Documents de travail

N° M2020/01
Octobre 2020

L'économétrie en grande dimension

Jérémy L'HOUR (Insee, Crest)

Insee
Mesurer pour comprendre

L'économétrie en grande dimension

Jérémy L'Hour
Insee et CREST

13 octobre 2020

Résumé

Ce document de travail est une courte introduction aux principaux problèmes que l'on rencontre lorsque l'on souhaite faire de l'économétrie en grande dimension, c'est-à-dire lorsque $p > n$ – pour chaque observation, on dispose d'un nombre de caractéristiques potentiellement proportionnel ou plus grand que la taille de l'échantillon. La première partie présente les solutions standards de régression pénalisée (Lasso et Ridge). La seconde partie illustre et traite du problème de l'inférence post-sélection et du biais de régularisation. La dernière partie traite de la détection des effets hétérogènes dans les expériences aléatoires au moyen d'algorithmes de machine learning. Des illustrations en code `R` sont disponibles sur le répertoire GitHub associé : https://github.com/InseeFrLab/grandedim.

Mots-clés : économétrie, statistique en grande dimension, sélection de variables, effets hétérogènes, apprentissage automatique
Classification JEL : C01, C52

Jérémy L'Hour, SSP Lab, DMCSI, Insee et CREST, ENSAE Paris. 88, avenue Verdier, CS 70058 92541 Montrouge Cedex, (+33) 1 87 69 55 75, jeremy.lhour@insee.fr.

Table des matières

1. Introduction et exemples

On qualifie de *grande dimension* les situations dans lesquelles on souhaite estimer un grand nombre de paramètres relativement à la taille de l'échantillon disponible et où les méthodes économétriques traditionnelles sont mises en défaut. Ce contexte doit donc ne pas être confondu avec ceux qui requièrent le traitement de données massives ("big data"). La grande dimension apparaît, de plus en plus fréquemment, dans une large variété de domaines :

1. L'analyse des déterminants de la croissance économique. Par exemple, Sala-I-Martin (1997) cherche à sélectionner de façon robuste les facteurs de la croissance économique en estimant les 30,856 régressions à sept variables avec toutes les combinaisons possibles de trois variables parmi 58, soit $\binom{58}{3} = 30856$, quatre étant fixées a priori.

2. La prédiction à partir d'un grand nombre de variables ou bien de transformations polynomiales des régresseurs. C'est rapidement le cas dès que l'on parle de données macroéconomiques d'autant plus que le nombre d'observations se trouve souvent limité, voir par exemple Stock and Watson (2012). Par exemple, Ferrara and Simoni (2019) utilisent la régression Ridge (cf. infra Section 2.2) pour estimer une équation de prévision du taux de croissance du PIB avec un grand nombre de variables provenant à la fois de sources de comptabilité nationale, de sondages d'opinion et de données mesurant les recherches Google.

3. L'estimation d'effets de traitement pour une multitude de traitements ou sur un grand nombre de sous-groupes – ou tout simplement, l'estimation de l'effet d'un grand nombre de variables explicatives. Par exemple, Abrams et al. (2012) cherchent à estimer l'effet de l'origine ethnique de l'accusé sur la sentence donnée par le juge, et ce pour un grand nombre de juges ; Chetty and Hendren (2018) étudient l'impact de la localisation géographique sur la mobilité sociale aux États-Unis, pour un grand nombre de localisations. Voir Abadie and Kasy (2018) pour d'autres exemples.

4. L'analyse textuelle et en particulier la classification de documents. Cette analyse se base généralement sur la constitution d'un dictionnaire établi à partir des mots retrouvés dans les documents inclus dans la base de données. Ce dictionnaire per-

met alors de construire la matrice (généralement de très grande dimension) terme-document qui compte, par document, le nombre de fois où un mot du dictionnaire est utilisé. Cette matrice est ensuite utilisée directement ou indirectement pour construire des variables explicatives puisque l'utilisation d'un terme plutôt que d'un autre reflète un document de nature particulière. Il y a donc autant de variables que de mots dans le dictionnaire.

5. La construction de scores polygéniques. Les scores polygéniques sont des mesures individuelles de la propension génétique à développer certaines maladies, caractéristiques physiques ou traits de personnalité. Ils sont constitués à partir de la mesure de Polymorphismes d'un Seul Nucléotide (PSN, ou SNP en Anglais), c'est-à-dire la variation d'une seule paire de bases du génome entre individus d'une même espèce. Ils sont fréquents, à hauteur d'une paire de bases sur mille dans le génome humain, ce qui constitue autant de potentielles variables explicatives pour établir un score. Les exemples d'applications, y compris en science sociale, sont extrêmement nombreux, voir par exemple Plomin (2018).

Ces exemples requièrent ou, dans tous les cas, peuvent tirer parti d'une approche qui pénalise la complexité du modèle. Cette pénalisation peut se faire de plusieurs façons, et il existe une méthodologie pour sélectionner le bon niveau de pénalisation de sorte par exemple, à minimiser l'erreur de prédiction sur un nouvel échantillon. C'est ce que nous verrons dans la section 2.

En micro-économétrie, la plupart des applications empiriques cherchent à établir un lien causal clair entre deux phénomènes bien définis, du type : "Est-ce que A cause B ?". Dans ce contexte, le problème de la grande dimension est moins immédiat, ou du moins, peut être surmonté grâce à des solutions ad-hoc. Il s'agit par exemple du choix des variables de contrôle, de transformations de celles-ci, ou bien de celui des instruments. On verra toutefois que même dans ces cas, les propriétés classiques des estimateurs résultant d'une étape préalable de sélection de variables telles que la convergence ou la normalité asymptotique ne sont pas garanties, ce qui peut se traduire par des résultats de test biaisés ou des intervalles de confiance ayant un taux de couverture moindre que le taux théorique. C'est ce que nous verrons dans la section 3.

Enfin, une tâche standard est la détection d'effets hétérogènes, c'est-à-dire de la mo-

dulation de l'intensité d'un lien causal par des facteurs confondants. Cette tâche répond à plusieurs objectifs : d'une part, on peut souhaiter évaluer l'efficacité d'un traitement ou d'une politique pour des sous-populations distinctes afin de décider quels individus traiter, d'autre part, cela permet d'établir la validité externe des résultats en évaluant l'impact d'une politique sur une autre population, différente de celle ayant servie à l'expérimentation. Cette tâche est également de grande dimension dans la mesure où on peut vouloir considérer des sous-populations générées par un croisement de multiples caractéristiques. Le danger est alors de faire des découvertes fallacieuses concernant les dimensions selon lesquelles l'effet du traitement varie, en se lançant dans de la recherche de spécification (ou *pêche aux p-values*). Une solution peut être alors d'utiliser des algorithmes de machine learning (ML) afin d'étudier de façon automatique l'hétérogénéité d'un traitement. C'est ce que nous verrons dans la section 4.

Comme les exemples évoqués plus haut le montrent, le terme de *grande dimension* recouvre des contextes et des tâches statistiques divers. Ce document ne peut pas tous les couvrir. En particulier, on se concentre principalement sur un contexte microéconomique standard où les observations sont indépendantes et identiquement distribuées, bien que la section 2 puisse également s'appliquer à des données de séries chronologiques. En outre, la section 3 discute essentiellement l'estimation d'un paramètre d'intérêt qui est de petite dimension et n'évoque pas les sujets tels que le paradoxe de James-Stein et les tests multiples. Pour cela, Belloni et al. (2018); Abadie and Kasy (2018) constituent d'excellentes références. Enfin, pour certains exemples évoqués, le traitement de la grande dimension est très spécifique et ne peut pas être raisonnablement exposé ici : c'est le cas de l'analyse textuelle (Jurafsky and Martin, 2019) ou de l'analyse des réseaux (de Paula, 2015 ; Chandrasekhar, 2016 ; Graham, 2019).

Ce document est partiellement basé sur le cours de troisième année de l'ENSAE Paris "Machine Learning for Econometrics" disponible en ligne (Gaillac and L'Hour, 2019), notamment pour les sections 3 et 4. Belloni et al. (2018) consitute une référence complémentaire. Une bonne référence statistique pour la section 2 est (Hastie et al., 2009, Chapitres 3 et 18).

Mise en pratique 0: Notebooks

Des notebooks `R` sont disponibles dans le dossier GitHub correspondant : https://github.com/InseeFrLab/grandedim. Des encadrés faisant référence à ces notebooks sont insérés dans le texte pour vous aider à mettre en pratique les méthodes présentées.

La section 2 passe en revue les différents types de pénalisation couramment utilisés, en s'intéressant en particulier au cas de la régression linéaire. Cette section est générale et peut également s'appliquer pour des tâches de prédiction avec des séries temporelles. La section 3 étudie le problème de l'inférence après une première étape de sélection de variables. Elle est plus spécifique aux questions standards de micro-économétrie. La section 4 montre comment on peut étudier l'hétérogénéité des effets de traitement dans les expériences aléatoires à partir de proxies construits grâce à des algorithmes de machine learning.

2. La régression pénalisée

Dans cette partie, nous allons étudier les alternatives à l'estimateur des Moindres Carrés Ordinaires (MCO) dans le cas de la grande dimension.

2.1. La régression linéaire mise en défaut

Prenons le cadre de régression linéaire simple. On suppose que l'on observe la suite de données iid $(Y_i, X_i)_{i=1,\dots,n}$ où $Y_i = X_i'\beta_0 + \varepsilon_i$, et $X_i \perp\!\!\!\perp \varepsilon_i$. On suppose que X_i est un vecteur aléatoire de dimension p avec $p < n$ – on est donc dans un cas de petite dimension – et on note $X_{i,j}$ la j-ième composante de X_i. L'estimateur des Moindres Carrés Ordinaires (MCO) de β_0 est donné par :

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i'\beta\right)^2 . \tag{MCO}$$

Ce programme admet une solution analytique unique sous la forme :

$$\widehat{\beta} = \left[\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i,$$

à condition que la matrice $\sum_{i=1}^{n} X_i X_i'/n$ (la matrice de Gram) soit inversible ce qui suppose en particulier, que les colonnes de la matrice de dimension $n \times p$ $(X_i')_{i=1,\dots,n}$ soient linéairement indépendantes.

Dans le contexte précédent, on entend par *grande dimension*, le fait d'avoir à disposition un grand nombre de régresseurs, *i.e.* $p > n$ ou bien, simplement, p proportionnel à n. Deux problèmes apparaissent alors : (i) la précision de l'estimateur (MCO) se dégrade (plus forte variance) à cause de la multicolinéarité, et même (ii) impossibilité de le calculer (si la matrice de Gram, $\sum_{i=1}^{n} X_i X_i'/n$, n'est plus inversible). La statistique dite "en grande dimension" a développé tout un lot de techniques répondant à ce problème, et c'est ce que nous allons voir dans ce document.

2.2. Régression Ridge

Pour un niveau de pénalité $\lambda \geq 0$, on définit l'estimateur (RIDGE) comme solution du programme de minimisation :

$$\widehat{\beta}^R(\lambda) = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i'\beta\right)^2 + \lambda \left\|\beta\right\|_2^2 , \tag{RIDGE}$$

où $\|\beta\|_2 = \sqrt{\sum_{j=1}^p \beta_j^2}$. Après calcul, on trouve :

$$\widehat{\beta}^R(\lambda) = \left[\frac{1}{n}\sum_{i=1}^n X_i X_i' + \lambda \mathbb{I}_p\right]^{-1} \frac{1}{n}\sum_{i=1}^n X_i Y_i,$$

où $\mathbb{I}_p$ désigne la matrice identité de dimension p. À travers la solution analytique, on peut voir que la régression Ridge résout le problème de la grande dimension en ajoutant le terme $\lambda\mathbb{I}_p$ afin de rendre inversible la matrice $\sum_{i=1}^n X_i X_i'/n + \lambda\mathbb{I}_p$ quand $\lambda > 0$ même si $\sum_{i=1}^n X_i X_i'/n$ n'est pas de plein-rang. Plus le coefficient λ est grand et plus l'estimateur (RIDGE) va réduire l'estimateur (MCO) vers le vecteur nul. Si $\lambda = 0$, on retrouve la solution (MCO). Inversement, si $\lambda \to \infty$, $\widehat{\beta}^R(\lambda) \to 0$.

De façon pratique, il est important de noter que la solution est sensible à l'échelle des régresseurs. Il est donc préférable de les normaliser, par exemple en les divisant par leur écart-type ou en les ramenant à l'intervalle $[0, 1]$ par la transformation $x \to (x - \min(x))/(\max(x) - \min(x))$. Il est également possible de pénaliser chaque élément de β de façon différente ou de ne pas tous les pénaliser. Typiquement, si X_i contient une constante en tant que premier élément, on ne pénalise généralement pas le coefficient associé, de sorte que l'on résout le programme

$$\min_\beta \frac{1}{n}\sum_{i=1}^n \left(Y_i - \beta_1 - X_{i,-1}'\beta_{-1}\right)^2 + \lambda\sum_{j=2}^p \beta_j^2,$$

où pour un vecteur x de taille p la notation x_{-1} désigne le vecteur x privé de sa première composante.

Mise en pratique 1: Ridge avec `glmnet`

Le notebook `RidgeLasso-glmnet` met en pratique la régression Ridge avec le package `glmnet` à partir de données simulées.

2.3. Régression Lasso

Pour un niveau de pénalité $\lambda \geq 0$, on définit l'estimateur (LASSO) comme solution du programme de minimisation :

$$\widehat{\beta}^L(\lambda) \in \arg\min_\beta \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\beta)^2 + \lambda \|\beta\|_1, \qquad \text{(LASSO)}$$

où $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. De même que dans le cas de l'estimateur Ridge, λ définit le niveau de pénalisation. Notons que $\widehat{\beta}^L(\lambda)$ peut ne pas être unique et que (LASSO) ne possède pas de solution analytique en général. En revanche, à λ fixé, la prédiction $X_i'\widehat{\beta}^L(\lambda)$ est unique pour chaque $i = 1, \ldots, n$. Il existe des algorithmes performants pour résoudre ce programme d'optimisation, dont par exemple le Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) de Beck and Teboulle (2014). Du fait de la non différentiabilité en zéro de la pénalisation par la norme ℓ_1, la solution obtenue est très souvent *parcimonieuse* ou *sparse* au sens où un certain nombre d'éléments de $\widehat{\beta}^L(\lambda)$ vont être exactement égaux à zéro. Cette propriété fait du Lasso un outil régulièrement utilisé pour faire de la sélection de variables (cf. infra, Section 2.4).

Quelles sont les garanties théoriques de l'estimateur Lasso en terme d'estimation et de prédiction ?

