Économétrie & *Machine Learning*

Arthur Charpentier

Université de Rennes 1 & CREM

Emmanuel Flachaire

Aix-Marseille Université, AMSE, CNRS & EHESS

et

Antoine Ly

Université Paris-Est

Résumé

L'économétrie et l'apprentissage machine semblent avoir une finalité en commun: construire un modèle prédictif, pour une variable d'intérêt, à l'aide de variables explicatives (ou *features*). Pourtant, ces deux champs se sont développés en parallèle, créant ainsi deux cultures différentes, pour paraphraser Breiman (2001a). Le premier visait à construire des modèles probabilistes permettant de décrire des phénomèmes économiques. Le second utilise des algorithmes qui vont apprendre de leurs erreurs, dans le but, le plus souvent de classer (des sons, des images, etc). Or récemment, les modèles d'apprentissage se sont montrés plus efficaces que les techniques économétriques traditionnelles (avec comme prix à payer un moindre pouvoir explicatif), et surtout, ils arrivent à gérer des données beaucoup plus volumineuses. Dans ce contexte, il devient nécessaire que les économètres comprennent ce que sont ces deux cultures, ce qui les oppose et surtout ce qui les rapproche, afin de s'approprier des outils développés par la communauté de l'apprentissage statistique, pour les intégrer dans des modèles économétriques.

JEL Code: C18; C52; C55
Key-words: apprentissage; données massives; économétrie; modélisation; moindres carrés;

Juillet 2017

Contents

1 Introduction

L'utilisation de techniques quantitatives en économie remonte probablement au 16ème siècle, comme le montre Morgan (1990). Mais il faudra attendre le début du XXième siècle pour que le terme "économétrie" soit utilisé pour la première fois, donnant naissance à l'*Econometric Society* en 1933. Les techniques de *machine learning* (apprentissage machine) sont plus récentes. C'est à Arthur Samuel, considéré comme le père du premier programme d'auto-apprentissage, que l'on doit le terme "*machine learning*" qu'il définit comme "*a field of study that gives computer the ability without being explicitly programmed*". Parmi les premières techniques, on peut penser à la théorie des assemblées de neurones proposée dans Hebb (1949) (qui donnera naissance au *perceptron* dans les années 1950, puis aux réseaux de neurones) dont Widrow & Hoff (1960) montreront quinze ans plus tard les liens avec les méthodes de moindres carrées, aux SVM (*support vector machine*) et plus récemment aux méthodes de *boosting*. Si les deux communautés ont grandi en parallèle, les données massives imposent de créer des passerelles entre les deux approches, en rapprochant les "deux cultures" évoquées par Breiman (2001a), opposant la statistique mathématique (que l'on peut rapprocher de l'économétrie traditionnelle, comme le note Aldrich (2010)) à la statistique computationnelle, et à l'apprentissage machine de manière générale.

1.1 La Modélisation économétrique

L'économétrie et les techniques d'apprentissage statistique supervisé sont proches, tout en étant très différentes. Proches au départ, car toutes les deux utilisent une base (ou un tableau) de données, c'est à dire des observations $\{(y_i, \boldsymbol{x}_i)\}$, avec $i = 1, \cdots, n$, $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^p$ et $y_i \in \mathcal{Y}$. Si y_i est qualitative, on parlera d'un problème de classification, et dans le cas contraire, d'un problème de régression. Proches à l'arrivée, car dans les deux cas, on cherche à construire un "modèle", c'est à dire une fonction $m : \mathcal{X} \mapsto \mathcal{Y}$ qui sera interprétée comme une prévision.

