A1, A2, A4 et A3. Concernant le TPR, le Gets modelling sans test et les deux implémentations du Lasso ont des écarts inter-quartiles généralement assez proches : leurs performances sont du même ordre de stabilité. En revanche, le Gets modelling avec test tend à avoir un écart inter-quartile plus élevé : dans la configuration avec 16 variables pertinentes, 40 candidates et un bruit de 10, alors qu'il a un meilleur résultat médian du point de vue du TPR que les deux méthodes lasso, il est largement moins stable (son écart inter-quartile est environ cinq fois plus large et son premier quartile est ainsi largement inférieur à celui des deux méthodes lasso). Dans cette configuration, on serait tentés de sacrifier un peu d'efficacité contre de la stabilité, et donc de préférer les méthodes lasso au getsm avec test. Du point de

vue du FNR, les résultats ont tendance à être très stables, avec des écarts inter-quartiles très rapprochés (par exemple sur la figure au centre à gauche des graphiques 4), les écarts les plus grands s'observent surtout avec l'OLS sur GUM ainsi que, dans une moindre mesure, avec le Lasso, ce qui renforce les observations sur les moyennes des FNR par configuration. En ce qui concerne l'efficacité de la prévision, mesurée par le RMSFE, elle est également plutôt stable avec les différentes méthodes, quoique là encore l'OLS sur GUM et le Lasso aient tendance à l'être un peu moins.

4.3.3 Un succès à nuancer néanmoins

Getsm sans test l'emporte donc largement du point de vue des trois critères, en moyenne ainsi qu'en écart inter-quartile, suivi du getsm avec test du point de vue de la moyenne, mais sans distinction claire entre le getsm avec test et le lasso lorsque l'on prend en compte la stabilité des méthodes. On soulignera que dans le cadre d'une limitation du volume de données, le getsm avec test est bien moins efficace, en moyenne et en écart inter-quartile pour le TPR et le RMSFE. Pour autant, il convient de nuancer cette conclusion en apportant deux limites à ces simulations.

Premièrement, le Gets modelling a été évalué dans son cadre le plus propice, puisqu'effectivement, le vrai modèle était inclus dans le modèle dont on partait (le GUM), ce qui se reflète dans les résultats pas si mauvais du simple OLS sur le GUM. Dans la pratique, il n'est absolument pas garanti que cela soit le cas, et le Gets modelling risque alors d'être bien moins efficace. En comparaison, le lasso n'a pas besoin de cette hypothèse et il est d'une efficacité plutôt acceptable dans les simulations : s'il conserve la même efficacité lorsque cette hypothèse n'est pas vérifiée alors que le Gets modelling se dégrade, on peut s'attendre à ce qu'il soit relativement aussi, voire plus, efficace que le Gets modelling.

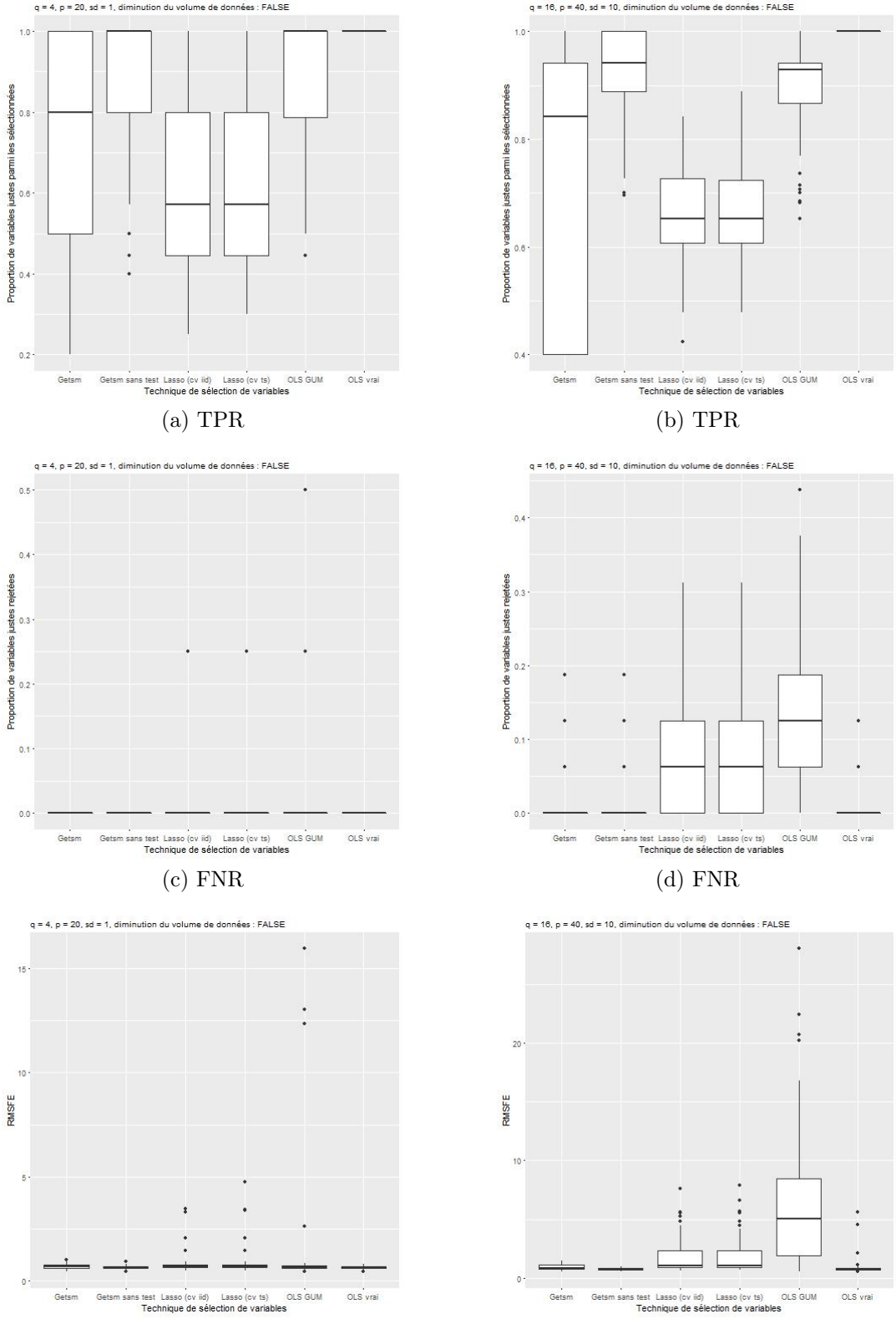
Deuxièmement, le lasso est bien plus rapide que le Gets modelling.