Estimation. Rappelons que dans le cas de l'estimateur (MCO), on peut montrer que sous des hypothèses de régularité et si p est fixe :

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}_p\left(0, \mathbb{E}[XX']^{-1}\mathbb{E}[\varepsilon^2 XX']\mathbb{E}[XX']^{-1}\right), \text{ lorsque } n \to \infty.$$

La grande dimension impose de considérer un cadre asymptotique où $p \to \infty$ lorsque $n \to \infty$, sinon le problème est de petite dimension dans l'asymptotique – si p est fixe alors systématiquement n le dépasse lorsque $n \to \infty$. En outre, le bon fonctionnement de l'estimateur Lasso repose sur deux hypothèses de régularité supplémentaires. La première est la parcimonie ou *sparsité* du vecteur β_0 dans le sens où l'on suppose $\|\beta_0\|_0 = \sum_{j=1}^{p} \mathbb{1}\{\beta_{0j} \neq 0\} < s$ pour un entier fixe $s << p$. Cela signifie que l'on présuppose que seul un petit nombre de régresseurs possèdent effectivement un coefficient non nul. Cela revient à changer de perspective pour percevoir le problème de la grande dimension comme un problème de sélection de variables, ce qui n'était pas le cas pour l'estimateur Ridge. La seconde hypothèse, dite de *valeur propre restreinte* (*restricted eigenvalue* en Anglais), suppose l'existence d'une borne inférieure positive sur la plus petite valeur propre des matrices de Gram calculées à partir des vecteurs contenant au plus s éléments de X_i. Il s'agit d'une hypothèse technique mais vérifiée sous des conditions raisonnables sur la distribution de X_i. Sous ces deux hypothèses, et en choisissant une valeur

particulière à λ, de l'ordre de $\lambda_n \propto \sqrt{\log(p)/n}$, on peut montrer que l'erreur d'estimation est de la forme $\left\|\widehat{\beta}^L(\lambda_n) - \beta_0\right\|_2 \lesssim \sqrt{s\log(p)/n}$ à comparer avec celle de l'estimateur (MCO) dans le cas de s régresseurs qui est de la forme $\left\|\widehat{\beta} - \beta_0\right\|_2 \lesssim \sqrt{s/n}$. Autrement dit : on "paye" un prix $\sqrt{\log(p)}$ pour notre ignorance concernant l'identité des s variables pertinentes par rapport aux p variables incluses dans le modèle. A noter que l'estimateur $\widehat{\beta}^L(\lambda_n)$ ne possède pas de distribution asymptotique gaussienne. Pour s'en rendre compte, il suffit de voir que certains coefficients sont strictement nuls avec une probabilité positive.

Sélection de variables. Qu'en est-il des propriétés de sélection de variables de l'estimateur (LASSO) ? On s'intéresse, dans ce cas, à l'erreur qui compte le nombre de faux positifs ou de faux négatifs produits par l'estimateur Lasso. Soit $S_0 = \{j, \beta_{0,j} \neq 0\}$ et $\widehat{S} = \left\{j, \widehat{\beta}^L_j(\lambda) \neq 0\right\}$, on s'intéresse à l'évènement :

$$\left\{S_0 = \widehat{S}\right\}.$$

On peut montrer que sous des conditions techniques, en particulier si les régresseurs contenus dans X_i ne sont pas trop corrélés entre eux et que les coefficients non-nuls sont suffisamment différents de zéro (condition appelée *beta-min* dans la littérature), alors le Lasso sélectionne le vrai modèle avec probabilité tendant vers 1 lorsque $n \to \infty$ pour un certain λ bien choisi (Zhao and Yu, 2006). Attention toutefois, Yang (2005) montre que pour que n'importe quelle procédure de sélection de modèles soit convergente, elle doit se comporter de façon sous-optimale pour estimer la fonction de régression, et vice-versa. Autrement dit : il n'existe pas un unique outil qui permette à la fois d'estimer la fonction $x \to E[Y|X = x] = x'\beta_0$ et l'ensemble $\{j, \beta_{0j} \neq 0\}$ puisque l'optimalité du Lasso pour ces deux objectifs dépend de valeurs différentes de la pénalité λ. On rappelle également que comme pour un modèle de régression classique, en l'absence d'hypothèse forte, l'interprétation causale n'est pas non plus garantie.

Biais et Post-Lasso. L'estimateur Lasso souffre d'un biais de pénalisation à distance finie dans la mesure où même les coefficients non-nuls sont "tirés" vers zéro. Afin de diminuer ce biais, on peut mettre en œuvre une seconde étape dite "Post-Lasso" (Belloni and Chernozhukov, 2013) en réestimant β_0 par moindres carrés ordinaires, ayant au préalable sélectionné uniquement les régresseurs correspondant à un coefficient non nul dans $\widehat{\beta}^L(\lambda)$.

$$\widehat{\beta}^{PL}(\lambda) = \underset{\beta:\beta_j=0 \text{ si } \widehat{\beta}^L_j(\lambda)=0}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2. \qquad \text{(POST-LASSO)}$$

Mise en pratique 2: Lasso avec `glmnet`

Le notebook `RidgeLasso-glmnet` met en pratique la régression Lasso avec le package `glmnet` à partir de données simulées.

2.4. Ridge ou Lasso ?

Le type de pénalisation (RIDGE) ou (LASSO) n'est pas anodin et va donner lieu a des solutions de natures différentes, reflétant des *a priori* différents sur le paramètre β_0 sous-jacent. Dans le cas de la régression Ridge, la solution est dite *dense* dans la mesure où les éléments de $\widehat{\beta}^R(\lambda)$ vont prendre des valeurs petites mais jamais exactement zéro. Dans le cas de la régression Lasso, la solution est dite *parcimonieuse* (ou *sparse*) dans la mesure où généralement $\widehat{\beta}^L(\lambda)$ est un vecteur dont beaucoup d'éléments vont valoir exactement zéro. En ce sens, seul le Lasso permet de faire de la sélection de variables. Il est à noter que le Lasso peut être vu comme une convexification de la pénalisation par la norme ℓ_0 qui compte le nombre de coefficients non nuls $\|\beta\|_0$. Néanmoins le programme :

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2 + \lambda\|\beta\|_0,$$

est NP-difficile et donc computationnellement long. En pratique, cela signifie que que l'on va vouloir utiliser le (RIDGE) quand on pense que les variables ont toutes une influence mais qu'elle est faible. Inversement, le (LASSO) est utilisé pour sélectionner des variables, lorsque l'on pense que seul un nombre limité d'éléments de β_0 sont réellement différents de zéro, ou que l'on veut obtenir une fonction de régression interprétable facilement. Ainsi, dans l'exemple des déterminants de la croissance de Sala-I-Martin (1997), on voudrait plutôt utiliser le Lasso, tandis que Ferrara and Simoni (2019) utilisent un Ridge reflétant le fait que parmi les nombreuses séries temporelles beaucoup contiennent un fragment d'information pertinente.

Ce choix peut également être interprété sous l'angle bayésien. De ce point de vue, les solutions (RIDGE) et (LASSO) sont deux Maximum A Posteriori (MAP) provenant de distributions *a priori* de β_0 différentes.

Lemme 1 (Interprétation bayésienne) *Dans le modèle* $Y_i = X_i'\beta_0 + \varepsilon_i$ *avec* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ *:*

— *L'estimateur* (RIDGE) *est le MAP provenant de l'a priori* $\beta_0 \sim \mathcal{N}_p\left(0, (\sigma^2/\lambda)\mathbb{I}_p\right)$,

— *L'estimateur* (LASSO) *est le MAP provenant de l'a priori* $\beta_0 \sim \mathcal{L}(1/\lambda)^{\otimes p}$, *où* $\mathcal{L}(1/\lambda)$ *désigne la loi de Laplace de paramètre* $1/\lambda$ *dont la densité est* $x \to (\lambda/2)\exp(-\lambda|x|)$.

Notons cependant que ces deux distributions *a priori* font l'hypothèse que le coefficient se concentre autour du vecteur nul, $\mathbb{E}[\beta_0] = 0$. On remarque que la variance de la loi *a priori* varie en sens inverse de la pénalité λ : plus λ est grand et plus la loi *a priori* sera concentrée autour du vecteur nul [1].

Notons finalement l'existence de l'estimateur dit *elastic net* combinant les deux types de pénalisation et prenant la forme, pour $\alpha \in [0, 1]$ et $\lambda \geq 0$:

$$\widehat{\beta}^E(\lambda, \alpha) = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2), \quad \text{(ELASTIC-NET)}$$

Avec évidemment, $\widehat{\beta}^E(\lambda, 1) = \widehat{\beta}^L(\lambda)$ et $\widehat{\beta}^E(\lambda, 0) = \widehat{\beta}^R(\lambda)$. Dès que $\alpha > 0$, la solution proposée sera sparse, comme le Lasso.

Mise en pratique 3: Elastic Net avec `glmnet`

Le notebook `RidgeLasso-glmnet` met en pratique la régression Elastic-Net avec le package `glmnet` à partir de données simulées. Les notations de ce document correspondent à celles du package, ainsi $\alpha = 0$ donne la solution Ridge et $\alpha = 1$ la solution Lasso.

Il existe d'autres types de pénalisation plus sophistiqués, reflétant certains schémas de sparsté définis *a priori*, tels que le Group-Lasso. Le Group-Lasso définit des groupes de variables qui sont pensées comme étant non-nulles en même temps mais suppose qu'il existe une sparsité au niveau des *groupes* de variables. Cela peut être le cas si l'on a des régresseurs de base que l'on souhaite croiser avec des indicatrices de groupes (sociaux-démographiques). Dans ce cas, on aura tendance à penser que si un des coefficient est différent de zéro pour un groupe donné, il le sera aussi pour les autres groupes. Soit

1. La variance d'une loi de la Laplace de paramètre $1/\lambda$ vaut $2/\lambda^2$.

$\mathcal{G} = \{G_1, \ldots, G_G\}$ une partition de $\{1, \ldots, p\}$ en G groupes et β_{G_g} le vecteur β dont les entrées non contenues dans G_g sont égales à 0. Le Group-Lasso est défini par :

$$\widehat{\beta}^{GL}(\lambda, w_1, \ldots, w_G) = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2 + \lambda \sum_{g=1}^{G} w_g \left\|\beta_{G_g}\right\|_2, \quad \text{(GROUP-LASSO)}$$

avec w_g la pénalité spécifique au groupe g, généralement fixée de façon proportionnelle à $\sqrt{\text{Card}(G_g)}$.

Au delà de ces considérations théoriques, quand il s'agit d'une tâche de prédiction, le choix entre différents types de pénalisation se fait par comparaison de la perte moyenne sur un échantillon de test.

2.5. Choix de λ par validation croisée

Les garanties théoriques sur l'estimateur Lasso reposent sur des choix de pénalisation λ théoriques et dans la plupart des cas infaisables car dépendant de quantités inconnues. Certains auteurs ont développé des algorithmes permettant d'aboutir à une pénalisation asymptotiquement optimale (pour l'estimation) qui marchent assez bien en pratique (*e.g.* Belloni et al., 2014a, disponible dans le package `hdm` pour `R`).

Néanmoins, dans la grande majorité des cas et en particulier quand il s'agit d'une tâche de prédiction, on utilise une procédure empirique dite de *validation croisée* pour choisir λ. L'idée est de séparer les données en deux parties disjointes, l'une sur laquelle on va calculer l'estimateur pour un λ donné, et l'autre sur laquelle on va optimiser ce λ de sorte à minimiser l'erreur hors de l'échantillon (*out-of-sample*). Cette procédure a pour but d'éviter le sur-apprentissage, c'est-à-dire éviter de trop "coller" aux données d'apprentissage et d'obtenir un estimateur de β_0 qui soit peu performant sur un échantillon qui n'a pas été utilisé pour le calculer. Pour se convaincre de l'utilité de cette procédure, il suffit de noter, par exemple, que $\lambda = 0$ est solution de :

$$\min_{\lambda} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i'\widehat{\beta}^{L}(\lambda)\right)^2,$$

car $\widehat{\beta}^{L}(0) = \widehat{\beta}$ est solution de l'équation (MCO). Pour autant, cela ne garantit pas que l'estimateur (MCO) sera nécessairement celui qui produit une erreur quadratique moyenne minimale sur un échantillon n'ayant pas été utilisé pour son calcul. Pour choisir

λ, il est donc nécessaire de d'utiliser un échantillon différent de celui utilisé pour calculer $\widehat{\beta}^L(\lambda)$ à λ donné.

La procédure est la suivante. Pour simplifier les notations, on supposera que $n = K \times n_0$ pour deux entiers K et n_0.

1. Pour un entier K, tirer au hasard une partition de $\{1, \ldots, n\}$ en K groupes de tailles égales n_0 (*folds*). On note $G_i \in \{1, \ldots, K\}$ le groupe d'appartenance de l'observation i.

2. Pour chaque $k = 1, \ldots, K$, en utilisant uniquement les données qui n'appartiennent pas au groupe k, on calcule l'estimateur Lasso ou Ridge :

$$\widehat{\beta}_k^R(\lambda) = \arg\min_\beta \frac{1}{(K-1)n_0} \sum_{G_i \neq k} (Y_i - X_i'\beta)^2 + \lambda \|\beta\|_2^2 ,$$
$$\widehat{\beta}_k^L(\lambda) = \arg\min_\beta \frac{1}{(K-1)n_0} \sum_{G_i \neq k} (Y_i - X_i'\beta)^2 + \lambda \|\beta\|_1 .$$

3. Pour chaque $k = 1, \ldots, K$, on calcule l'erreur sur le groupe k :

$$\frac{1}{n_0} \sum_{G_i = k} \left(Y_i - X_i'\widehat{\beta}_k^R(\lambda)\right)^2 ,$$
$$\frac{1}{n_0} \sum_{G_i = k} \left(Y_i - X_i'\widehat{\beta}_k^L(\lambda)\right)^2 .$$

4. On agrége les erreurs précédentes et on minimise en λ :

$$\widehat{\lambda}^R = \arg\min_\lambda \frac{1}{K} \sum_{k=1}^K \frac{1}{n_0} \sum_{G_i = k} \left(Y_i - X_i'\widehat{\beta}_k^R(\lambda)\right)^2 ,$$
$$\widehat{\lambda}^L = \arg\min_\lambda \frac{1}{K} \sum_{k=1}^K \frac{1}{n_0} \sum_{G_i = k} \left(Y_i - X_i'\widehat{\beta}_k^L(\lambda)\right)^2 .$$

En pratique, on prend régulièrement des valeurs $K = 5$ ou $K = 10$.

Mise en pratique 4: Validation croisée

Le notebook `RidgeLasso-glmnet` met en pratique la méthode de la validation croisée, à la fois telle que décrite ici, et directement à l'aide du package `glmnet` (fonction `cv.glmnet`) – ce qui est bien plus efficient.

2.6. Extensions

L'approche pénalisée ne se limite pas à la perte quadratique (modèle de régression linéaire), mais peut également être adaptée à tout estimateur qui minimise un risque ou maximise une vraisemblance. Pour s'en convaincre, voici quelques exemples.

Exemple 1 (Régression logistique) *Dans ce cas, la variable cible est binaire,* $Y_i \in \{0,1\}$ *et on suppose le lien suivant* $P[Y_i = 1|X_i] = \exp(X_i'\gamma_0) / (1 + \exp(X_i'\gamma_0))$*, soit encore* $P[Y_i = y|X_i] = \exp(yX_i'\gamma_0) / (1 + \exp(X_i'\gamma_0))$ *pour* $y \in \{0,1\}$*. La contribution individuelle à la vraisemblance s'écrit ainsi :* $\exp(Y_iX_i'\gamma_0) / (1 + \exp(X_i'\gamma_0))$*, d'où l'estimateur standard du maximum de vraisemblance de* γ_0 *:*

$$\widehat{\gamma}^{EMV} = \arg\min_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(X_i'\gamma)) - Y_iX_i'\gamma. \tag{LOGIT}$$

Une version pénalisée populaire de (LOGIT) *est le* (LOGIT-LASSO) *(voir par exemple Van de Geer, 2008) :*

$$\widehat{\gamma}^{L}(\lambda) = \arg\min_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(X_i'\gamma)) - Y_iX_i'\gamma + \lambda \|\gamma\|_1 . \tag{LOGIT-LASSO}$$

□

Mais on peut également utiliser une pénalisation par la norme ℓ_2, et les concepts développés au début de cette section (validation croisée etc.) s'y appliquent également.