Mais entre le départ et l'arrivée, il existe de réelles différences. Historiquement, les modèles économétriques s'appuient sur une théorie économique, avec le plus souvent des modèles paramétriques. On a alors recours aux outils classiques de l'inférence statistique (comme le maximum de vraisemblance, ou la méthode des moments) pour estimer les valeurs d'un vecteur de paramètres $\boldsymbol{\theta}$, dans un modèle paramétrique $m_{\boldsymbol{\theta}}(\cdot)$, par une valeur $\widehat{\boldsymbol{\theta}}$. Comme en statistique, avoir des estimateurs sans biais est important car on peut quantifier une borne inférieure pour la variance (borne de Cramér-Rao). La théorie asymptotique joue alors un rôle important (développements de Taylor, loi des grands nombres, et théorème central limite). En apprentissage statistique, en revanche, on construit souvent des modèles non-paramétriques, reposant presque exclusivement sur les données (i.e. sans hypothèse de distribution), et les méta-paramètres utilisés (profondeur de l'arbre, paramètre de pénalisation, etc) sont optimisés par validation croisée.

Au delà des fondements, si l'économétrie étudie abondamment les propriétés (souvent asymptotiques) de $\widehat{\boldsymbol{\theta}}$ (vu comme une variable aléatoire, grâce à la représentation stochastique sous-jacente), l'apprentissage statistique s'intéresse davantage aux propriétés du modèle optimal $m^\star(\cdot)$ (suivant un critère qui reste à définir), voire simplement $m^\star(\boldsymbol{x}_i)$ pour quelques observations i jugées d'intérêt (par exemple dans une population de test). Le problème de choix de modèle est aussi vu sous un angle assez différent. Suivant la loi de Goodhart ("*si une mesure devient un objectif, elle cesse d'être une mesure*"), les économètres utilisent des critères de type AIC ou BIC pour choisir un modèle optimal (pénalisant la qualité d'ajustement d'un modèle par sa complexité, ex-post, lors de la phase de validation ou de choix), alors qu'en apprentissage statistique, c'est la fonction objectif qui tiendra compte d'une pénalisation, comme pour le LASSO, ressemblant à une forme de pénalisation ex-ante.

1.2 Applications

Avant de revenir sommairement sur l'évolution des modèles économétriques, c'est à Francis Galton que l'on doit le terme "régression", comme le rappelle Koenker (1998). Si le terme est parfois devenu synonyme de "modèle économétrique", il avait été introduit dans le contexte de "*regression towards mediocraty inhereditary stature*", pour reprendre le titre de l'article paru en 1886. Galton utilisait un modèle linéaire pour modéliser la taille moyenne d'un garçon (à l'âge adulte) en fonction de la taille de son père. Si cette technique de régression était connue par les économistes, il a fallu attendre les années 1930 pour voir surgir le concept de "modèle" économique. Comme le note Debreu (1986), la première étape a été de formuler des affirmations économiques dans un language mathématique. Les différentes grandeurs sont vues comme des variables, et dans les années 1930, on verra apparaître des concepts comme "*statistical demand curves*", pour reprendre la terminologie d'Henry Schultz. Cette approche statistique permettra d'aller plus loin que les travaux pionners

de Engel (1857) qui étudiait empiriquement la relation entre la consommation et le revenu des ménages, par exemple, dans une approche uniquement descriptive.

Les modèles économétriques se sont développés en parallèle des modèles macro-économiques. Les premiers travaux de la Commission Cowles ont porté sur l'identification des modèles économiques, et l'estimation de modèles à équations simultanées. Ces développements vont aboutir à un âge d'or de l'économétrie, dans les années 1960, où les modèles économétriques seront utilisés afin d'améliorer les prévisions macroéconomiques. On va alors voir apparaître tout un ensemble de "lois" qui sont souvent traduites comme des relations linéaires entre des grandeurs agrégées, telle que la "loi de Okun" introduite dans Okun (1962) qui postule une relation linéaire entre le variation du nombre de demandeurs d'emploi et de la croissance du PIB,

$$\Delta\text{Chômage}_t = \beta_0 + \beta_1 \text{Croissance}_t + \varepsilon_t,$$