1. Une recherche exhaustive requière 2^p régressions OLS, soit un coût exponentiel.
2. Getsm a un coût quadratique (voir ci-dessous), de p^2 régressions OLS au maximum.
3. Le Lasso a un coût computationnel invariant. Il est équivalent à une régression OLS [34], auquel s'ajoutent autant de régressions que de feuilles pour la cross-validation.

En effet, Getsm requiert en première étape une régression sur le GUM. Ensuite, chaque variable non significative définit un chemin. On note u le nombre de chemins, qui vaut au maximum p. Chaque chemin commence par une estimation du GUM dont on ôte la variable qui définit le chemin (soit u régressions). Au sein de chaque chemin, on effectue des régressions en ôtant la variable la moins significative tant que le modèle passe les tests, soit $p-1$ régressions par chemin au maximum, et donc $u * (p-1)$ régressions au maximum. En sommant le tout, on effectue donc : $1 + u + u(p-1) = 1 + up$ régressions, soit p^2 régressions si $u = p$ dans le pire des cas. De ce point de vue d'ailleurs, getsm sans test étant moins restrictif, il est assurément plus coûteux computationnellement que gestm avec test.

34. Tibshirani, op. cit.

q = 4, p = 20, sd = 1, volume inchangé q = 16, p = 40, sd = 10, volume inchangé



(a) TPR

(b) TPR

(c) FNR

(d) FNR

(e) RMSFE

(f) RMSFE

FIGURE 4 – Boîte de Tuckey des résultats des simulations pour chaque configuration (100 simulations par configuration)
Note de lecture : chaque colonne indique la dispersion des résultats pour chacune des trois métriques (taux de vrais positifs, de faux négatifs et RMSFE) au sein d'une configuration.

5 La sélection de variables en pratique

Les conseils donnés dans la présente partie ont été obtenus après des tests pratiques de sélection de variables sur de vrais agrégats économiques. Ces tests ont été effectués à l'aide d'une fonction implémentée dans pRev, la librairie R de la Synthèse conjoncturelle.

5.1 Les séries en entrée

On suppose les séries en entrée correctement préparées, désaisonnalisées et stationnarisées. Il convient ensuite d'observer d'une part le début des séries candidates et d'autre part leur fin. En effet, il faut garder à l'esprit que la variable qui commence le plus tard est limitante, que sa date de début limitera la date de début de toutes les variables candidates. Il faut de plus bien évidemment que la date de fin de toutes les séries soit postérieure à celle de l'agrégat à prévoir. Dans le cas où une série finirait trop tard, on peut envisager, si cela fait sens économiquement, de la retarder.

Il peut arriver que le nombre de séries candidates soit supérieur au nombre de périodes d'observation. En ce cas, le Gets modelling, puisqu'il repose sur des régressions linéaires, ne fonctionne pas. En revanche, le Lasso accepte une telle configuration.