Exemple 2 (Modèle de durée avec censure) *Supposons que l'on s'intéresse à la durée de vie d'un individu (e.g être humain, voiture, etc.) que l'on note* $T_i \geq 0$*. Cependant, pour les individus encore en vie au moment de l'étude, on n'observe qu'une borne inférieure sur cette durée, que l'on note* C_i*. On note la variable observée* $Y_i = \min(T_i, C_i)$ *et on note* $D_i = \mathbb{1}\{T_i < C_i\}$*. On suppose* $T_i \perp\!\!\!\perp C_i|X_i$ *et* $T_i \sim \mathcal{E}(\exp(X_i'\theta_0))$*. L'estimateur du maximum de vraisemblance de* θ_0 *est :*

$$\begin{aligned}
\widehat{\theta}^{EMV} &= \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} D_i(\exp(X_i'\theta)Y_i - X_i'\theta) + (1 - D_i)\exp(X_i'\theta)Y_i, \\
&= \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \exp(X_i'\theta)Y_i - D_iX_i'\theta.
\end{aligned}$$

On peut préférer une version pénalisée de cet estimateur :

$$\widehat{\theta}^R(\lambda) = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n \exp(X_i'\theta)Y_i - D_i X_i'\theta + \lambda \|\theta\|_2^2 .$$

□

Mise en pratique 5: Différentes fonctions de perte avec `glmnet`

Le notebook `RidgeLasso-glmnet` met en pratique ces méthodes d'estimations pour différentes fonctions de perte via `glmnet`. Il suffit de changer l'argument `family` : *binomial* correspond à un modèle Logit, *poisson* à une régression de Poisson, *cox* à un modèle de durée à hasard proportionnel.

3. L'inférence post-sélection

Dans cette section, on traite de la question de l'inférence post-sélection, c'est-à-dire de la construction d'intervalles de confiance et de tests à partir d'un modèle ayant fait l'objet de sélection dans une étape précédente, par exemple suite à une régression Lasso.

Imaginons par exemple que l'on s'intéresse à l'effet d'une variable spécifique X_1 (*e.g.* le niveau d'éducation, un traitement particulier) sur une mesure de résultat Y (*e.g.* le salaire, le retour à l'emploi) en prenant en compte un certain nombre d'autres caractéristiques X_2 qui sont potentiellement de grande dimension. Pour cela, on souhaite estimer le modèle $Y = X_1\tau_0 + X_2'\beta_0 + \varepsilon$. Par souci de parcimonie, de rigueur, de précision, pour obtenir des résultats conformes à son intuition, ou tout simplement pour des questions de faisabilité (si X_2 est de grande dimension), le chargé d'étude peut vouloir sélectionner parmi les variables X_2 et reporter les résultats d'un modèle restreint. Il va par exemple utiliser une première étape du type Lasso sans pénaliser le coefficient associé à X_1 puisqu'il s'agit de la variable d'intérêt :

$$\widehat{\beta}^L(\lambda) \in \arg\min_{\tau,\beta} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_{1,i}\tau - X_{2,i}'\beta\right)^2 + \lambda\left\|\beta\right\|_1,$$

puis ensuite reporter les résultats d'un modèle "court" où il aura éliminé les variables associées à un coefficient nul dans la première étape :

$$(\widehat{\tau},\widehat{\beta}) = \arg\min_{\tau,\beta:\beta_j=0 \text{ si } \widehat{\beta}_j^L(\lambda)=0} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_{1,i}\tau - X_{2,i}'\beta\right)^2.$$

Le problème est que $\widehat{\tau}$ a de très mauvaise propriétés et utiliser une procédure d'inférence standard sur cet estimateur va donner des résultats faux. Il faut le corriger pour l'immuniser contre les erreurs de sélection et qu'*in fine* il soit asymptotiquement gaussien.

En pratique, les économistes empiriques sélectionnent souvent leurs variables par tâtonnement, guidés par l'intuition et reportent leurs résultats sous l'hypothèse implicite que le modèle sélectionné est le bon. Leurs résultats sont ensuite mis à l'épreuve par des analyses de sensitivité ou des tests de robustesse. Cela dit, l'étape de sélection de variables dans les travaux empiriques est couramment passée sous silence quand bien même elle n'est pas anodine. Leamer (1983) était l'un des premiers à tirer la sonnette d'alarme, en plaisantant : "il y a deux choses dont on préfère ne pas savoir comment elles sont faites :

ce sont les saucisses et les estimations économétriques". Pour une présentation moderne, voir Leeb and Pötscher (2005) et, dans un contexte d'évaluation des politiques publiques, Belloni et al. (2014a). Cette section reprend en partie le premier chapitre de Gaillac and L'Hour (2019) et l'introduction de L'Hour (2019).

3.1. Le problème de l'inférence post-sélection

Commençons par analyser la méthode d'inférence en deux étapes telle que décrite précédemment, *i.e.* d'abord sélectionner le modèle, puis reporter les résultats de ce modèle en faisant comme si c'était le vrai, dans un contexte de petite dimension. L'intuition s'étend facilement au cas de la grande dimension. Cette section est basée sur le travail de Leeb and Pötscher (2005).

Hypothèse 1 (Modèle linéaire gaussien possiblement sparse) *Considérons la séquence de variables aléatoires indépendantes et identiquement distribuées* $(Y_i, X_i)_{i=1,\dots,n}$ *telles que :*

$$Y_i = X_{i,1}\tau_0 + X_{i,2}\beta_0 + \varepsilon_i,$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, σ^2 *est connue,* $X_i = (X_{i,1}, X_{i,2})$ *est un vecteur de dimension deux,* $\varepsilon_i \perp\!\!\!\perp X_i$, *et* $\mathbb{E}(X_i X_i')$ *est inversible. On utilise la notation suivante pour la matrice de variance-covariance de l'estimateur des moindre carrés :*

$$\begin{bmatrix} \sigma_\tau^2 & \sigma_{\tau,\beta} \\ \sigma_{\tau,\beta} & \sigma_\beta^2 \end{bmatrix} := \sigma^2 \left[\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right]^{-1}.$$

Le modèle vrai le plus parcimonieux est codé par M_0, *une variable aléatoire prenant la valeur* R *("restreint") si* $\beta_0 = 0$ *et* U *sinon.*

L'économètre souhaite faire de l'inférence sur le paramètre τ_0 et se demande s'il devrait inclure ou non $X_{i,2}$ dans la régression. Dans une seconde étape, il reporte le résultat du modèle $\widehat{M}$ qu'il a sélectionné dans un premier temps. En évaluation des politiques publiques, $X_{i,1}$ correspond généralement au traitement d'intérêt et $X_{i,2}$ à une variable de contrôle. On note $\widehat{\tau}(U)$ et $\widehat{\beta}(U)$ les estimateurs MCO du modèle sans restriction (modèle U) et $\widehat{\tau}(R)$ et $\widehat{\beta}(R) = 0$ les estimateurs MCO du modèle restreint (modèle R).

Tout dans cette section sera conditionnel aux variables $(X_i)_{1\leq i\leq n}$ mais cette dépendance est laissée masquée. En particulier, conditionnellement aux variables, l'estimateur

sans restriction a une distribution gaussienne :

$$\sqrt{n}\begin{bmatrix}\widehat{\beta}(U)-\beta_0\\ \widehat{\tau}(U)-\tau_0\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma_\beta^2 & \sigma_{\tau,\beta}\\ \sigma_{\tau,\beta} & \sigma_\tau^2\end{bmatrix}\right).$$

L'économètre inclut $X_{i,2}$ dans la régression si la statistique de Student correspondant au test de l'hypothèse nulle H_0 :"$\beta_0 = 0$" est suffisamment grande :

Hypothèse 2 (Règle de décision)

$$\widehat{M}=\begin{cases}U & \textit{if } |\sqrt{n}\widehat{\beta}(U)/\sigma_\beta|>c_n\\ R & \textit{sinon,}\end{cases}$$

avec $c_n\to\infty$ *et* $c_n/\sqrt{n}\to 0$ *lorsque* $n\to\infty$.

Le critère AIC correspond à $c_n=\sqrt{2}$ et le BIC à $c_n=\sqrt{\log n}$. On va voir que cette méthode est convergente, c'est-à-dire qu'elle sélectionne le bon modèle avec une probabilité tendant vers 1, mais que cela n'est pas suffisant pour faire fonctionner un cadre d'inférence standard sur le modèle classique. Cela est en particulier dû aux coefficients qui sont proches de zéro, ce qui rend difficile leur détection.

Lemme 2 (Convergence de la sélection de modèle) *Pour* $M_0\in\{U,R\}$,

$$\mathbb{P}_{M_0}\left(\widehat{M}=M_0\right)\to 1,$$

quand $n\to\infty$, *où* $\mathbb{P}_{M_0}$ *indique la distribution de* $\widehat{M}$ *sous le vrai modèle* M_0.

Toutes les preuves se trouvent dans Gaillac and L'Hour (2019) ou bien dans l'article original de Leeb and Pötscher (2005). Le Lemme 2 peut amener à penser qu'une procédure de sélection de modèle convergente permet de réaliser une inférence "comme d'habitude", c'est-à-dire en négligeant l'étape de sélection du modèle. Cependant, pour toute taille d'échantillon donnée n, la probabilité de sélectionner le vrai modèle peut être très faible si β_0 est proche de zéro sans être exactement à zéro. Par exemple, supposons que $\beta_0=\delta\sigma_\beta c_n/\sqrt{n}$ avec $|\delta|<1$ alors : $\sqrt{n}\beta_0/\sigma_\beta=\delta c_n$ et la probabilité dans la preuve du Lemme 2 est égale à $1-\Phi(c_n(1+\delta))+\Phi((\delta-1)c_n)$, et tend vers zéro bien que le vrai modèle soit U parce que $\beta_0\neq 0$! Cette analyse rapide nous indique que la procédure de sélection de modèle est aveugle aux petits écarts par rapport au modèle restreint ($\beta_0=0$) qui sont de l'ordre de $c_n/\sqrt{n}$. Les statisticiens disent que dans ce cas, la procédure de

sélection de modèle n'est pas *uniformément convergente* par rapport à β_0. Pour l'économètre appliqué, cela signifie que la procédure d'inférence classique, *i.e.*, qui suppose que le modèle sélectionné est le vrai et utilise la normalité asymptotique pour effectuer des tests peut nécessiter des échantillons de très grande taille pour être précis. De plus, cette taille d'échantillon requise dépend du paramètre inconnu β_0.

Plus intéressant encore, Leeb and Pötscher (2005) analysent la distribution de l'estimateur de post-sélection $\tilde{\tau}$ défini par :

$$\tilde{\tau} := \widehat{\tau}(\widehat{M}) = \widehat{\tau}(R)\mathbf{1}_{\widehat{M}=R} + \widehat{\tau}(U)\mathbf{1}_{\widehat{M}=U}.$$

Compte tenue de la mise en garde émise dans le paragraphe précédent, une procédure de sélection de modèle convergente est-elle suffisante pour lever les préoccupations relatives à l'approche de post-sélection ? En effet, en utilisant le Lemme 2, il est tentant de penser que, $\tilde{\tau}$ sera asymptotiquement distribué comme une gaussienne et que l'inférence asymptotique standard peut être utilisée pour estimer le comportement de l'estimateur à distance finie. Cependant, on va voir que sa distribution en échantillon fini peut être très différente d'une distribution gaussienne standard. En particulier, le Lemma 3 ci-dessous montre que la distribution de l'estimateur de post-sélection peut être complexe et non centrée sur la vraie valeur du paramètre τ_0, souffrant tout particulièrement d'un biais de variable omise.

Lemme 3 (Densité de l'estimateur de post-sélection, Leeb, 2006) *La densité à distance finie (conditionnelle à $(X_i)_{i=1,\dots,n}$) de $\sqrt{n}(\tilde{\tau}-\tau_0)$ est donnée par :*

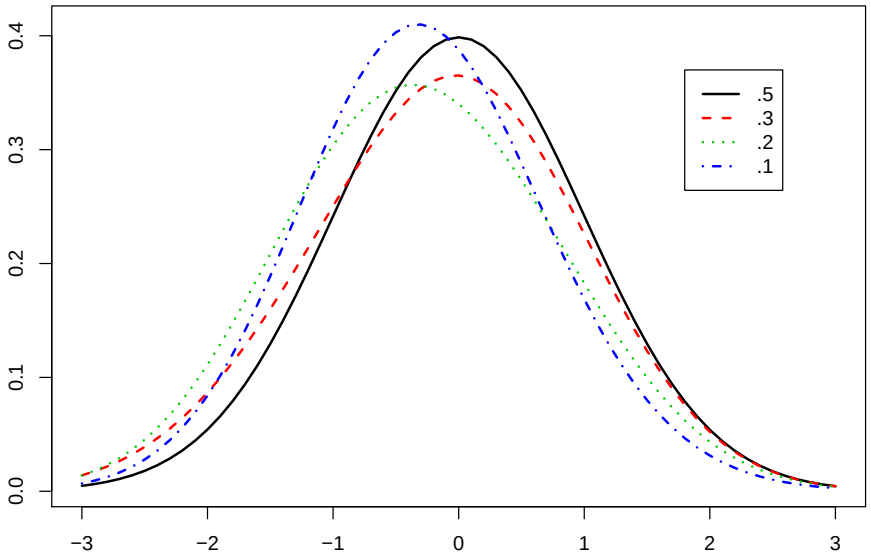
$$f_{\sqrt{n}(\tilde{\tau}-\tau_0)}(x) = \Delta\left(\sqrt{n}\frac{\beta_0}{\sigma_\beta}, c_n\right)\frac{1}{\sigma_\tau\sqrt{1-\rho^2}}\varphi\left(\frac{x}{\sigma_\tau\sqrt{1-\rho^2}} + \frac{\rho}{\sqrt{1-\rho^2}}\frac{\sqrt{n}\beta_0}{\sigma_\beta}\right)$$
$$+\left[1-\Delta\left(\frac{\sqrt{n}\beta_0/\sigma_\beta+\rho x/\sigma_\tau}{\sqrt{1-\rho^2}}, \frac{c_n}{\sqrt{1-\rho^2}}\right)\right]\frac{1}{\sigma_\tau}\varphi\left(\frac{x}{\sigma_\tau}\right),$$

où $\rho = \sigma_{\tau,\beta}/\sigma_\tau\sigma_\beta$, $\Delta(a,b) := \Phi(a+b)-\Phi(a-b)$ et φ and Φ sont respectivement la densité et la fonction de répartition de la loi $\mathcal{N}(0,1)$.

Notons que le biais correspond au biais de variable omise habituel puisque

$$-\beta_0\rho\sigma_\tau/\sigma_\beta \xrightarrow{p} \beta_0 \underbrace{\mathrm{Cov}(X_{i,1}, X_{i,2})/\mathrm{V}(X_{i,1})}_{\substack{\text{Coefficient de } X_1 \text{ dans la} \\ \text{régression de } X_2 \text{ sur } X_1.}}.$$

FIGURE 1 – Densité à distance finie de $\sqrt{n}(\tilde{\tau} - \tau_0)$, $\rho = .4$



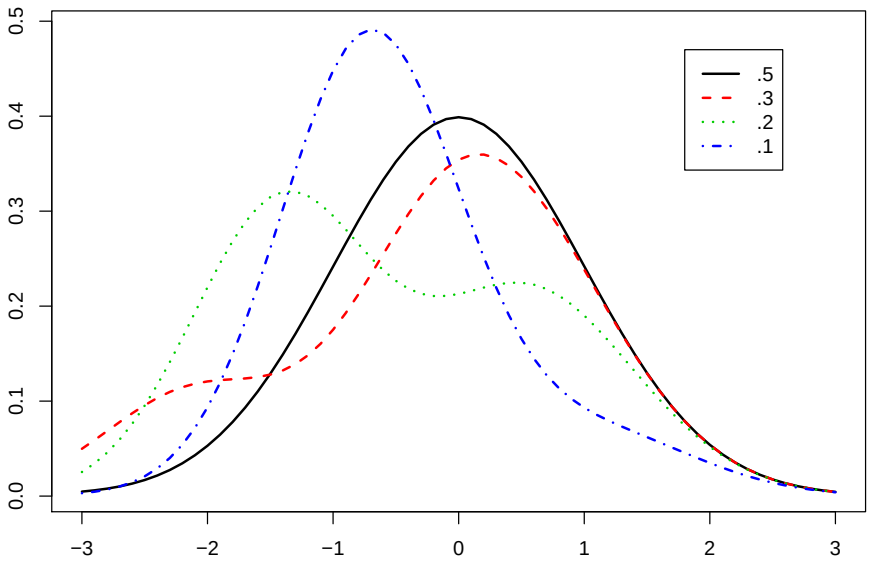
Note : Densité de l'estimateur de post-sélection $\tilde{\tau}$ pour différentes valeurs de β_0/σ_β, voir la légende. Les autres paramètres sont fixés à : $c_n = \sqrt{\log n}$, $n = 100$, $\sigma_\tau = 1$ et $\rho = .4$. Voir Lemme 3 pour la formule.