quand on étudie ces grandeurs au cours du temps (t), ou la loi de "Feldstein-Horioka" introduite dans Feldstein & Horioka (1980) qui suppose une relation linéaire entre les taux d'investissement et d'épargne, relativement au revenu national,

$$\frac{\text{investissement}_i}{\text{revenu national}_i} = \beta_0 + \beta_1 \frac{\text{épargne}_i}{\text{revenu national}_i} + \varepsilon_i$$

quand on modèlise les liens entre les allocations investissement-épargne pour plusieurs pays (i). Cet âge d'or correspond aussi à un questionnement profond, suite à la critique de Lucas (1976), s'interrogeant sur l'inefficacité de ces outils à expliquer et à prévoir des crises. La principale explication était alors le manque de fondement micro-économiques de ces modèles, ce qui donnera un second souffle aux modèles micro-économétriques. On pourra rappeler que cette critique dite "de Lucas" avait été formulée dans Orcutt (1952), qui avançait l'idée que les données macroéconomiques posaient des problèmes insolubles d'identification. La solution passait forcément par de l'économétrie sur données individuelles (au lieu de données obtenues par aggrégation), ce qui sera reformulé quelques années plus tard par Koopmans (1957).

Malheureusement, les modèles micro-économétriques sont généralement plus complexes, car ils se doivent de tenir compte d'une éventuelle censure dans les données, avec par exemple le modèle introduit par Tobin (1958), d'erreurs sur les variables (qui pourront être corrigées par des instruments avec une technique initiée par Reiersøol (1945)) ou avoir été collectée avec un biais de sélection, avec les techniques proposées par Heckman (1979). On notera que les économètres se sont beaucoup interrogés sur la manière dont les données étaient construites, et ne se sont jamais contenté de "construire des modèles". Un exemple peut être l'évaluation des politiques publiques, largement détaillé dans Givord (2010). Dans ce cas, en effet, deux écoles se sont opposées (initiant un débat que l'on verra resurgir tout au long de l'article sur les méthodes d'apprentissage statistique). La première, dite "structuraliste", cherchera à construire un modèle complet afin de décrire le comportement des agents économiques. La seconde, souvent qualifiée d'"empiriste", vise à tester l'effet d'une mesure sans pour autant expliciter les mécanismes sous-jacents. C'est ce qu'expliquent très bien Angrist & Krueger (1991), lorsqu'ils affirment "*research in a structuralist style relies heavily on economic theory to guide empirical work* $[\cdots]$ *An alternative to structural modeling,* $[\cdots]$ *the 'experimentalist' approach,* $[\cdots]$ *puts front and center the problem of identifying causal effects from specific events or situations*".

En parallèle, alors que l'analyse économétrique (en particulier à des fins de politique économique) s'est développée plus récemment autour de l'inférence causale, les techniques d'apprentissage machine ont été vues, traditionnellement, autour de la prédiction (où la recherche de corrélations suffisamment fortes entre variables suffit) d'où leur popularité dans des usages plus industriels de classification, comme la reconnaissance de caractères, de signature, d'images, ou de traduction, comme le rappelle Bishop (2006). En biologie, ces techniques ont été appliquées pour créer des classifications d'espèces animales en fonction d'analyse d'ADN, ou dans le domaine militaire et sécuritaire pour l'identification de cibles ou de terroristes (potentiels). Il faudra attendre les années 1990 pour voir des applications en finance avec Altman *et al.* (1994) par exemple, ou Herbrich *et al.* (1999) pour une revue de littérature sur les applications potentielles en économie. Si des applications sont aujourd'hui nombreuses, et si ces techniques concurrencent les modèles de micro-économétrie (on pourra penser au scoring bancaire, à la détection de fraude fiscale ou assurantielle, ou à l'identification de prospects en marketing), les algorithmes d'apprentissage sont devenus très populaires en reconnaissance de parole, puis d'images, et plus récemment avec les applications en ligne et les applications aux jeux (d'échec, et plus récemment de go). Si l'économétrie s'est développée au confluent des mathématiques et de l'économie, l'apprentissage machine (que l'on pourrait avoir tendance à rapprocher de l'intelligence artificelle) s'est développé à la frontière des mathématiques et de l'informatique (avec des résultats fondamentaux en optimisation - en particulier autour des méthodes de gradient stochastique - et sur les espaces sparses).