Cette première phase est à effectuer manuellement, en gardant toujours à l'esprit la signification économique des étalonnages rendus possibles par le choix des séries candidates : toute série ou combinaison linéaire de séries peut être retenue par le Gets modelling ou le Lasso.

5.2 La sélection : gets modelling ou lasso ?

En pratique, l'hypothèse que les variables à l'origine de l'agrégat à prévoir sont toutes inclues dans l'ensemble des variables candidates est bien souvent difficile à justifier. En conséquence, le Gets modelling et le Lasso n'ont pas toujours les mêmes performances. La fonction implémentée dans pRev sélectionne donc plusieurs étalonnages, un par Gets modelling et un par Lasso, et les évalue en fonction de leur R^2 et de leur RMSFE.

5.3 En sortie : interpréter les résultats et choisir le bon étalonnage

Le Lasso et le Gets modelling (avec ou sans test) sélectionnent deux ou trois étalonnages, qui chacun au regard d'un certain critère sont parmi les meilleurs de l'ensemble des 2^p étalonnages possibles. Néanmoins, il reste ensuite au prévisionniste à choisir parmi ces deux ou trois modèles en ayant deux éléments potentiellement antagonistes en tête :

- La performance pour la prévision. En théorie, plus le R^2 est élevé et plus le RMSFE est faible, meilleur est l'étalonnage. Il faut toutefois garder à l'esprit qu'un étalonnage trop efficace peut être le signe d'un suraprentissage.
- L'interprétabilité. Le modèle doit faire sens économiquement : les coefficients doivent être du bon signe, y compris au sein de combinaisons linéaires (voir plus haut 1.3 en page 4).

Conclusion

Gets modelling ou Lasso ? Cela dépend de la situation... En présence de variables hautement corrélées, le Lasso perd son efficacité. Il n'en est pas moins la seule possibilité lorsque l'on dispose de moins d'observations que l'on a de variables candidates. Réciproquement, le Gets modelling, et particulièrement sa version sans test sur l'hétéroscédasticité des résidus, obtient de bien meilleures performances lors de simulations, que ce soit pour retenir les bonnes variables (TPR), ne pas écarter de variable pertinente (FNR) et fournir un étalonnage prédictif (RMSFE),

mais ses performances sont souvent comparables à celle du Lasso lorsque l'ensemble des variables pertinentes n'est pas inclus dans les variables candidates.

En pratique, il est souvent plus prudent d'effectuer plusieurs sélections, en ayant recours à chacune de ces méthodes, pour écarter les étalonnages moins bons et ensuite, à l'aide d'une analyse économique et statistique, sélectionner l'étalonnage qui semble conjuguer le mieux performance de prévision et interprétabilité. L'automatisation permet certes d'accélérer la recherche, mais la décision humaine reste encore au coeur de la sélection d'étalonnages pour la prévision économique.

Références

D. Hendry, H. Krolzig. « Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez ». In : *Econometrics Journal* 2.2 (1999), p. 202–219.

E. Dubois, E. Michaux. « Etalonnages à l'aide d'enquêtes de conjoncture : de nouveaux résultats ». In : *Economie et prévision* 172 (2006), p. 11–28.

Epprecht, Camila et al. *Variable selection and forecasting via automated methods for linear models : LASSO/adaLASSO and Autometrics*. Documents de travail du Centre d'Economie de la Sorbonne. Université Panthéon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne, nov. 2013.

G. Sucarrat, A. Escribano. « Automated Financial Model Selection : General-to- Specific Modelling of the Mean and Volatility Specifications ». In : *Oxford Bulletin of Economics and Statistics* (2012), p. 716–735.

H. Krolzig, D. Hendry. « Computer Automation of General-to-Specific Model Selection Procedures ». In : *Journal of Economic Dynamics and Control* (2001).

K. Hoover, S. Perez. « Data mining reconsidered : encompassing and the general-to-specific approach to specification search ». In : *Econometrics Journal* 2 (1999), p. 1–25.

Lovell, Michael C. « Data Mining ». In : *The Review of Economics and Statistics* 65.1 (1983), p. 1–12.

N. Hsu H. Hung, C. Ya-Mei. « Subset Selection for Vector Autoregressive Processes Using Lasso ». In : 52 (mar. 2008), p. 3645–3657.

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

Pagan, A. « Three Econometric Methodologies : A Critical Appraisal ». In : *Journal of Economic Surveys* 1.1 (1987), p. 3–24.

S. Grégoir, E. Le Rey. « La pratique des étalonnages dans l'analyse conjoncturelle ». In : *INSEE Méthodes* 33 (1995), p. 52–98.

Savin, I. *A Comparative Study of the Lasso-Type and Heuristic Model Selection Methods*. Working Paper 04. COMISEF, 2010.