Le problème fondamental est le biais de variable omise que l'estimateur de post-sélection ne peut surmonter que si $\beta_0 = 0$ ou bien si $\rho = 0$. En effet, lorsque $\rho = 0$, $\sqrt{n}(\tilde{\tau} - \tau_0) \sim \mathcal{N}(0, \sigma_\tau^2)$; alors que quand $\beta_0 = 0$, $\sqrt{n}(\tilde{\tau} - \tau_0) \sim \mathcal{N}(0, \sigma_\tau^2/(1-\rho^2))$ (environ), car $\Delta(0, c_n) \geq 1 - \exp(-c_n^2/2)$ - la probabilité de sélection du modèle restreint est grande. Les figures 1 et 2 tracent la densité en échantillon fini de l'estimateur de post-sélection pour plusieurs valeurs de β_0/σ_β dans les cas $\rho = .4$ et $\rho = .7$, respectivement. La figure 1 montre une distorsion légère mais significative par rapport à une distribution gaussienne standard. L'estimateur de post-sélection présente clairement un biais. À mesure que la corrélation entre les deux variables s'intensifie, la densité de l'estimateur post-sélection devient extrêmement non gaussienne, présentant même deux modes. Suite à cette analyse, il est clair que l'inférence basée sur les quantiles gaussiens standards donnera en général une image très différente de la véritable distribution illustrée par la figure 2.

Mise en pratique 6: Densité de l'estimateur de post-sélection

Le notebook `LeebPotscher` permet de reproduire les figures 1 et 2.

FIGURE 2 – Densité à distance finie de $\sqrt{n}(\tilde{\tau} - \tau_0)$, $\rho = .7$



Note : Voir Figure 1. $\rho = .7$.

3.2. Biais de régularisation : simulations

Illustrons l'intuition de la section précédente avec des simulations sur un cas qui se rapproche de la pratique.

Le processus générateur des données est : $Y_i = D_i\tau_0 + X_i'\beta_0 + \varepsilon_i$, où $\tau_0 = .5$, $\varepsilon_i \perp\!\!\!\perp X_i$, and $\varepsilon_i \sim \mathcal{N}(0,1)$. L'équation de traitement suit un modèle Probit, $D_i|X_i \sim \text{Probit}\,(X_i'\delta_0)$. On simule les variables de contrôle comme des Gaussiennes, $X_i \sim \mathcal{N}(0,\Sigma)$, où chaque entrée de la matrice de variance-covariance est fixée de la façon suivante : $\Sigma_{j,k} = 0.5^{|j-k|}$. Un élément de X_i sur deux est remplacé par 1 si $X_{i,j} > 0$ et par 0 sinon. La partie la plus intéressante du processus générateur des données est δ_0 et β_0 :

$$\beta_{0j} = \begin{cases} \rho_d(-1)^j/j^2,\ j < p/2 \\ 0,\ \text{sinon} \end{cases}, \delta_{0j} = \begin{cases} \rho_y(-1)^j/j^2,\ j < p/2 \\ \rho_y(-1)^{j+1}/(p-j+1)^2,\ \text{sinon} \end{cases}$$

Pour les deux équations, nous sommes dans un scénario dit approximativement sparse. ρ_y and ρ_d sont des constantes qui fixent le ratio signal/bruit, au sens où plus elles sont grandes, plus les variables X_i vont jouer un rôle important dans chaque équation. L'astuce est que certaines variables qui jouent fortement dans l'assignation au traitement,

n'ont qu'un très faible effet dans la fonction de régression de Y. L'intuition est que les procédures de sélection à équation unique vont manquer certaines variables pertinentes dans l'équation de Y, ce qui va créer un biais et/ou une distribution non normale.

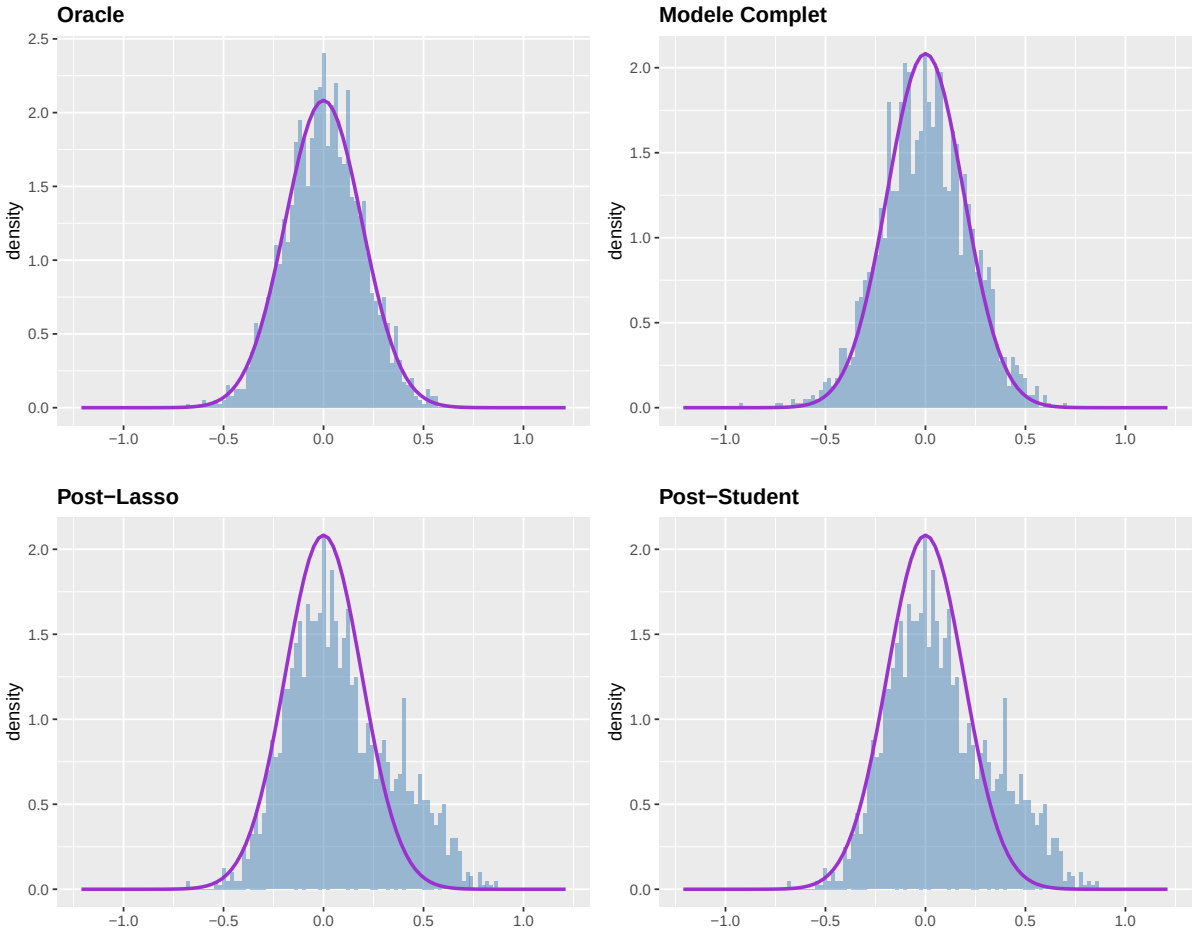
On compare le bias, l'erreur quadratique moyenne et le taux de couverture (le pourcentage des simulations dans lesquelles τ_0 est bien contenu dans l'intervalle de confiance) de quatre estimateurs de τ_0 dont deux correspondent à la pratique :

1. L'estimateur oracle de τ_0 qui consiste à régresser Y sur D et les éléments de X qui correspondent à des coefficients non-nuls dans β_0. Évidemment, cet estimateur est inutilisable en pratique puisqu'il repose sur la connaissance du schéma de sparsité de β_0, mais il sert de d'étalon pour évaluer la performance des autres modèles,
2. L'estimateur des moindres carrés de τ_0 dans le modèle complet $Y_i = D_i\tau_0 + X_i'\beta_0 + \varepsilon_i$,
3. L'estimateur (POST-LASSO) de τ_0 dans le modèle $Y_i = D_i\tau_0 + X_i'\beta_0 + \varepsilon_i$ où on l'on n'a pas pénalisé le coefficient associé à D dans la première étape. En outre, on sélectionne λ par validation croisée de sorte à minimiser l'erreur de prédiction, ce qui est complètement sous optimal pour une tâche d'estimation. Cet estimateur correspond à la pratique décrite au début de la section,
4. L'estimateur "Post-Student" de τ_0 qui correspond à l'estimateur des moindres carrés dans un modèle linéaire où l'on régresse Y sur D et les éléments de X qui ont une p-value associée inférieure à .01 dans le modèle complet $Y_i = D_i\tau_0 + X_i'\beta_0 + \varepsilon_i$. Cet estimateur est semblable au précédent, mais la procédure de sélection est différente.

D'après l'intuition développée dans la section précédente, seuls les deux premiers estimateurs devrait être sans biais (mais potentiellement avec une forte variance), alors que les deux autres devraient être biaisés et afficher une distribution loin de la loi normale.

Les résultats sont présentés dans la Figure 3 et la Table 1. On peut constater que conformément à la théorie, l'estimateur des MCO dans le modèle complet a un biais quasi nul et une distribution approximativement gaussienne – affichant des performances proches de l'Oracle. Cela n'est pas le cas pour les deux autres estimateurs qui souffrent à la fois d'un biais mais également d'une distribution qui n'est pas gaussienne.

FIGURE 3 – Distribution simulée des estimateurs de τ_0



Note : Chaque graphique représente la distribution simulée d'un estimateur de τ_0. Le graphique en haut à gauche représente l'oracle et celui d'en haut à droite, l'estimateur MCO dans le modèle complet. Ces estimateurs sont asymptotiquement gaussiens. Les deux autres estimateurs souffrent d'un biais de régularisation, c'est-à-dire d'un biais de variable omise. On notera que ce problème est particulièrement criant pour la seconde méthode (post-Lasso) mais le dernier estimateur (post-Student) souffre également d'un biais et d'une queue plus épaisse à droite. La courbe mauve représente la distribution gaussienne théorique centrée en zéro d'écart-type celui de l'oracle. Résultats de 2000 tirages, pour un taille d'échantillon de 200 observations et de 30 variables explicatives.

Mise en pratique 7: Répliquer ces simulations

Le notebook `RegularizationBias` permet de reproduire cet exercice. Il donne également accès à des fonctions que vous pouvez réutiliser pour mettre en pratique ces estimateurs.

TABLE 1 – Estimation de τ_0

	Estimateur :			
	Oracle (1)	*Modèle complet* (2)	*Post-Lasso* (3)	*Post-Student* (4)
Biais	-0.001	0.001	0.339	0.084
$\sqrt{EQM}$	0.192	0.217	0.432	0.274
Taux de couverture	0.951	0.944	0.312	0.774
Note :	Biais, racine de l'erreur quadratique moyenne et taux de couverture pour les quatre estimateurs décrits précédemment.			

3.3. Méthode de l'orthogonalisation

Maintenant, on suppose un ensemble de variables de contrôle de grande dimension : dans le modèle de la section précédente, on suppose que $p := \dim(X_{i,2})$ est grand, potentiellement plus grand que l'échantillon taille n, *i.e.* β_0 est un paramètre de nuisance de grande dimension. Les paramètres de nuisance de grande dimension nécessitant l'utilisation d'outils non standard tels que le Lasso, le cadre d'inférence standard peut s'en retrouver perturbé de façon similaire à ce que nous avons illustré dans les sections précédentes.

Le message transmis dans la section précédente était un message de prudence concernant l'utilisation de dispositifs de sélection tels que le Lasso dans les applications empiriques : l'inférence, sans prendre en compte l'étape de sélection de variables, peut être extrêmement trompeuse. Même en l'absence de référence explicite à la sélection de variables, l'estimation d'un paramètre de nuisance de grande dimension à l'aide d'algorithmes ML ne conduira généralement pas à un estimateur $\sqrt{n}$-convergent et entraîne ce que l'on appelle un biais de *régularisation* dans la seconde étape (Belloni et al., 2014b).

Maintenant, supposons que l'estimation $\widehat{\beta}$ de β_0 soit issue d'un algorithme de machine learning quel qu'il soit, mais probablement non convergent à la vitesse $\sqrt{n}$. Remarquons que l'équation normale implicitement utilisée pour définir τ_0 dans l'hypothèse 1 est

$$E\left[\left(Y_i - X_{i,1}\tau_0 - X_{i,2}'\beta_0\right) X_{i,1}\right] = 0, \tag{1}$$

et l'estimateur correspondant se décompose de la façon suivante :

$$\widehat{\tau} = \frac{\sum_{i=1}^n \left(Y_i - X_{i,2}'\widehat{\beta}\right) X_{i,1}}{\sum_{i=1}^n X_{i,1}^2}$$

$$
\begin{aligned}
&= \frac{\sum_{i=1}^n \left(Y_i - X_{i,2}'\beta_0\right) X_{i,1}}{\sum_{i=1}^n X_{i,1}^2} + \left(\beta_0 - \widehat{\beta}\right)' \frac{\sum_{i=1}^n X_{i,2}X_{i,1}}{\sum_{i=1}^n X_{i,1}^2} \\
&= \tau_0 + \frac{\sum_{i=1}^n \varepsilon_i X_{i,1}}{\sum_{i=1}^n X_{i,1}^2} + \underbrace{\left(\beta_0 - \widehat{\beta}\right)' \frac{\sum_{i=1}^n X_{i,2}X_{i,1}}{\sum_{i=1}^n X_{i,1}^2}}_{\text{Terme problématique.}},
\end{aligned}
$$

pour un estimateur machine learning $\widehat{\beta}$.

Le biais de régularisation provient du fait que $\widehat{\tau}$ n'est pas insensible au premier ordre aux erreurs dans l'estimation de β_0, *i.e.* en toute généralité, la quantité $\sum_{i=1}^n X_{i,2}X_{i,1} / \sum_{i=1}^n X_{i,1}^2$ ne converge pas vers zéro. Cette insensibilité aux erreurs d'estimation de β_0, combinée au fait qu'en général $\widehat{\beta}$ ne possède pas la propriété $\sqrt{n}(\beta_0 - \widehat{\beta}) = O_P(1)$ (ce qui est le cas pour la plupart des procédure de machine learning [2]), signifie qu'en général $\sqrt{n}\left(\widehat{\tau} - \tau_0\right)$ ne sera pas asymptotiquement gaussien de moyenne nulle. Bien sûr, lorsque β_0 est de petite dimension, cela ne pose pas de problème, car il suffit de remplacer $\widehat{\beta}$ par un estimateur MCO. Mais lorsque p est très grand, cela peut ne pas être possible ni même souhaitable.