1.3 Les Données massives

Dans cet article, une variable sera un vecteur de $\mathbb{R}^n$, de telle sorte qu'en concaténant les variables ensemble, on puisse les stocker dans une matrice $\boldsymbol{X}$, de taille $n \times p$, avec n et p potentiellement grands[1]. Le fait que n soit grand n'est, a priori, pas un problème en soi, au contraire. De nombreux théorèmes en économétrie et en statistique sont obtenus lorsque $n \to \infty$ (c'est la théorie asymptotique). En revanche, le fait que p soit grand est problématique, en particulier si $p > n$. Les deux dimensions sont à distinguer, car elles vont engendrer des problèmes relativement différents.

Portnoy (1988) a montré que l'estimateur du maximum de vraisemblance conserve la propriété de normalité asymptotique si p reste petit devant n, ou plus précisément, si $p^2/n \to 0$ lorsque $n, p \to \infty$. Aussi, il n'est pas rare de parler de données massives dès lors que $p > \sqrt{n}$. Un autre concept important est celui de sparcité, qui repose non pas sur la dimension p mais sur la dimension effective, autrement dit le nombre de variables effectivement significatives. Il est alors possible d'avoir $p > n$ tout en ayant des estimateurs convergents.

La grande dimension en terme de nombre de variables, p, peut faire peur à cause de la malédiction de la dimension, introduit par Bellman (1957). L'explication de cette malédiction est que le volume de la sphère unité, en dimension p, tend vers 0 lorsque $p \to \infty$. On dit alors que l'espace est *sparse*, c'est à dire que la probabilité de trouver un point proche d'un autre devient de plus en plus faible (on pourrait parler d'espace clairsemé). Ou de manière duale, pour reprendre la formulation de Hastie *et al.* (2009), le volume qu'il convient de considérer pour avoir une proportion donnée d'observations augmente avec p. L'idée de réduire la dimension en considérant une analyse en composante principale peut paraître séduisante, mais l'analyse souffre d'un certain nombre de défauts en grande dimension. La solution est alors souvent la sélection de variables, qui pose le problème des tests multiples, ou du temps de calcul, pour sélectionner k variables parmi p, lorsque p est grand.

1.4 Statistique computationnelle et non-paramétrique

L'objet de ce papier est d'expliquer les différences majeures entre l'économétrie et l'apprentissage statistique, correspondant aux deux cultures mentionnées par Breiman (2001a), lorsqu'il évoque en modélisation statistique la "*data modeling culture*" (reposant sur un modèle stochastique, comme la régression logistique ou le modèle de Cox) et la "*algorithmic modeling culture*" (reposant sur la mise en œuvre d'un algorithme, comme dans les forêts aléatoires ou les supports vecteurs machines) . Mais la frontière entre les deux est très poreuse. À l'intersection se retrouve l'économétrie non-paramétrique. Cette dernière repose sur un modèle probabiliste (comme l'économétrie), tout en insistant davantage sur les algorithmes (et leurs performances) plutôt que sur des théorèmes asymptotiques.