The joint harmonised EU programme of business and consumer surveys. User Guide. European Commission, 2016. URL : `https://ec.europa.eu/info/files/user-guide-joint-harmonised-eu-programme-business-and-consumer-surveys_en`.

Tibshirani, B. Efron T. Hastie I. Johnstone R. « Least angle regression ». In : *Ann. Statist.* 32.2 (avr. 2004), p. 407–499.

Tibshirani, R. « Regression Shrinkage and Selection via the Lasso ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), p. 267–288.

TABLE A1 – Effectifs par enquêtes au sein du BCS français
Note de lecture : en France, l'enquête industrie du BCS a un effectif de 4000 entreprises.

	Industrie	Services	Consommateurs	Commerce de détail	Construction
France	4000	4500	3300	3750	2500

Annexes

TABLE A2 – Liste des configurations testées pour la simulation
Note de lecture : chaque ligne décrit une configuration. Ainsi, dans la première configuration, la variable à prédire a été construite à partir de 4 variables et d'un bruit centré et d'écart-type 1, le volume de données a été réduit et il y avait 20 variables candidates (parmi lesquels il y avait les 4 variables pertinentes.

sd	p	q	Diminution du volume de données
1	20	4	VRAI
1	40	4	VRAI
1	20	4	FAUX
1	20	8	FAUX
10	20	4	FAUX
10	20	8	FAUX
1	40	4	FAUX
1	40	16	FAUX
10	40	4	FAUX
10	40	16	FAUX

TABLE A3 – Résultats moyens de chaque méthode de sélection pour les 10 configurations
Note de lecture : dans la première configuration (sd=1, p=20, q=4 et pas de réduction du volume de données), 73 % des variables gardées l'étaient à raison, aucune variable pertinente n'a été omise et l'erreur moyenne de prévision est de 0,7.

Méthode	sd	p	q	Réduction	TPR	FNR	RMSFE
Getsm	1	20	4	FAUX	73 %	0 %	0,7
Getsm sans test	1	20	4	FAUX	88 %	0 %	0,6
Lasso (cv iid)	1	20	4	FAUX	60 %	2 %	0,8
Lasso (cv ts)	1	20	4	FAUX	60 %	2 %	0,8
OLS GUM	1	20	4	FAUX	86 %	1 %	1,1
OLS vrai	1	20	4	FAUX	100 %	0 %	0,6
Getsm	10	20	4	FAUX	84 %	0 %	0,7
Getsm sans test	10	20	4	FAUX	91 %	0 %	0,7
Lasso (cv iid)	10	20	4	FAUX	60 %	2 %	0,7
Lasso (cv ts)	10	20	4	FAUX	61 %	2 %	0,8
OLS GUM	10	20	4	FAUX	86 %	1 %	0,9
OLS vrai	10	20	4	FAUX	100 %	0 %	0,7
Getsm	1	40	4	FAUX	46 %	1 %	0,9
Getsm sans test	1	40	4	FAUX	73 %	2 %	0,7
Lasso (cv iid)	1	40	4	FAUX	50 %	2 %	0,8
Lasso (cv ts)	1	40	4	FAUX	51 %	2 %	0,8
OLS GUM	1	40	4	FAUX	72 %	11 %	3,2
OLS vrai	1	40	4	FAUX	100 %	0 %	0,7
Getsm	10	40	4	FAUX	44 %	1 %	0,9
Getsm sans test	10	40	4	FAUX	71 %	1 %	0,7
Lasso (cv iid)	10	40	4	FAUX	46 %	3 %	0,8
Lasso (cv ts)	10	40	4	FAUX	47 %	3 %	0,8
OLS GUM	10	40	4	FAUX	69 %	15 %	5,4
OLS vrai	10	40	4	FAUX	100 %	0 %	0,7
Getsm	1	20	8	FAUX	87 %	1 %	0,7
Getsm sans test	1	20	8	FAUX	95 %	1 %	0,7
Lasso (cv iid)	1	20	8	FAUX	72 %	3 %	0,9
Lasso (cv ts)	1	20	8	FAUX	72 %	3 %	1,0
OLS GUM	1	20	8	FAUX	94 %	1 %	1,1
OLS vrai	1	20	8	FAUX	100 %	0 %	0,7

TABLE A4 – Résultats moyens de chaque méthode de sélection pour les 10 configurations (suite)