L'astuce consiste à remplacer l'équation de moment (1) par une autre, $E[\psi(Y_i, X_i, \tau_0, \eta_0)] = 0$, pour un moment ψ et le paramètre de nuisance η_0 tel que $E[\partial_\eta \psi(Y_i, X_i, \tau_0, \eta_0)] = 0$. Cette dernière condition, qui est une condition d'orthogonalité, garantit que le moment d'estimation est insensible au premier ordre aux déviations du paramètre de nuisance par rapport à sa valeur vraie. Cela aide à "immuniser" l'estimateur de τ_0 contre une estimation de première étape qui utilise des outils de ML. Dans le cas de la régression linéaire, ψ prend la forme

$$
E[\psi(Y_i, X_i, \tau, \eta)] = E\Bigg[\underbrace{\left(Y_i - X_{i,1}\tau - X_{i,2}'\beta\right)}_{\substack{\text{Résidu de la régression} \\ \text{de } Y \text{ sur } X_1 \text{ et } X_2}} \quad \underbrace{\left(X_{i,1} - X_{i,2}'\delta\right)}_{\substack{\text{Résidu de la} \\ \text{régression de } X_1 \text{ sur } X_2}} \Bigg], \tag{2}
$$

avec $\eta = (\beta, \delta)$, et δ_0 tel que $E\left[X_{i,2}\left(X_{i,1} - X_{i,2}'\delta_0\right)\right] = 0$. Il est simple de vérifier que $E[\partial_\beta \psi(Y_i, X_i, \tau_0, \eta_0)] = E[\partial_\delta \psi(Y_i, X_i, \tau_0, \eta_0)] = 0$. Notons que l'orthogonalisation nécessite également l'estimation d'un autre paramètre de nuisance, δ_0, de même dimension que β_0. Pour obtenir le score orthogonalisé dans des cas plus généraux, on peut se référer à Chernozhukov et al. (2018a); Farrell (2015); Bléhaut et al. (2020)

Appelons $\check{\tau}$ l'estimation de τ_0 basée sur l'équation précédente, c'est-à-dire que $\check{\tau}$ est

2. Cela signifie par exemple que $\sqrt{n}(\beta_0 - \widehat{\beta})$ ne va pas tendre vers une loi Normale centrée, comme cela serait le cas avec un estimateur plus standard.

tel que

$$\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i, X_i, \check{\tau}, \widehat{\eta}) = 0.$$

Belloni et al. (2014b) (Théorème 1) montrent que si l'équation de moment ψ se décompose de la manière $\psi(Y_i, X_i, \tau, \eta) = \Gamma_1(Y_i, X_i, \eta)\tau - \Gamma_2(Y_i, X_i, \eta)$ pour des fonctions Γ_j, $j = 1, 2$, telles que leurs dérivées secondes par rapport à η sont constantes sur l'ensemble des valeurs possibles de η, alors sous des conditions de régularité :

$$\sqrt{n}\,(\check{\tau} - \tau_0) \xrightarrow{d} \mathcal{N}(0, \sigma_\Gamma^2),$$

avec $\sigma_\Gamma^2 := \mathbb{E}[\psi(Z_i, \tau_0, \eta_0)^2]/\mathbb{E}[\Gamma_1(Z_i, \eta_0)]^2$. Ce résultat est également valable sous des conditions plus générales, *e.g.* Chernozhukov et al. (2018a).

3.4. Application dans le cas linéaire : la double-sélection

L'équation (2) évoque le théorème de Frish-Waugh-Lovell selon lequel il est équivalent, pour estimer τ_0 d'estimer complétement le modèle de régression, ou bien de d'abord régresser $X_{i,1}$ sur $X_{i,2}$ puis de prendre le coefficient associé au résidu de cette régression dans la régression de Y_i sur ce même résidu. En effet, il suffit de voir que (2) prise en (τ_0, η_0) peut se réexprimer de la façon suivante :

$$E\Big[\left(Y_i - X_{i,1}\tau_0 - X_{i,2}'\beta_0\right)\left(X_{i,1} - X_{i,2}'\delta_0\right)\Big] = E\Big[\left(Y_i - (X_{i,1} - X_{i,2}'\delta_0)\tau_0\right)\left(X_{i,1} - X_{i,2}'\delta_0\right)\Big] = 0,$$

car on a supposé que $E\left[X_{i,2}\left(X_{i,1} - X_{i,2}'\delta_0\right)\right] = 0$.

L'idée de l'othogonalisation, bien qu'ancienne, a été développée et étendue dans une multitude d'articles écrits par Victor Chernozhukov et ses co-auteurs, par exemple Chernozhukov et al. (2015); Chernozhukov et al. (2015); Belloni et al. (2017); Chernozhukov et al. (2017, 2018a), et apparait sous le nom de *double sélection* quand le problème provient de l'utilisation du Lasso dans un modèle linéaire, estimateur *immunisé* ou *Neyman-orthogonal* dans un cadre général ou *double machine learning* quand il est spécifiquement adapté à des estimateurs de machine learning. En règle générale, lorsque l'on utilise le Lasso dans un cadre linéaire, la procédure prend la forme suivante :

1. Effectuer la régression Lasso X_1 sur X_2, récupérer $\widehat{\delta}^L$. Notons $\widehat{S}_D := \left\{j = 1, ..., p, \widehat{\delta}_j^L \neq 0\right\}$ l'ensemble des variables sélectionnées,

2. Effectuer la régression Lasso Y sur X_2, récupérer $\widehat{\beta}^L$. Notons $\widehat{S}_Y := \left\{j = 1, ..., p, \widehat{\beta}_j^L \neq 0\right\}$,
3. Effectuer la régression MCO de Y sur X_1 et les $\widehat{s} = |\widehat{S}_D \cup \widehat{S}_Y|$ éléments de X_2 qui correspondent aux indices $j \in \widehat{S}_D \cup \widehat{S}_Y$.

Et alors $\check{\tau}$ est le coefficient associé à X_1 dans la régression de l'étape (3). Définissions les estimateurs de post-selection $\widehat{\beta}$ et $\widehat{\delta}$:

$$\widehat{\beta} = \underset{\beta:\beta_j=0,\forall j\notin\widehat{S}_D\cup\widehat{S}_Y}{\arg\min} \sum_{i=1}^{n}(Y_i - X_{i,1}\widehat{\tau} - X_{i,2}'\beta)^2, \tag{3}$$

$$\widehat{\delta} = \underset{\delta:\delta_j=0,\forall j\notin\widehat{S}_D\cup\widehat{S}_Y}{\arg\min} \sum_{i=1}^{n}(X_{i,1} - X_{i,2}'\delta)^2. \tag{4}$$

Basé sur l'équation (2), l'estimateur de post-double-sélection $\check{\tau}$ possède la forme explicite :

$$\check{\tau} = \frac{n^{-1}\sum_{i=1}^{n}(Y_i - X_{i,2}'\widehat{\beta})(X_{i,1} - X_{i,2}'\widehat{\delta})}{n^{-1}\sum_{i=1}^{n} X_{i,1}(X_{i,1} - X_{i,2}'\widehat{\delta})}, \tag{POST-DOUBLE-SELEC}$$

et sera asymptotiquement gaussien. L'intuition est qu'une procédure de sélection basée sur deux étapes n'est pas sujet au biais de variable omise. Ici la variance asymptotique se calcule simplement :

$$\sigma_\Gamma^2 = \frac{\mathbb{E}\left[\left(X_{i,1} - X_{i,2}'\delta_0\right)^2\left(Y_i - \tau_0 X_{i,1} - X_{i,2}'\beta_0\right)^2\right]}{\mathbb{E}\left[\left(X_{i,1} - X_{i,2}'\delta_0\right)^2\right]^2},$$

et peut être estimée de façon convergente par :

$$\widehat{\sigma}_\Gamma^2 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(X_{i,1} - X_{i,2}'\widehat{\delta}\right)^2\right]^{-2} \frac{1}{n-\widehat{s}-1}\sum_{i=1}^{n}\left(X_{i,1} - X_{i,2}'\widehat{\delta}\right)^2\left(Y_i - \widehat{\tau}X_{i,1} - X_{i,2}'\widehat{\beta}\right)^2,$$

avec les estimateurs de post-double-sélection définis par (3) et (4). Le package `hdm` disponible sur `R` permet la mise en place de cet estimateur de façon simple.

Mise en pratique 8: Double-sélection

Le notebook `DoubleSelection` donne les clés pour mettre en oeuvre l'estimateur de double-sélection en utilisant des valeurs théoriques de λ données par Belloni et al. (2014b). On montre également comment calculer l'écart-type.

3.5. Partitionnement d'échantillon (*sample-splitting*)

Lorsque l'on utilise une première étape de type machine learning, il peut être également intéressant de mettre en place du *sample-splitting* c'est-à-dire séparer l'échantillon

entre un échantillon auxiliaire sur lequel on estime les paramètres de nuisance β_0 et δ_0 et un échantillon principal sur lequel on applique ces estimateurs pour calculer la quantité d'intérêt, τ_0. Cet outil requiert l'indépendance des données entre les observations contenues dans les deux échantillons. Outre le fait de faciliter l'analyse théorique des estimateurs, le but de cette technique est d'éviter le biais de sur-apprentissage qui peut avoir lieu lorsque l'échantillon d'entraînement est le même que l'échantillon de test. Mise en application dans notre cadre, cette technique revient à utiliser un échantillon auxiliaire pour estimer $\widehat{\delta}$ et $\widehat{\beta}$ dans les étapes (1) et (2), puis d'utiliser l'échantillon principal pour réaliser l'étape (3). Pour éviter les pertes de précision dues à la réduction de la taille de l'échantillon, on peut ensuite inverser le rôle des deux échantillons et moyenner les résultats. C'est ce que l'on appelle le *cross-fitting* Chernozhukov et al. (2018a).

Mise en pratique 9: *Sample-splitting*

Le notebook `DoubleSelection` montre comment mettre en oeuvre le sample-splitting pour l'estimateur de double-sélection.

3.6. Simulations

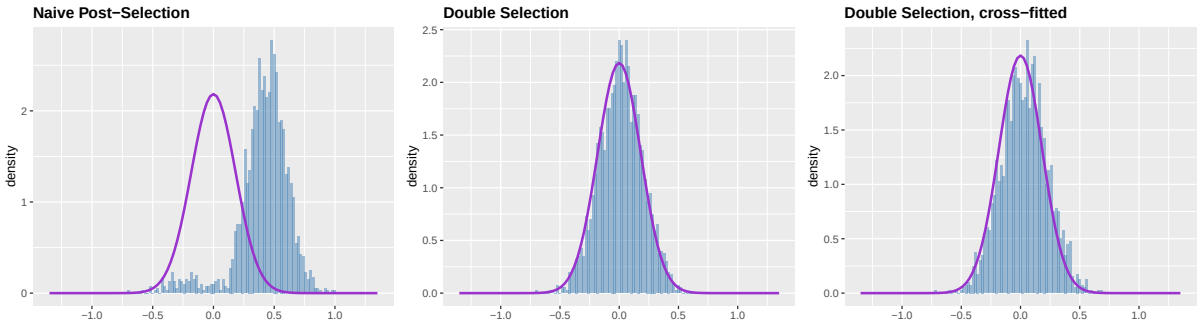
On considère le même processus générateur des données qu'à la section 3.2, mais cette fois avec un échantillon de 200 observations et $p = 300$ variables. On compare les trois estimateurs suivants :

1. l'estimateur (POST-LASSO) de τ_0 dans le modèle $Y_i = D_i\tau_0 + X_i'\beta_0 + \varepsilon_i$ où on l'on n'a pas pénalisé le coefficient associé à D dans la première étape (*estimateur naïf*),
2. l'estimateur de double-sélection (POST-DOUBLE-SELEC) sans partitionnement de l'échantillon (*double sélection*),
3. le même estimateur mais avec un partitionnement d'échantillon (*double sélection avec cross-fitting*).

La Figure 4 illustre les résultats obtenus. Comme prévu, l'estimateur naïf qui ne repose que sur seule étape de sélection de variable souffre d'un fort biais et n'affiche pas une distribution proche d'une gaussienne. Le taux de couverture est particulièrement mauvais, alors qu'il est très proche du taux espéré de .95 pour les deux autres estimateurs. Les

estimateurs reposant sur une double sélection, avec ou sans partitionnement de l'échantillon, affichent des performances comparables et des distributions qui se rapprochent d'une distribution normale.

FIGURE 4 – Distribution simulée des estimateurs de τ_0



Note : Chaque graphique représente la distribution simulée d'un estimateur de τ_0. Le graphique de gauche représente l'estimateur naïf avec une seule étape de sélection. Les deux autres estimateurs ne souffrent pas de biais de régularisation, c'est-à-dire d'un biais de variable omise. On notera que les résultats sont très comparables pour ces deux estimateurs. Résultats de 2000 tirages, pour un taille d'échantillon de 200 observations et de 150 variables explicatives.

TABLE 2 – Estimation de τ_0

	Estimateur :		
	Sélection naïve	*Double sélection*	*Double sélection, avec cross-fitting*
	(1)	(2)	(3)
Biais	0.410	0.015	0.032
$\sqrt{EQM}$	0.464	0.182	0.196
Taux de couverture	0.197	0.946	0.941
Note :	Biais, racine de l'erreur quadratique moyenne et taux de couverture pour les trois estimateurs décrits précédemment.		

Mise en pratique 10: Répliquer ces simulations

Le notebook `DoubleSelection` permet de reproduire cet exercice.

3.7. Application empirique : l'effet du diplôme sur le salaire

Nous appliquons les concepts décrits précédemment pour tenter de quantifier l'impact du niveau de diplôme sur le salaire grâce aux données de l'enquête emploi. Une autre

référence qui peut servir d'exemple est Bach et al. (2018), où les auteurs appliquent la méthode de la double-sélection pour évaluer l'hétérogénéité de l'écart salarial homme-femme aux États-Unis.

Pour cela, on utilise une variable catégorielle indiquant le niveau de diplôme en seize modalités. On souhaite alors estimer un modèle linéaire où la variable dépendante est le log du salaire mensuel, et les variables explicatives sont constituées par niveau de diplôme en variables binarisées (quinze variables au total) et d'autres variables de contrôles. On considère un total de 393 autres variables de contrôles, parmi lesquelles des déterminants usuels du salaire tels que le nombre d'heures travaillées, l'expérience au sein de l'entreprise, l'âge de l'individu, ainsi qu'une grande quantité de variables socio-démographiques et géographiques (*e.g.* sexe, origine sociale, statut matrimonial, nationalité, nombre d'enfants). Quand cela a un sens, on considère également des transformations pertinentes de ces variables (*e.g.* mise au carré). Finalement, on supprime automatiquement les variables engendrant de la multi-colinéarité.

Dans le cas présent, nous avons quinze paramètres d'intérêt, c'est-à-dire un par niveau de diplôme. On pourrait donc directement appliquer la méthode de la double-sélection en répliquant quinze fois l'étape 1 décrite à la Section 3.4, qui consiste à estimer une régression Lasso de la variable d'intérêt sur les variables de contrôles. Cependant, on préfère estimer une seule régression empilée au moyen d'un Group-Lasso. En effet, étant donnée que chacune des variables binaires représente un niveau de diplôme différent, on peut *a priori* penser que l'essentiel de leurs déterminants sont partagés et qu'en conséquence le schéma de sparsité est le même pour chaque équation. L'approche (GROUP-LASSO) permet de combiner l'information à travers chacune des équations pour aboutir à une sélection de variables plus pertinente. Cela est d'autant plus important qu'une approche où l'on analyse séparément chaque modalité de diplôme est susceptible d'échouer car certaines modalités de diplôme sont très rares dans la population (moins de 2 %). La méthodologie suivie pour adapter la méthode de la double-sélection à un cadre où le paramètre d'intérêt est multi-dimensionnel, incluant l'approche (GROUP-LASSO) utilisée, est décrite en annexe de ce document.