La statistique non-paramétrique repose sur sur des décompositions dans des bases fonctionnelles. L'économétrie linéaire consiste à approcher la fonction $m : \boldsymbol{x} \mapsto \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ par une fonction linéaire. Mais plus généralement, on peut considérer une décomposition dans une base fonctionnelle, et s'intéresser à une approximation obtenue sur un nombre fini de termes :

$$m(\boldsymbol{x}) = \sum_{j=0}^{\infty} \omega_j g_j(\boldsymbol{x}) \quad \text{et} \quad \widehat{m}(\boldsymbol{x}) = \sum_{j=0}^{h^\star} \widehat{\omega}_j g_j(\boldsymbol{x}),$$

où les poids ω_j sont estimés, alors que le nombre de composantes h est optimisé. On retrouvera ici les modèles additifs (dits GAM), par exemple, étudiés dans Hastie & Tibshirani (1990). Une autre solution consiste à considérer un modèle simple, mais local. Par exemple un modèle constant, au voisinage de $\boldsymbol{x}$, obtenu en considérant seulement les observations proches de $\boldsymbol{x}$:

$$\widehat{g}(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\omega}_{\boldsymbol{x}} y_i \quad \text{par exemple} \quad \widehat{g}(\boldsymbol{x}) = \frac{1}{n_{\boldsymbol{x}}} \sum_{i:\|\boldsymbol{x}_i - \boldsymbol{x}\| \leq h} y_i$$

où $n_{\boldsymbol{x}}$ est le nombre d'observations au voisinage de $\boldsymbol{x}$. En mettant des poids fonctions de la distance à $\boldsymbol{x}$, on retrouve ici le modèle obtenu par Nadaraya (1964) et Watson (1964), ou les méthodes de régression locale.

Les différentes méthodes reposent sur des méta-paramètres - correspondant paramètres de lissage - c'est à dire h dans les exemples précédents. Pour un économètre, le paramètre "optimal" pour h est obtenu soit

[1]Là encore, des extensions sont possibles, en particulier dans les données médicales avec des images de type IRM comme variables prédictives, ou des données climatiques avec des cartes en variables prédictives, ou plus généralement une variable tensorielle en dimension plus ou moins grande. Comme le montre Kolda & Bader (2009) il est toutefois possible de se ramener dans le cas usuel (de données sous formes de vecteurs) en utilisant la décomposition de Tucker.

à l'aide de théorèmes asymptotiques, soit à l'aide de techniques de validation, comme en *machine learning*. On obtient alors une valeur numérique, mais on n'a pas d'interprétation en lien avec la taille de l'échantillon, ou les variances des différentes grandeurs.

1.5 Plan de l'article

Pour reprendre le titre de Varian (2014), l'objet de cet article est de présenter les différences fondamentales entre l'économétrie et l'apprentissage statistique, et surtout de voir comment ces deux techniques peuvent apprendre l'une de l'autre, dans un contexte où les bases de données deviennent massives. La Section 2 reviendra sur la construction du modèle linéaire. Le modèle sera introduit ici à partir du modèle Gaussien "homoscédastique". Ce modèle présente l'avantage d'avoir une élégante interprétation géométrique, en terme de projection sur le sous-espace des combinaisons linéaires des variables explicatives. La première extension que nous verrons est le passage du modèle linéaire à un modèle non-linéaire, tout en construisant un prédicteur linéaire. La seconde extension proposera de construire un modèle non-gaussien, pour modéliser une variable indicatrice ou un comptage Poissonnien, par exemple, donnant naissance aux modèles linéaires généralisés (construits pour des variables dans la famille exponentielle).

Une fois rappelé l'origine des outils économétriques standards, dans la Section 3 nous présenterons les outils et techniques développés dans le contexte de l'apprentissage machine. Si l'outil central des modèles économétriques est la distribution de la variable dépendante, Y, les techniques d'apprentissage reposent sur une fonction de perte, ℓ, représentant une "distance" entre la variable d'intérêt y, et le modèle $m(\cdot)$. Nous présenterons tout d'abord l'algorithme de boosting, reposant sur l'idée d'un apprentissage lent, en modélisant séquentiellement les résidus. Le danger des méthodes d'apprentissage est qu'il est aisé de construire un modèle "parfait", dont le pouvoir prédictif serait faible. Nous évoquerons alors les techniques de pénalisation, utilisées pour éviter le sur-apprentissage. Nous évoquerons en particulier les notions d'*in-sample* et *out-of-sample*, et les techniques de validation croisées. Pour conclure cette section, nous reviendrons sur les interprétations probabilistes des outils d'apprentissage, qui permettront de faire le lien entre les différentes approches, tout en restant sur une discussion générale sur la philosophie de ces deux cultures.