Méthode	sd	p	q	Réduction	TPR	FNR	RMSFE
Getsm	10	20	8	FAUX	80 %	0 %	0,7
Getsm sans test	10	20	8	FAUX	94 %	0 %	0,7
Lasso (cv iid)	10	20	8	FAUX	73 %	3 %	1,0
Lasso (cv ts)	10	20	8	FAUX	73 %	4 %	1,0
OLS GUM	10	20	8	FAUX	93 %	2 %	1,4
OLS vrai	10	20	8	FAUX	100 %	0 %	0,7
Getsm	1	40	16	FAUX	70 %	2 %	1,0
Getsm sans test	1	40	16	FAUX	92 %	2 %	0,8
Lasso (cv iid)	1	40	16	FAUX	69 %	7 %	1,6
Lasso (cv ts)	1	40	16	FAUX	69 %	7 %	1,7
OLS GUM	1	40	16	FAUX	91 %	12 %	6,1
OLS vrai	1	40	16	FAUX	100 %	1 %	0,8
Getsm	10	40	16	FAUX	71 %	2 %	0,9
Getsm sans test	10	40	16	FAUX	93 %	2 %	0,8
Lasso (cv iid)	10	40	16	FAUX	66 %	8 %	1,8
Lasso (cv ts)	10	40	16	FAUX	66 %	8 %	1,8
OLS GUM	10	40	16	FAUX	90 %	14 %	6,0
OLS vrai	10	40	16	FAUX	100 %	1 %	0,9
Getsm	1	20	4	VRAI	59 %	0 %	7,4
Getsm sans test	1	20	4	VRAI	85 %	0 %	1,1
Lasso (cv iid)	1	20	4	VRAI	56 %	2 %	1,5
OLS GUM	1	20	4	VRAI	88 %	2 %	2,2
OLS vrai	1	20	4	VRAI	100 %	0 %	1,1
Getsm	1	40	4	VRAI	24 %	4 %	164,5
Getsm sans test	1	40	4	VRAI	58 %	5 %	1,7
Lasso (cv iid)	1	40	4	VRAI	45 %	3 %	1,7
OLS GUM	1	40	4	VRAI	68 %	31 %	20,3
OLS vrai	1	40	4	VRAI	100 %	1 %	1,1