Les quatre trimestres des années 2017, 2018 et 2019 de l'enquête emploi sont utilisés, ce qui représente un total de 162,254 observations. Les écart-types sont estimés à partir

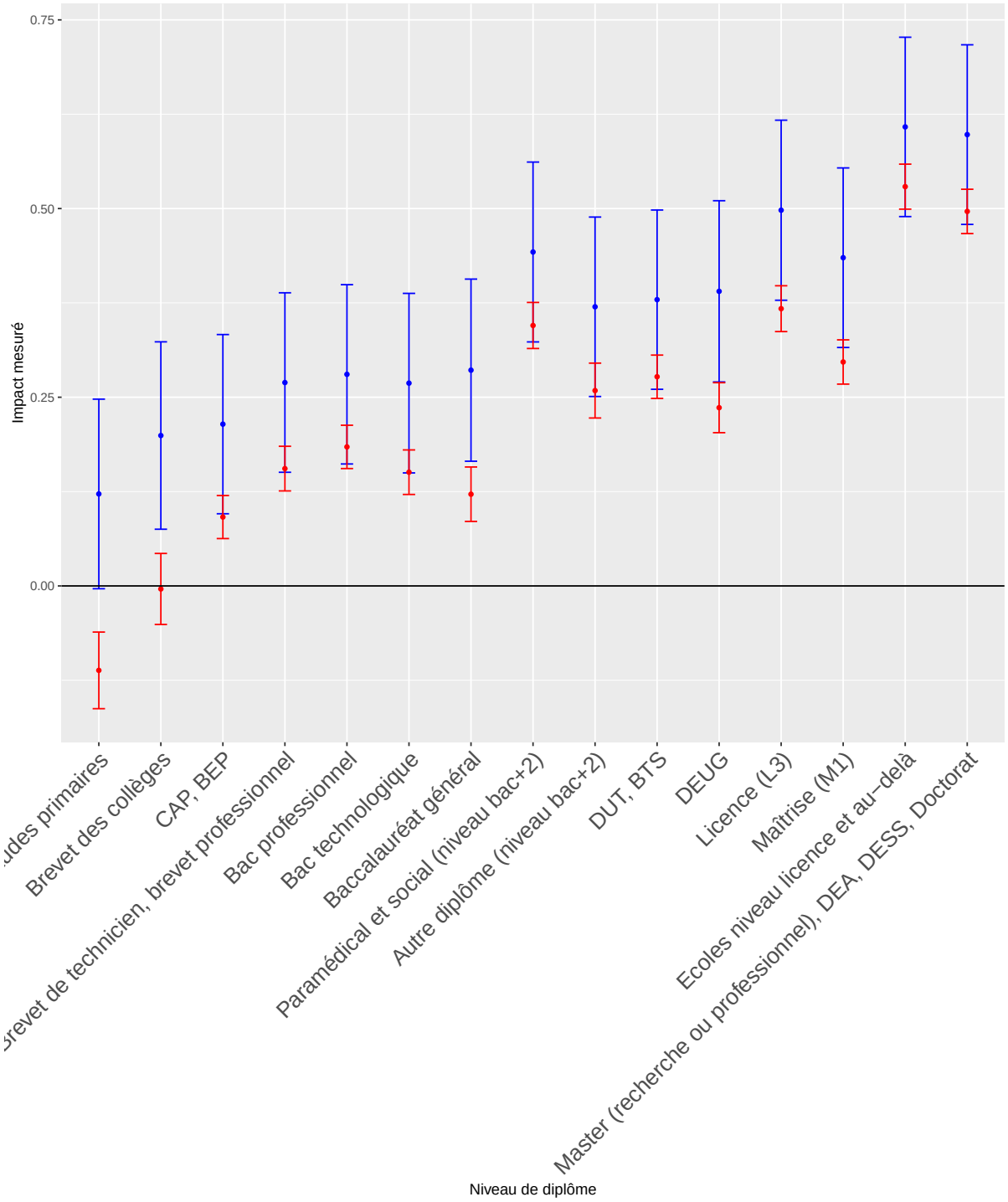
de clusters au niveau du ménage.

Les résultats sont présentés sur la Figure 5. On reporte les estimations issues de deux modèles différents : (i) le modèle complet incluant les 393 variables de contrôle estimé par MCO (en bleu), (ii) le modèle estimé par double-sélection (en rouge). La première étape de la double-sélection, celle qui consiste à sélectionner les variables de contrôle en lien avec le log du salaire, aboutit à sélectionner 71 variables, tandis que la seconde, celle qui consiste à sélectionner les variables de contrôle en lien avec le niveau de diplôme via un Group-Lasso, aboutit à la sélection de 68 variables. A total, 105 variables de contrôles uniques sont sélectionnées. Le niveau de pénalisation a été fixé de manière théorique, proportionnellement à $\sqrt{\log p/n}$ où p désigne le nombre de paramètres du modèle soit $393 + 1$ dans la première étape et $15 \times (393 + 1)$ dans la seconde. Les intervalles de confiance ont un niveau de 95% et sont construits à partir des écart-types robustes. On peut d'abord constater que la méthode de la double-sélection aboutit à un estimateur qui est bien plus précis que dans le modèle complet. En outre, les effets estimés sont globalement de magnitude moindre, si bien que certains intervalles de confiance pour les deux modèles ont une intersection vide, et on distingue quatre groupes de diplômes : (i) pas de diplôme / jusqu'au brevet des collèges, (ii) Bacs, niveau qui correspond à un écart de salaire de +15% par rapport à (i), (iii) de Bac +2 à maîtrise, ce qui correspond à un écart de salaire d'entre 25 et 35 %, et enfin, (iv) master 2, écoles et au delà, associés à un écart de salaire de +50%.

Mise en pratique 11: Répliquer cet exemple

Le notebook `ApplicationEnqueteEmploi.ipynb` permet de reproduire cet exemple empirique.

FIGURE 5 – Impact du niveau de diplôme sur le salaire



Note : Impact mesuré et intervalles de confiance asymptotiques à 95 % de chaque niveau de diplôme sur le log du salaire mensuel, avec écarts-types clusterisés au niveau du ménage. L'intervalle bleu est obtenu en estimant le modèle complet, sans aucune sélection de variable. L'intervalle rouge est obtenu par double-sélection.

4. Détecter l'hétérogénéité des effets

4.1. Motivation

L'étude des effets hétérogènes dans les expériences randomisées est particulièrement utile pour la validité externe, c'est-à-dire l'extension à une population différente de celle ayant fait l'objet de l'expérimentation. Elle est également importante pour l'étude de l'allocation optimale du traitement (*i.e.* répondre à la question : qui doit être traité ?). Pour cela il est nécessaire d'estimer l'effet conditionnel moyen du traitement (*Conditional Average Treatment Effect*, ou CATE), qui mesure l'espérance de l'effet du traitement conditionnellement à des caractéristiques individuelles observables. Cependant, le principal obstacle réside dans le fait qu'il faut fractionner l'échantillon selon plusieurs dimensions, ce qui aboutit à réduire la taille de l'échantillon et expose à un risque de sur-apprentissage (ou de *pêche aux p-values*).

Actuellement, l'approche par défaut est de mettre en place un plan de pré-analyse où le chercheur s'engage à étudier l'hétérogénéité uniquement selon certains dimensions pré-spécifiées. Cela aboutit à potentiellement gaspiller des données puisque l'on se prive de les explorer selon des dimensions non spécifiées à l'avance (Olken, 2015).

Une autre solution repose sur des algorithmes d'apprentissage automatique qui permettent de "laisser parler" les données et permettent de découvrir les dimensions selon lesquelles l'effet du traitement diffère même si on ne le suspectait pas auparavant. Cette solution présente deux problèmes : (i) les estimateurs issus d'algorithmes d'apprentissage automatique ne permettent généralement pas d'estimer le CATE de manière convergente sans hypothèses très fortes et non vérifiables, et (ii) ces algorithmes sont de toute façon mis en œuvre de manière sensiblement différente de leurs équivalents théoriques, ce qui ne permet donc pas de garantir les propriétés éventuellement établies par la théorie (par exemple, les paramètres de pénalisation, tels que le λ dans la section 2.5, sont choisis par validation croisée).

Chernozhukov et al. (2018b) adoptent une approche agnostique et générique en considérant les estimations issues de procédure de machine learning du CATE comme de simples *proxies* du véritable CATE et en mettant l'accent sur l'inférence sur des fonctionnelles du véritable CATE plutôt que sur le CATE lui-même (*e.g.* moyenne, moyenne

par groupe, corrélation avec le proxy obtenu par machine learning).

4.2. BLP, GATES, CLAN et cie.

Dans cette section, nous adoptons le cadre de résultats potentiels de Rubin (1974). Soit D la variable qui code le traitement, c'est-à-dire $D = 1$ pour un individu traité et $D = 0$ sinon. Soit Y_1 et Y_0 deux variables aléatoires représentant les résultats potentiels avec traitement et sans aucun traitement, respectivement. L'effet du traitement est $Y_1 - Y_0$. Soit X un vecteur aléatoire de caractéristiques individuelles observables. Le score de propension est défini par $p(X) := P[D = 1|X]$ et est supposé connu. On fait les deux hypothèses standard suivantes : (i) $D \perp\!\!\!\perp (Y_0, Y_1)|X$ *(Indépendance conditionnelle)* et (ii) $p(X) \in (0, 1)$ *(Support Commun)*. L'hypothèse (i) signifie que l'assignation au traitement est indépendante des résultats potentiels conditionnellement aux caractéristiques individuelles. L'hypothèse (ii) signifie que pour n'importe quelle strate de la population, on peut trouver un individu non-traité et un individu traité (afin de les comparer). En pratique, il pourra être nécessaire d'éliminer les observations pour lesquelles le score de propension est trop extrême de sorte à assurer que les distributions des traités et des non-traités couvrent le même support. Le résultat observé est $Y = Y_0 + D(Y_1 - Y_0)$, *i.e.* pour un individu donné, on n'observe jamais à la fois Y_1 et Y_0. Les hypothèses (i) et (ii) aboutissent au modèle :

$$Y = b_0(X) + D\tau_0(X) + U, \text{ où } E[U|X, D] = 0, \tag{5}$$

avec $b_0(X) = E[Y_0|X]$ et $\tau_0(X) = E[Y_1 - Y_0|X]$, le CATE.

On a à disposition le jeu de données iid suivant $Data = (Y_i, D_i, X_i)_{i=1}^n$ qui est aléatoirement partitionné entre un échantillon principal (M), $Data_M = (Y_i, D_i, X_i)_{i\in M}$, et un échantillon auxiliaire (A), $Data_A = (Y_i, D_i, X_i)_{i\in A}$. L'échantillon auxiliaire est utilisé pour estimer des prédicteurs ML de $b_0(X)$ et $\tau_0(X)$ (aussi appelés *predicteurs proxy*) :

$$x \to B(x) = B(x, Data_A) \text{ et } x \to T(x) = T(x, Data_A). \tag{6}$$

On considère ces fonctions comme fixes lorsque l'on travaille sur l'échantillon principal (on conditionne par rapport à $Data_A$). Par exemple, $B(\cdot)$ peut être construite comme la fonction de régression de Y sur X dans le groupe des non-traités, telle qu'estimée par

une méthode de Gradient Boosting. $T(\cdot)$ peut être construit comme la différence entre la fonction de régression de Y sur X dans le groupe des traités, telle qu'estimée par un Lasso et $B(\cdot)$. Ces prédicteurs proxy sont des prédicteurs potentiellement biaisés et bruités de $b_0(X)$ and $\tau_0(X)$. Nous ne sommes donc pas enclins à supposer qu'ils sont très performants. Peut-on tout de même dire quelque chose sur $b_0(X)$ et $\tau_0(X)$? Chernozhukov et al. (2018b) présentent trois quantités intéressantes.

4.2.1. Meilleur Prédicteur Linéaire du CATE

Le Best Linear Predictor (BLP) en anglais est une quantité qui va permettre de tester de façon jointe l'existence de l'hétérogénéité et la pertinente de l'algorithme de machine learning proposé.

On n'observe jamais $\tau_0(X)$, revanche, on possède (i) un proxy ML de $\tau_0(X)$, c'est-à-dire $T(X)$ et (ii) un signal sans biais de $\tau_0(X)$ qui est $w(X)(D - p(X))Y$ avec $w(X) = [p(X)(1 - p(X))]^{-1}$ au sens où $E[w(X)(D - p(X))Y|X] = \tau_0(X)$. A partir de ces deux quantités, on peut obtenir le meilleur prédicteur linéaire de $\tau_0(X)$ sachant $T(X)$, *i.e.* la projection L^2 de $\tau_0(X)$ sur une constante et $T(X)$:

$$\text{BLP}[\tau_0(X)|T(X)] = E[\tau_0(X)] + \frac{Cov(\tau_0(X), T(X))}{V(T(X))} (T(X) - E[T(X)]) . \tag{7}$$

Pour obtenir un estimateur de $Cov(\tau_0(X), T(X))/V(T(X))$, on régresse $w(X)(D-p(X))Y$ sur $T(X) - E[T(X)]$ au sein de l'échantillon principal. C'est-à-dire que l'on estime l'équation :

$$w(X)(D - p(X))Y = \beta_1 + \beta_2(T(X) - E[T(X)]) + \varepsilon,\ E[\varepsilon(1, (T(X) - E[T(X)]))'] = 0. \tag{BLP}$$

Théorème 1 (BLP 2) *Considérons $x \to T(x)$ comme une fonction non-aléatoire. Supposons que Y possède un moment d'ordre deux et que $V(T(X)) \neq 0$. Alors (β_1, β_2) définis par (BLP) sont les coefficients du meilleur prédicteur linéaire de $\tau_0(X)$ sachant $T(X)$:*

$$\beta_1 = E[\tau_0(X)] \text{ et } \beta_2 = \frac{Cov(\tau_0(X), T(X))}{V(T(X))}.$$

A quoi cela sert-il ? Considérons le cas polaire $\beta_2 = 0$ alors soit $\tau_0(X)$ est constant (il n'y a pas d'hétérogénéité de traitement) ou $T(X)$ n'est pas corrélé au CATE. En

conséquence, tester $\beta_2 = 0$ permet de vérifier l'existence d'hétérogénéité. Si nous rejetons cette hypothèse, cela signifie qu'il existe une certaine hétérogénéité et que $T(X)$ est corrélé au CATE. A noter que l'on peut ne pas rejeter l'hypothèse "$\beta_2 = 0$" dans le cas où il y a de l'hétérogénéité mais où le ML proxy ne la capte pas. Il peut donc être intéressant de considérer plusieurs algorithmes de machine learning.

De façon plus fondamentale, ce résultat permet de transformer un problème non supervisé – comme on n'observe jamais $\tau_0(X)$, il n'est pas possible de savoir si on le prédit bien – en un problème supervisé, puisque l'on est capable de calculer la covariance $Cov(\tau_0(X), T(X))$ et donc de chercher l'algorithme qui la maximise. Les auteurs proposent de choisir l'algorithme qui maximise la quantité

$$\Lambda = |\beta_2|^2 V(T(X)) = Corr(\tau_0(X), T(X))^2 V(\tau_0(X)),$$

que l'on peut calculer en pratique. Cela est équivalent à maximiser le R^2 de la régression de $\tau_0(X)$ sur $T(X)$ – régression qui est infaisable en pratique !