Après cette section sur la philosophie des méthodes de *machine learning*, nous reviendrons dans la section 4 sur quelques algorithmes importants : les réseaux de neurones, les supports vecteurs machine (SVM) et enfin les méthodes de type arbres et forêts.

La Section 5 proposera des exemples concrets de comparaison entre les différentes techniques, dans le cas de classifications (binaires) pour des variables $y \in \{0, 1\}$ (achat d'assurance, non-remboursement d'un crédit) et dans un contexte de régression (lorsque la variable d'intérêt n'est plus qualitative - ce que nous simplifierons en notant $y \in \mathbb{R}$). Nous reviendrons avant sur les courbes ROC, outils importants pour juger de la qualité d'un classifieur, malheureusement peu utilisés en économétrie. Nous verrons en particulier les méthodes de bagging, forêts aléatoires ou boosting. Nous reviendrons aussi sur les méthodes de choix de modèles et des méta-paramètres. À travers ces exemples d'application, nous verrons comment les modèles de type *machine learning* peuvent être utilisés pour mieux détecter la mauvaise spécification des modèles de régression paramétriques, à cause de non-linéarités, et/ou d'intéractions manquées.

2 Économétrie et modèle probabiliste

L'importance des modèles probabilistes en économie trouve sa source dans les questionnements de Working (1927), les tentatives de réponses apportées dans les deux tomes de Tinbergen (1939), surtout les travaux qu'ils ont engendré (comme le rappelle Duo (1993) dans son ouvrage sur les fondements de l'économétrie, et plus particulièrement dans le premier chapitre "*The Probability Foundations of Econometrics*"). Rappelons que Trygve Haavelmo a reçu le prix Nobel d'économie en 1989 pour sa "*clarification des fondations de la théorie probabiliste de l'économétrie*". Car comme l'a montré Haavelmo (1944) (initiant un changement profond dans la théorie économétrique dans les années 1930, comme le rappelle le chapitre 8 de Morgan (1990)) l'économétrie repose fondamentalement sur un modèle probabiliste, et ceci pour deux raisons essentielles. Premièrement, l'utilisation de grandeurs (ou "mesures") statistiques telles que les moyennes, les erreurs-types et les coefficients de corrélation à des fins inférentielles ne peut se justifier que si le processus générant les données peut être exprimé en termes de modèle probabiliste. Deuxièmement, l'approche par les probabilitées est relativement générale, et se trouve être particulièrement adaptée à l'analyse des observations "dépendantes" et "non homogènes", telles qu'on les trouve souvent sur des données économiques. On va alors supposer qu'il existe un espace probabiliste $(\Omega, \mathcal{F}, \mathbb{P})$ tel que les observations $(y_i, \boldsymbol{x}_i)$ sont vues comme des réalisations de variables aléatoires $(Y_i, \boldsymbol{X}_i)$. En pratique, la loi jointe du couple $(Y, \boldsymbol{X})$ nous intéresse toutefois peu : la loi de $\boldsymbol{X}$ est inconnue, et c'est la loi de Y conditionnelle à $\boldsymbol{X}$ qui nous intéressera. Dans

la suite, nous noterons x une observation, $\boldsymbol{x}$ un vecteur d'observations, X une variable aléatoire, et $\boldsymbol{X}$ un vecteur aléatoire et la matrice des observations $\boldsymbol{x}_i$ (suivant le contexte).