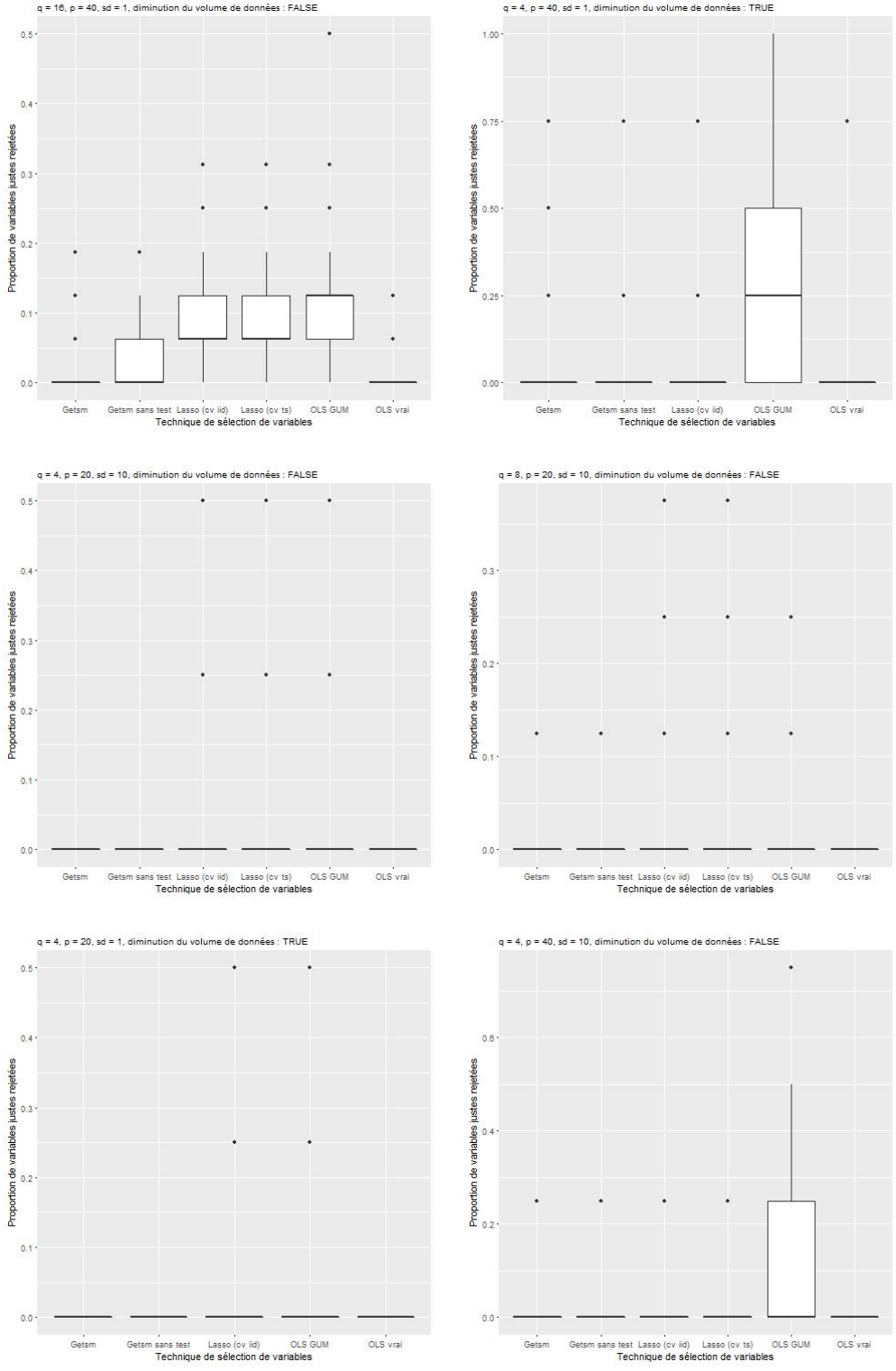


FIGURE A1 – Boîte de Tuckey des FNR des simulations pour chaque configuration (100 simulations par configuration)
getsm : GETS modelling, lasso_cv_iid : lasso par cross-validation classique en 10 groupes, lasso_cv_ts : lasso par cross-validation adaptée aux séries temporelles, ols_gum : régression OLS sur le GUM, ols_true : régression sur le vrai modèle