4.2.2. Autres quantités (GATES, CLAN)

Nous pouvons également diviser le support du prédicteur proxy en régions disjointes pour définir des groupes de réponse au traitement similaire et faire de l'inférence sur les effets attendus du traitement au sein de ces groupes :

$$E[\tau_0(X)|G_1] \leq \ldots \leq E[\tau_0(X)|G_K],$$

pour $G_k = \mathbf{1}\{\ell_{k-1} \leq T(X) < \ell_k\}$ avec $-\infty = \ell_0 \leq \ell_1 \leq \ldots \leq \ell_K = +\infty$. Pour cela, il suffit de faire la régression de $w(X)(D - p(X))Y$ sur $G_1, \ldots, G_K$. Les coefficients associés sont les effets moyens du traitements ordonnés (*Sorted Group Average Treatment Effect*, ou GATES), $E[\tau_0(X)|G_k]$.

Finalement, on peut aussi vouloir estimer les caractéristiques des groupes les plus et les moins affectés, en estimant de façon triviale – c'est-à-dire par des moyennes conditionnelles – $E[X_j|G_1]$ et $E[X_j|G_K]$ pour une caractéristique donnée X_j puisque qu'elle est observée. C'est ce que Chernozhukov et al. (2018b) appellent le CLAN pour *CLassification ANalysis*.

4.3. Inférence

Pour faire de l'inférence sur les paramètres présentés, c'est-à-dire β_2, les GATES et les paramètres issus de la CLAN, il faut prendre en compte deux sources d'incertitude : la première est l'aléa classique d'échantillonage, la seconde est l'aléa introduit par la division en deux des données (aléa "variationnel").

Prenons d'abord l'aléa standard. Soit une partition (A, M), le paramètre d'intérêt θ_A (*e.g.* $Cov(\tau_0(X), T(X))/V(T(X))$) dépend de $Data_A$ via le proxy machine learning $T(X)$. En conséquence, l'inférence doit être effectuée sur l'échantillon principal, conditionnellement à $Data_A$. Dans la plupart des cas, cela n'est pas un problème si bien que l'estimateur décrit plus haut va satisfaire un certain type de Théorème Central Limite conditionnel, $\widehat{\sigma}_A^{-1}\sqrt{|Data_M|}(\widehat{\theta}_A - \theta_A)|Data_A \xrightarrow{d} \mathcal{N}(0,1)$, où $|Data_M|$ désigne le nombre d'observations contenues dans $Data_M$. Par conséquent, conditionnellement à $Data_A$, $P[\theta_A \ni [L_A, U_A]|Data_A] = 1 - \alpha + o_P(1)$ pour $[L_A, U_A] := [\widehat{\theta}_A \pm \Phi^{-1}(1 - \alpha/2)\widehat{\sigma}_A/\sqrt{|Data_M|}]$ qui constitue donc un intervalle de confiance asymptotique de niveau α.

Il subsiste ensuite l'aléa de fractionnement de l'échantillon. Différentes partitions (A, M) des données produisent différentes cibles θ_A, donc conditionnellement à $Data_A$, θ_A est une variable aléatoire car ce paramètre dépend de (A, M). De même pour $\widehat{\theta}_A$. Pour prendre en compte cette incertitude, on reporte l'estimateur $\widehat{\theta} := Med[\widehat{\theta}_A|Data_A]$ défini comme la médiane des estimateurs obtenus chacun avec une répartition de données différente et cela pour un nombre B de partitions. L'intervalle de confiance correspondant au niveau $1-2\alpha$ est donné par $[Med[L_A|Data_A], Med[U_A|Data_A]]$ où $[L_A, U_A]$ est l'intervalle de confiance asymptotique habituel de niveau $1 - \alpha$ construit sur $\hat{\theta}_A$. Ainsi pour obtenir le niveau de confiance habituel de 95 %, il faut avoir, pour chaque partition, construit des intervalles à 97.5%. On note donc que l'on paye le prix du fractionnement de l'échantillon en transformant l'intervalle à 97.5 % sur chaque partition, en intervalle de confiance à 95 %.

De façon analogue, le fractionnement occasionne un coût de même nature pour les tests. Prenons par exemple le test de l'hypothèse H_0 :"$\beta_2 = 0$" versus l'hypothèse alternative H_1 :"$\beta_2 \neq 0$". Soit p_A la p–value associée à ce test sur l'échantillon principal, pour

une partition des données fixe, $p_A = \Phi\left(\widehat{\sigma}_A^{-1}\sqrt{|Data_M|}\widehat{\beta}_A\right)$. On rejette H_0 si pour au moins la moitié des partitions, la p-value p_A est plus petite que $\alpha/2$, c'est-à-dire que le test de niveau α s'écrit :

$$\phi_\alpha = \mathbf{1}\left\{\text{Med}(p_A) \leq \alpha/2\right\}.$$

Ce coût est certes gênant, mais on ne peut s'en abstraire car le partitionnement de l'échantillon original permet de considérer des estimateurs issus de procédure de machine learning sans avoir de garantie théorique sur leur fonctionnement.

4.4. Application empirique

Afin d'illustrer la mise en application du Generic Machine Learning, nous revisitons l'article de Bolsen et al. (2014), qui étudie la propension des gens qui votent régulièrement à contribuer aux biens publics, dans le cadre d'une expérimentation aléatoire à grande échelle. Durant une période de sécheresse au Sud-Est des Etats-Unis (Atlanta, Georgie) en 2007, 35 000 ménages, tirés aléatoirement parmi 106 000, ont reçu des prospectus les incitant à économiser l'eau courante. Il y avait trois sortes de prospectus : (i) un prospectus listant des astuces pour conserver l'eau, (ii) ce même prospectus augmenté d'un message citoyen encourageant les ménages à adopter un comportement "responsable" et une attitude solidaire pour utiliser l'eau courante "judicieusement", (iii) le même prospectus qu'au deuxième traitement, mais incluant en plus une information sur le quantile de consommation d'eau du ménage l'année précédente. Par souci de simplification, on considère ces trois traitements de façon groupée, et souhaite évaluer leur impact versus ne pas recevoir de prospectus concernant l'économie d'eau courante. La variable d'intérêt est la consommation d'eau entre juin et septembre 2007 en milliers de galons.

L'idée de Bolsen et al. (2014) est d'étudier la mesure selon laquelle les ménages dont les individus votent le plus ou appartiennent à une certaine couleur politique sont plus sensibles aux messages incitant à économiser l'eau. Il s'agit donc d'estimer des effets hétérogènes, en particulier selon trois variables : (i) la fréquence de vote au niveau du ménage définie comme le rapport entre la somme du nombre des fois où un individu du ménage est allé voter sur la somme du nombre des fois où un individu du ménage pouvait voter au cours des élections qui ont eu lieu entre 1990 et 2008, (ii) une variable indicatrice d'un ménage considéré comme Républicain, c'est-à-dire si le nombre de fois où

tous les membres du ménage ont voté dans une primaire républicaine est plus grand que le nombre de fois où tous les membres du ménage ont voté dans une primaire démocrate, (iii) une variable indicatrice d'un ménage considéré comme Démocrate, calculé de la façon symétrique à la précédente.

Outre ces trois variables, les variables de contrôle considérées sont : une indicatrice qui vaut un si aucun des membres du ménages n'est inscrit sur une liste électorale, la consommation d'eau courante entre juin et septembre 2006, la consommation d'eau courante entre avril et mai 2007, la valeur marchande de la résidence, l'âge de la résidence, une variable indicatrice qui vaut un si le ménage est propriétaire, ainsi qu'une variable indicatrice qui vaut un si le propriétaire a plus de 65 ans. Les régressions pour obtenir les coefficients d'intérêt sont faites au niveau du ménage, et incluent des effets fixes par circuit de compteur d'eau, niveau auquel a eu lieu la randomization.

Nous considérons quatre algorithmes de machine learning : un modèle de régression (ELASTIC-NET), un algorithme de gradient boosting machine (Friedman, 2001) pour un modèle de régression, un réseau de neurones avec une étape préalable d'analyse en composantes principales, ainsi qu'une forêt aléatoire, que nous entraînons sur une moitié de l'échantillon au moyen d'une validation croisée en deux parties, répétée deux fois. A chaque fois, on partitionne aléatoirement l'échantillon par strates de traitement, afin d'avoir un échantillon suffisamment grand pour entraîner un modèle sur la demi-population traitée et la demi-population non-traitée. On considère au total 30 partitions différentes de l'échantillon pour prendre en compte l'aléa variationnel. La Table 3 reporte la statistique $\Lambda = Corr(\tau_0(X), T(X))^2 V(\tau_0(X))$ permettant de juger de l'existence d'hétérogénéité et de la pertinence d'un algorithme, pour chacun des algorithmes. C'est le gradient boosting machine qui est le plus corrélé à l'effet du traitement conditionnel, autrement dit il capture le mieux l'hétérogénéité des effets attendus. Pour la suite, nous reportons uniquement les résultats du gradient boosting machine et du réseau de neurones, ce dernier algorithme étant performant sur une autre mesure non évoquée dans ce document.

La Table 4 reporte l'effet moyen du traitement, estimé à environ -952 gallons d'eau sur le troisième trimestre 2007 par ménage en moyenne, ainsi que le coefficient β_2 permettant de tester l'existence d'effets hétérogènes. On rejette le test de H_0 : "$\beta_2 = 0$" à 1% pour

TABLE 3 – Classement des algorithmes – Λ

Λ	Elastic Net	Gradient Boosting Machine	Réseau de Neurones	Forêt Aléatoire
Conso. Eau (T3 2007)	1.137	1.165	1.000	0.933

Note : Classement des algorithmes selon leur corrélation avec le véritable effet du traitement, $\Lambda = |\beta_2|^2 V(T(X)) = Corr(\tau_0(X), T(X))^2 V(\tau_0(X))$. On a considéré quatre algorithmes : un modèle de régression (ELASTIC-NET), un algorithme de gradient boosting machine pour un modèle de régression, un réseau de neurones avec une étape préalable d'analyse en composantes principales, ainsi qu'une forêt aléatoire. Résultats obtenus sur 30 partitions différentes des données.

le gradient boosting, ce qui confirme qu'il existe de l'hétérogénéité et que cet algorithme permet de la capturer en partie. Cela n'est pas le cas pour le réseau de neurone où l'on ne rejette pas H_0, ce qui laisse penser que le réseau de neurones capture mal l'hétérogénéité présente ici.

TABLE 4 – Meilleur prédicteur linéaire du CATE – BLP

	Gradient Boosting Machine		Réseau de Neurones	
	ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
Conso. Eau	-0.952	0.116	-0.902	0.058
(T3 2017)	(-1.278,-0.631)	(0.068,0.167)	(-1.233,-0.576)	(-0.031,0.146)
	[0.000]	[0.000]	[0.000]	[0.441]

Note : $\beta_1 = E[\tau_0(X)]$ et $\beta_2 = Cov(\tau_0(X), T(X))/V(T(X))$ pour l'agorithme de gradient boosting et le réseau de neurones. On reporte à chaque fois l'estimateur, l'intervalle de confiance à 90 % et la p-value. Résultats obtenus sur 30 partitions différentes des données.

La Table 5 reporte l'effet moyen du traitement par groupes classés selon les 20% les plus ou les 20% les moins affectés d'après les deux algorithmes de machine learning, ce qui est également visible dans le Figure 6. Pour l'ensemble de la population le traitement réduit bien la consommation d'eau, puisque même pour les moins affectés on rejette l'hypothèse de la nullité de l'effet à 5%, du moins quand on considère l'agorithme de gradient boosting. En revanche, il ne semble pas que l'effet moyen entre les 20 % les moins affectés et les 20% les plus affectés soit statistiquement significatif. Cela laisse penser que la variance de l'hétérogénéité est relativement modérée.

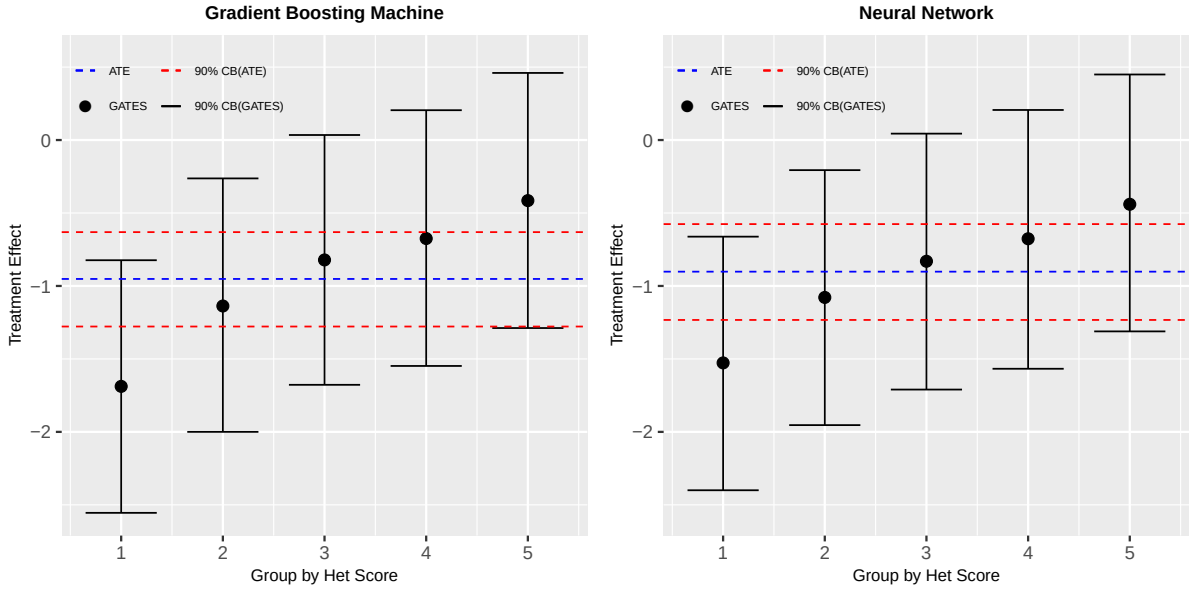
Finalement, la Table 6 reporte la CLAN réalisée selon les variables de contrôles utilisées dans le modèle. Ce sont principalement les trois premières qui nous intéressent. Plus précisément, il apparaît que les ménages les plus affectés par la politique votent en

TABLE 5 – Effet du traitement par groupe – GATES

	Gradient Boosting Machine			Réseau de Neurones		
	Moins affectés	Plus affectés	Différence	Moins affectés	Plus affectés	Différence
Conso. Eau	-0.953	-1.688	0.700	-0.707	-1.483	0.780
(T3 2007)	(-1.685,-0.217)	(-2.417,-0.960)	(-0.302,1.722)	(-1.459,0.050)	(-2.235,-0.730)	(-0.290,1.858)
	[0.023]	[0.000]	[0.342]	[0.135]	[0.000]	[0.307]

Note : $E[\tau_0(X)|G_{20}]$ et $E[\tau_0(X)|G_{80}]$, les GATES pour les 20% les moins affectés et les 20% les plus affectés selon un classement obtenu par les ML proxy issus du gradient boosting machine et du réseau de neurones. On reporte à chaque fois l'estimateur, l'intervalle de confiance à 90 % et la p-value. Résultats obtenus sur 30 partitions différentes des données.

FIGURE 6 – Effet moyen et effet par groupes – ATE et GATES



Note : Effet du traitement moyen (ATE) et $E[\tau_0(X)|G_{20}], \ldots E[\tau_0(X)|G_{80}]$, les GATES pour des groupes de 20 % de l'échantillon ordonné selon la valeur des ML proxies issus du gradient boosting machine et du réseau de neurones. On reporte à chaque fois l'estimateur et l'intervalle de confiance à 90 %. Résultats obtenus sur 30 partitions différentes des données.

moyenne plus que les moins affectés, mais également que la population générale (10%). En outre, les plus affectés par le traitement sont aussi plus inscrits sur les listes électorales. Cela semble confirmer l'hypothèse d'un sens civique qui dictera à la fois la participation au vote et la préservation du bien commun.