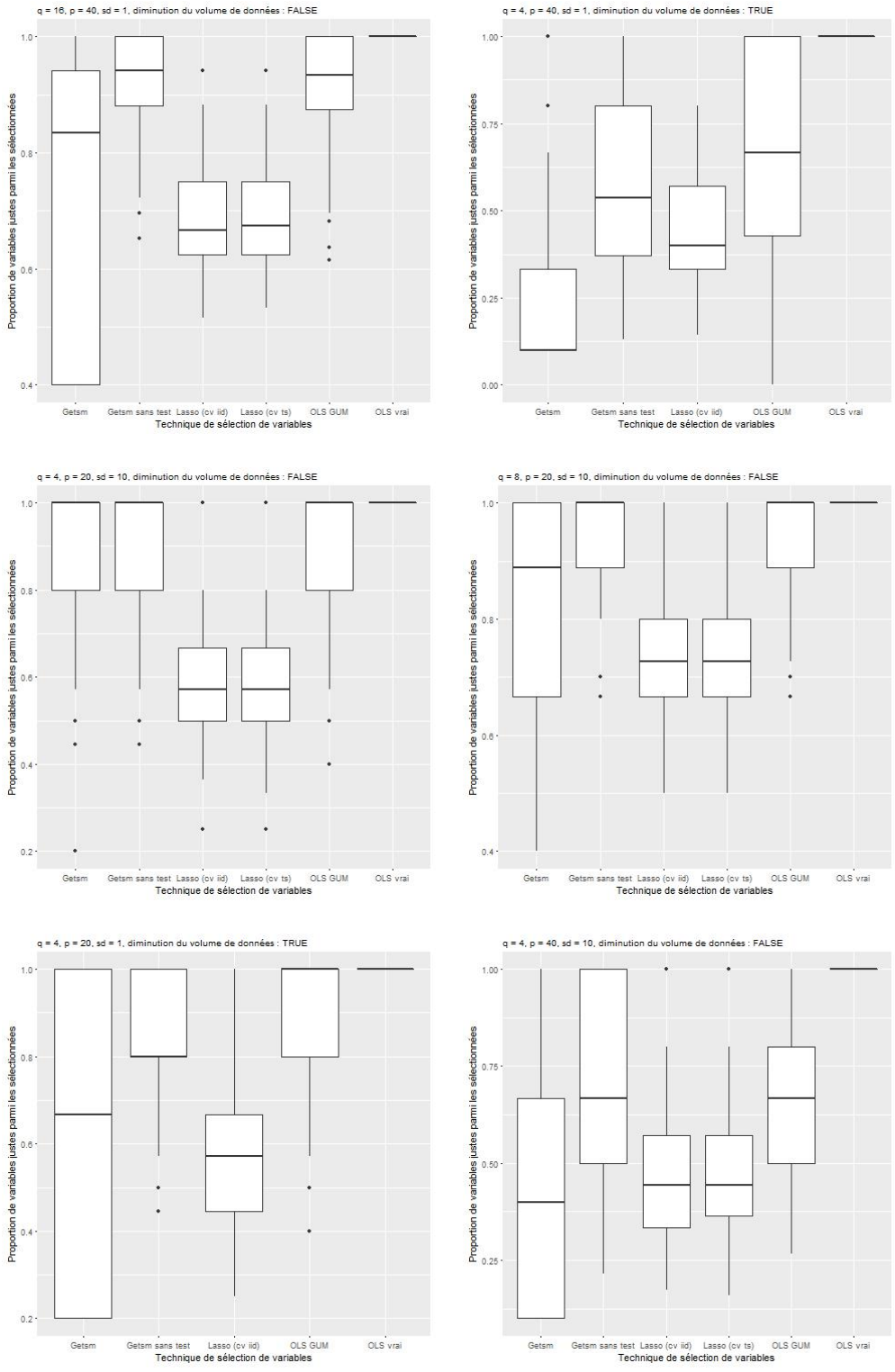


FIGURE A2 – Boîte de Tuckey des TPR des simulations pour chaque configuration (100 simulations par configuration)
getsm : GETS modelling, lasso_cv_iid : lasso par cross-validation classique en 10 groupes, lasso_cv_ts : lasso par cross-validation adaptée aux séries temporelles, ols_gum : régression OLS sur le GUM, ols_true : régression sur le vrai modèle

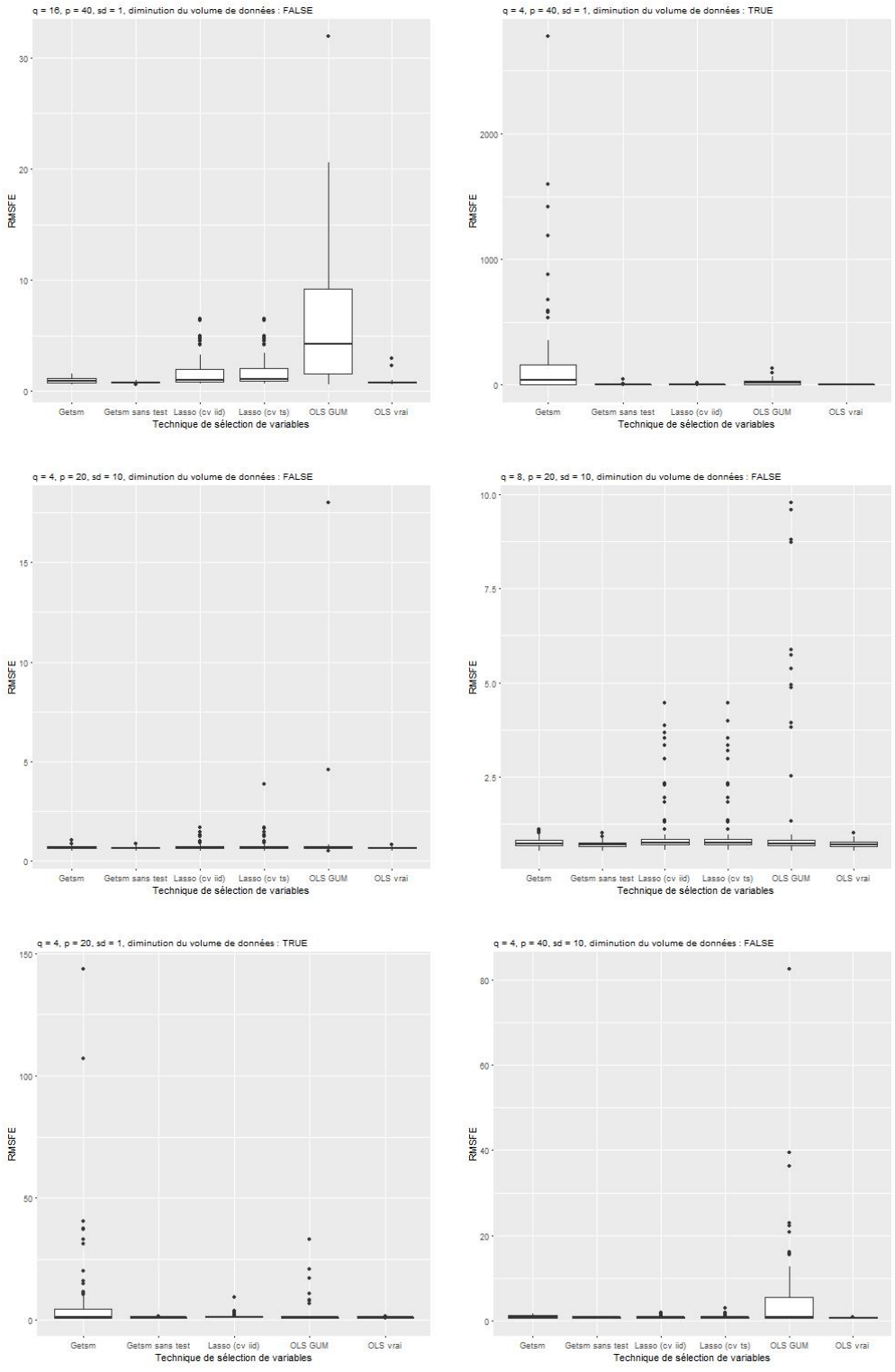


FIGURE A3 – Boîte de Tuckey des RMSFE des simulations pour chaque configuration (100 simulations par configuration)
getsm : GETS modelling, lasso_cv_iid : lasso par cross-validation classique en 10 groupes, lasso_cv_ts : lasso par cross-validation adaptée aux séries temporelles, ols_gum : régression OLS sur le GUM, ols_true : régression sur le vrai modèle

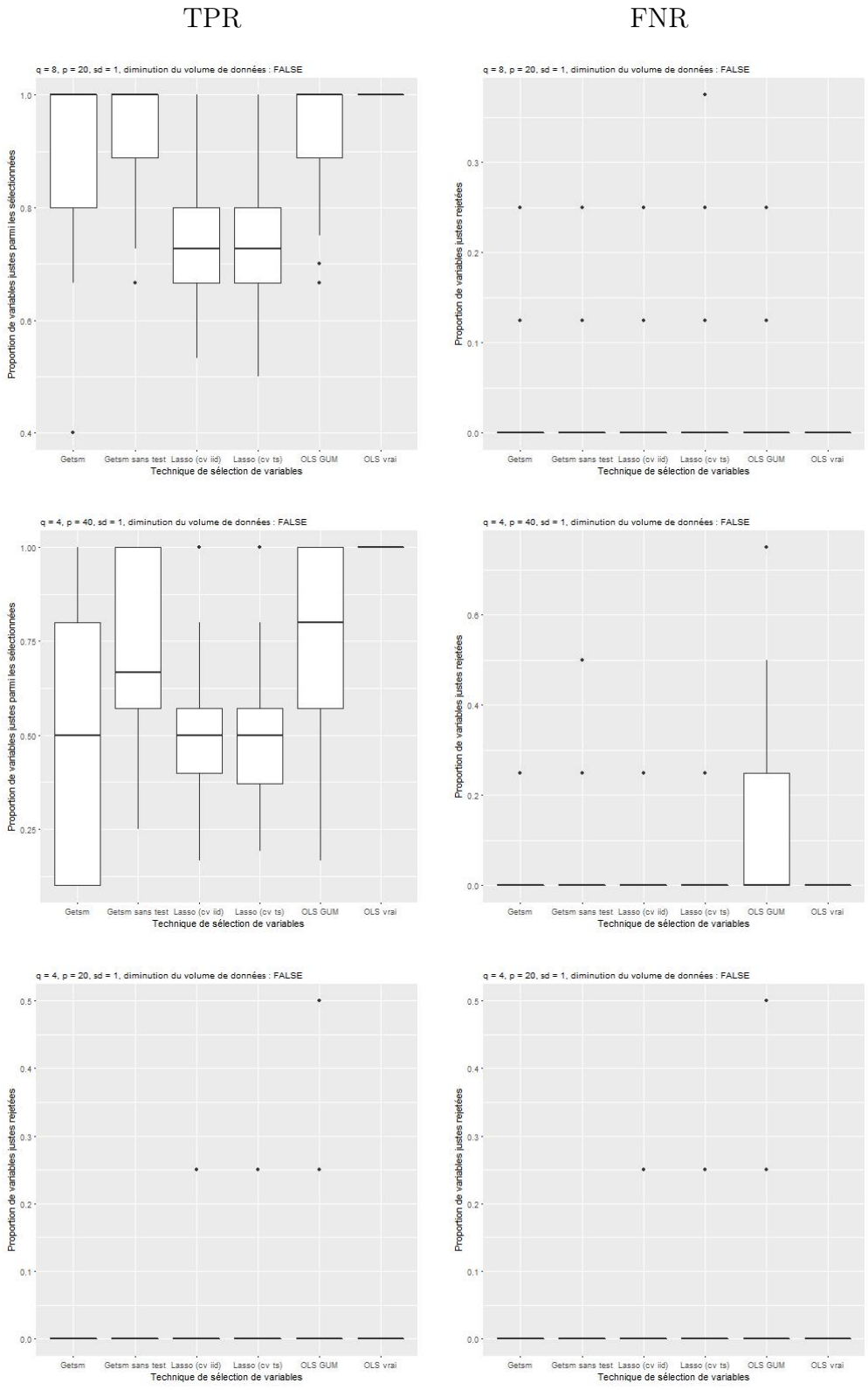


FIGURE A4 – Boîte de Tuckey des résultats des simulations pour chaque configuration (100 simulations par configuration)
getsm : GETS modelling, lasso_cv_iid : lasso par cross-validation classique en 10 groupes, lasso_cv_ts : lasso par cross-validation adaptée aux séries temporelles, ols_gum : régression OLS sur le GUM, ols_true : régression sur le vrai modèle

Table des matières