Qu'en est-il de la couleur politique ? Les ménages considérés comme démocrates sont également sur-représentés parmi les ménages qui répondent au traitement, 20 % contre 17 % parmi les ménages peu affectés par le traitement. A noter que puisque les groupes

sont de taille égales, cela se traduit également par une proportion plus fortes de ménages sensibles au traitement chez les démocrates comparativement aux ménages qui le sont faiblement (d'après la règle de Bayes, $P[Democrate|G_5]/P[Democrate|G_1] = P[G_5|Democrate]/P[G_1|Democrate]$). Cela n'est pas le cas pour les ménages qui sont de façon prédominante républicains puisque qu'ils sont à peu près aussi présents chez les moins affectés que chez les plus affectés. Ce résultat contredit en partie le résultat initial de Bolsen et al. (2014) qui ne reportait aucune différence entre la sensibilité des Démocrates et des Républicains aux messages incitant à préserver l'eau courante.

Mise en pratique 12: Répliquer cet exemple

Le code `GenericML-example` permet de reproduire cet exemple. Attention : la méthode du Generic Machine Learning étant très récente d'une part, et très générale d'autre part, il n'existe pas de package permettant d'automatiser facilement sa mise en oeuvre. Ce code repose sur l'utilisation intensive du package `caret` pour mettre en oeuvre les méthodes de machine learning nécessaires. Le temps d'exécution du code se compte en heures, même parallélisé. Il est probable qu'il soit plus efficient, à la fois en termes de temps de calcul et de propreté du code, de mettre en oeuvre cette méthode en `python`. Les données de Bolsen et al. (2014) sont disponibles en libre accès ici : https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QJD1OW.

Références

Abadie, A. and Kasy, M. (2018). Choosing among regularized estimators in empirical economics : The risk of machine learning. *The Review of Economics and Statistics*, 0(ja) :null.

Abrams, D. S., Bertrand, M., and Mullainathan, S. (2012). Do judges vary in their treatment of race ? *The Journal of Legal Studies*, 41(2) :347–383.

Bach, P., Chernozhukov, V., and Spindler, M. (2018). Closing the U.S. gender wage gap requires understanding its heterogeneity. *arXiv e-prints*, page arXiv :1812.04345.

Beck, A. and Teboulle, M. (2014). A fast dual proximal gradient algorithm for convex minimization and applications. *Operations Research Letters*, 42(1) :1–6.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2) :521–547.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018). High-Dimensional Econometrics and Regularized GMM. *arXiv e-prints*, page arXiv :1806.01888.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1) :233–298.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2) :29–50.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2) :608–650.

Bléhaut, M., D'Haultfoeuille, X., L'Hour, J., and Tsybakov, A. B. (2020). An alternative to synthetic control for models with many covariates under sparsity. *arXiv e-prints*.

Bolsen, T., Ferraro, P. J., and Miranda, J. J. (2014). Are voters more likely to contribute to other public goods ? evidence from a large-scale randomized policy experiment. *American Journal of Political Science*, 58(1) :17–30.

Chandrasekhar, A. (2016). Econometrics of network formation. *The Oxford Handbook of the Economics of Networks*, pages 303–357.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5) :261–65.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1) :C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2018b). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Working Paper 24678, National Bureau of Economic Research.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments. *American Economic Review*, 105(5) :486–90.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference : An elementary, general approach. *Annu. Rev. Econ.*, 7(1) :649–688.

Chetty, R. and Hendren, N. (2018). The Impacts of Neighborhoods on Intergenerational Mobility II : County-Level Estimates*. *The Quarterly Journal of Economics*, 133(3) :1163–1228.

de Paula, A. (2015). Econometrics of network models. CeMMAP working papers CWP52/15, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1) :1 – 23.

Ferrara, L. and Simoni, A. (2019). When are Google data useful to nowcast GDP ? An approach via pre-selection and shrinkage. Working Papers 2019-04, Center for Research in Economics and Statistics.

Friedman, J. H. (2001). Greedy function approximation : A gradient boosting machine. *Ann. Statist.*, 29(5) :1189–1232.

Gaillac, C. and L'Hour, J. (2019). Machine learning for econometrics. Notes de cours, ENSAE Paris.

Graham, B. S. (2019). Network data. Working Paper 26577, National Bureau of Economic Research.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning : Data mining, Inference and Prediction.* Springer, 2nd edition.

Jurafsky, D. and Martin, J. H. (2019). *Speech and Language Processing (3rd ed. draft).* Prentice Hall PTR, USA, 3rd edition.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1) :31–43.

Leeb, H. (2006). *The distribution of a linear predictor after model selection : Unconditional finite-sample distributions and asymptotic approximations*, volume Number 49 of *Lecture Notes–Monograph Series*, pages 291–311. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference : Facts and fiction. *Econometric Theory*, null :21–59.

L'Hour, J. (2019). *Policy evaluation, high-dimension and machine learning.* Thèses, Université Paris-Saclay.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3) :61–80.

Plomin, R. (2018). *Blueprint : How DNA Makes Us Who We Are.* Penguin Books Limited.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5) :688.

Sala-I-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, 87(2) :178–183.

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4) :481–493.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2) :614–645.

Yang, Y. (2005). Can the strengths of aic and bic be shared ? a conflict between model indentification and regression estimation. *Biometrika*, 92(4) :937–950.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7 :2541–2563.

Annexes

Extension de l'immunisation au cas où la variable d'intérêt est de dimension quelconque.

Cette partie est une annexe technique à la Section 3.7. On explique comment étendre la méthode de la double-sélection à un cas où le paramètre d'intérêt est de dimension quelconque. La portée de cette section est générale, mais on l'appliquera principalement au cas le paramètre d'intérêt mesure l'effet des différentes modalités d'une variable catégorielle (*e.g.* niveau de diplôme) ce qui justifie l'utilisation d'un Group-Lasso dans une étape intermédiaire.

Supposons que l'on considère le modèle :

$$Y_i = X_{i,1}'\tau_0 + X_{i,2}'\beta_0 + \varepsilon_i,$$

avec $X_{i,1}$ un vecteur aléatoire de dimension K et $X_{i,2}$ un vecteur aléatoire de dimension p. Par exemple, $X_{i,1}$ peut être une variable aléatoire discrète à plusieurs modalités que l'on encode en "one-hot" en enlevant une modalité pour éviter les problèmes de multi-colinéarité. τ_0 est toujours le paramètre d'intérêt alors que β_0 est le paramètre de nuisance. Le score orthogonalisé pour estimer τ_0 sera dans ce cas

$$E\left[\left(Y_i - X_{i,1}'\tau_0 - X_{i,2}'\beta_0\right)\left(X_1 - X_{i,2}'\Gamma_0\right)\right] = 0,$$

avec β_0 (dimension p) et Γ_0 (dimension $p \times K$) des paramètres de nuisance. Autrement dit :

$$\tau_0 = E\left[\left(X_1 - X_{i,2}'\Gamma_0\right)\left(X_1 - X_{i,2}'\Gamma_0\right)'\right]^{-1} E\left[\left(X_1 - X_{i,2}'\Gamma_0\right)\left(Y_i - X_{i,2}'\beta_0\right)\right].$$

Γ_0 est tel que :

$$E\left[\left(X_{i,1} - X_{i,2}'\Gamma_0\right)X_{i,2}'\right] = 0,$$

autrement dit la k-ème colonne de Γ_0 est constituée par les coefficients de la régression de $X_{i,1,k}$ sur $X_{i,2}$.

Dans le cas où $X_{i,1}$ encode les modalités d'une variable catégorielle, on peut estimer Γ_0 par (GROUP-LASSO). Par exemple si $X_{i,1}$ est un vecteur aléatoire indiquant le niveau de diplôme de l'individu i pour différentes modalités, on peut penser que les facteurs affectant l'assignation à une modalité jouent également sur l'affectation aux autres modalités. Cela constitue au total p groupes de K variables :

$$\begin{aligned}\widehat{\Gamma}^{GL}(\lambda) &= \arg\min_{\Gamma} \frac{1}{K \times n} \sum_{i=1}^{n} \left\|X_{i,1} - X_{i,2}'\Gamma\right\|^2 + \lambda \sum_{j=1}^{p} \left\|\Gamma_{j,.}\right\|_2 \\ &= \arg\min_{\Gamma} \frac{1}{K \times n} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(X_{i,1,k} - X_{i,2}'\Gamma_{.,k}\right)^2 + \lambda \sum_{j=1}^{p} \left\|\Gamma_{j,.}\right\|_2,\end{aligned}$$

où $\Gamma_{j,.}$ (resp. $\Gamma_{.,k}$) désigne la j-ème ligne (k-ème colonne) de Γ. β_0 est défini par :

$$E\left[\left(Y_i - X_{i,1}'\tau_0 - X_{i,2}'\beta_0\right)X_{i,2}\right] = 0.$$

Et on peut estimer β_0 par un LASSO classique.

L'estimateur vaut alors

$$\check{\tau} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(X_1 - X_{i,2}'\widehat{\Gamma}^{GL}(\lambda)\right)\left(X_1 - X_{i,2}'\widehat{\Gamma}^{GL}(\lambda)\right)'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_1 - X_{i,2}'\widehat{\Gamma}^{GL}(\lambda)\right)\left(Y_i - X_{i,2}'\widehat{\beta}\right)\right],$$

et est tel que :

$$\sqrt{n}\left(\check{\tau}-\tau_0\right) \xrightarrow{d} \mathcal{N}(0,\Sigma),$$

où

$$\begin{aligned}\Sigma :=& E\left[\left(X_{i,1}-X_{i,2}'\Gamma_0\right)\left(X_{i,1}-X_{i,2}'\Gamma_0\right)'\right]^{-1} E\left[\left(Y_i-X_{i,2}'\beta_0-X_{i,1}'\tau_0\right)^2\left(X_{i,1}-X_{i,2}'\Gamma_0\right)\left(X_{i,1}-X_{i,2}'\Gamma_0\right)'\right] \\ & \times E\left[\left(X_{i,1}-X_{i,2}'\Gamma_0\right)\left(X_{i,1}-X_{i,2}'\Gamma_0\right)'\right]^{-1}.\end{aligned}$$

□

Preuve du Théorème 1

On montre simplement que $\beta_2 = Cov(\tau_0(X), T(X))/V(T(X))$ puisque la preuve pour β_1 est très similaire. Les équations normales qui définissent (β_1, β_2) dans l'équation (BLP) donnent pour β_2 :

$$\beta_2 = \frac{Cov(w(X)(D-p(X))Y, T(X)-E(T(X)))}{V(T(X)-E(T(X)))}.$$

Le dénominateur est égal à $V(T(X))$. Puisque $T(X)-E(T(X))$ est de moyenne nulle, le numérateur est :

$$Cov(w(X)(D-p(X))Y, T(X)-E(T(X))) = E[w(X)(D-p(X))Y(T(X)-E(T(X)))].$$

Notons que $E[w(X)(D-p(X))D|X] = [w(X)(D-p(X))^2|X] = 1$ puisque $D|X \sim \mathcal{B}(p(X))$. Puisque $Y = b_0(X) + D\tau_0(X) + U$, la loi des espérances itérées :

$$E[w(X)(D-p(X))b_0(X)(T(X)-E(T(X)))] = E[w(X)b_0(X)(T(X)-E(T(X)))\underbrace{E[D-p(X)|X]}_{=0}] = 0,$$

$$E[w(X)(D-p(X))D\tau_0(X)(T(X)-E(T(X)))] = E[\tau_0(X)(T(X)-E(T(X)))] = Cov(\tau_0(X), T(X)),$$

$$E[w(X)(D-p(X))U(T(X)-E(T(X)))] = E[w(X)(D-p(X))\underbrace{E[U|X,D]}_{=0}(T(X)-E(T(X)))] = 0,$$

prouvant que $\beta_2 = Cov(\tau_0(X), T(X))/V(T(X))$. □

TABLE 6 – Caractéristiques moyennes des groupes – CLAN

	Gradient Boosting Machine			Réseau de Neurones		
	Moins affectés	Plus affectés	Différence	Moins affectés	Plus affectés	Différence
Fréquence vote	0.098	0.120	-0.017	0.096	0.121	-0.024
	(0.096,0.100)	(0.118,0.122)	(-0.020,-0.014)	(0.094,0.098)	(0.119,0.123)	(-0.027,-0.021)
	-	-	[0.000]	-	-	[0.000]
Démocrate	0.166	0.204	-0.044	0.147	0.242	-0.087
	(0.159,0.174)	(0.197,0.212)	(-0.054,-0.033)	(0.139,0.154)	(0.234,0.249)	(-0.097,-0.077)
	-	-	[0.000]	-	-	[0.000]
Républicain	0.408	0.448	-0.012	0.409	0.390	0.008
	(0.399,0.418)	(0.438,0.458)	(-0.025,0.001)	(0.400,0.419)	(0.380,0.399)	(-0.006,0.021)
	-	-	[0.159]	-	-	[0.531]
Non inscrit	0.148	0.089	0.058	0.174	0.092	0.068
sur les listes	(0.142,0.154)	(0.083,0.095)	(0.049,0.066)	(0.167,0.181)	(0.085,0.098)	(0.058,0.077)
	-	-	[0.000]	-	-	[0.000]
Conso. Eau (T3 2006)	65.23	89.82	-25.56	55.41	84.08	-31.35
	(64.24,66.21)	(88.86,90.78)	(-26.99,-24.13)	(54.43,56.40)	(83.09,85.08)	(-32.75,-29.95)
	-	-	[0.000]	-	-	[0.000]
Conso. Eau (T2 2007)	17.47	23.57	-6.150	14.14	23.04	-8.043
	(17.17,17.77)	(23.27,23.86)	(-6.564,-5.735)	(13.85,14.43)	(22.74,23.35)	(-8.458,-7.628)
	-	-	[0.000]	-	-	[0.000]
Valeur de la Résidence	279 975	354 577	-72 506	246 457	338 013	-89 671
	(275 861,284 090)	(350 235,358 919)	(-78 360,-66 652)	(24 2499,25 0415)	(333 817,342 208)	(-95 284,-84 057)
	-	-	[0.000]	-	-	[0.000]
Age de la résidence	19.81	18.64	1.095	21.55	19.12	2.470
	(19.55,20.06)	(18.39,18.89)	(0.735,1.451)	(21.29,21.81)	(18.86,19.39)	(2.103,2.836)
	-	-	[0.000]	-	-	[0.000]
Propriétaire	0.770	0.890	-0.131	0.710	0.905	-0.205
	(0.763,0.777)	(0.883,0.897)	(-0.141,-0.122)	(0.703,0.717)	(0.898,0.912)	(-0.215,-0.195)
	-	-	[0.000]	-	-	[0.000]
Personne âgée	0.073	0.093	-0.018	0.072	0.131	-0.046
	(0.067,0.078)	(0.087,0.098)	(-0.025,-0.010)	(0.066,0.078)	(0.125,0.136)	(-0.054,-0.038)
	-	-	[0.000]	-	-	[0.000]

Note : $E[X_j|G_{20}]$ et $E[X_{80}|G_K]$, les CLAN, c'est-à-dire les moyennes des caractéristiques pour les 20% les moins affectés et les 20 % les plus affectés. On reporte à chaque fois l'estimateur, l'intervalle de confiance à 90 % et la p-value. Résultats obtenus sur 30 partitions différentes des données.