2.1 Lois conditionnelles et vraisemblance

L'économétrie linéaire a été construite sous l'hypothèse de données individuelles, ce qui revient à supposer les variables $(Y_i, \boldsymbol{X}_i)$ indépendantes (s'il est possible d'imaginer des observations temporelles - on aurait alors un processus $(Y_t, \boldsymbol{X}_t)$). Plus précisément, on va supposer que conditionnellement aux variables explicatives $\boldsymbol{X}_i$, les variables Y_i sont indépendantes. On va également supposer que ces lois conditionnelles restent dans la même famille paramétrique, mais que le paramètre est une fonction de $\boldsymbol{x}$. Dans le modèle linéaire Gaussien on suppose que :

$$(Y|\boldsymbol{X}=\boldsymbol{x}) \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2) \quad \text{avec} \quad \mu(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^\top\boldsymbol{\beta}, \text{ et } \boldsymbol{\beta} \in \mathbb{R}^p. \tag{1}$$

On parle de modèle linéaire car $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] = \beta_0 + \boldsymbol{x}^\top\boldsymbol{\beta}$ est une combinaison linéaire des variables explicatives. C'est un modèle homoscédastique si $\text{Var}[Y|\boldsymbol{X}=\boldsymbol{x}] = \sigma^2$, où σ^2 est une constante positive. Pour estimer les paramètres, l'approche classique consiste à utiliser l'estimateur du Maximum de Vraisemblance, comme l'avait suggéré initialement Ronald Fisher. Dans le cas du modèle linéaire Gaussien, la log-vraisemblance s'écrit :

$$\log \mathcal{L}(\beta_0, \boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{x}) = -\frac{n}{2}\log[2\pi\sigma^2] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2.$$

Notons que le terme de droite, mesurant une distance entre les données et le modèle, va s'interpréter comme la déviance, dans les modèles linéaires généralisés. On va alors poser :

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = \text{argmax}\left\{\log \mathcal{L}(\beta_0, \boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{x})\right\}.$$

L'estimateur du maximum de vraisemblance est obtenu par minimisation de la somme des carrés des erreurs (estimateur dit des "moindres carrés") que nous retrouverons dans l'approche par *machine learning*.

Les conditions du premier ordre permettent de retrouver les équations normales, dont l'écriture matricielle est $\boldsymbol{X}^\top[\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}] = \boldsymbol{0}$, que l'on peut aussi écrire $(\boldsymbol{X}^\top\boldsymbol{X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{y}$. Si la matrice $\boldsymbol{X}$ est de plein rang colonne, alors on retrouve l'estimateur classique :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{\beta} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}$$

en utilisant une écriture basée sur les résidus (comme souvent en économétrie), $y = \boldsymbol{x}^\top\boldsymbol{\beta} + \varepsilon$. Le théorème de Gauss Markov assure que cette estimateur est l'estimateur linéaire sans biais de variance minimale. On peut alors montrer que $\widehat{\boldsymbol{\beta}} \overset{\mathcal{L}}{\sim} \mathcal{N}(\boldsymbol{\beta}, \sigma^2[\boldsymbol{X}^\top\boldsymbol{X}]^{-1})$, et en particulier :

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \quad \text{et} \quad \text{Var}[\widehat{\boldsymbol{\beta}}] = \sigma^2[\boldsymbol{X}^\top\boldsymbol{X}]^{-1}.$$

Lorsque l'on quitte le modèle linéaire Gaussien, il n'existe pas de forme explicite pour les estimateurs, qui sont calculés numériquement, mais qui continuent à vérifier un certain nombre de propriétés asymptotiques, par exemple (moyennant quelques hypothèses classiques) :

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{\mathcal{L}}{\to} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \quad \text{lorsque} \quad n \to \infty$$

pour une matrice $\boldsymbol{\Sigma}$ qui peut s'approcher, numériquement. Ces résultats asymptotiques permettent de faire des tests de significativité, que ce soit le test de Fisher, de Wald, ou du rapport de vraisemblance.

La condition d'avoir une matrice $\boldsymbol{X}$ de plein rang peut être (numériquement) forte en grande dimension. Si elle n'est pas vérifiée, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ n'existe pas. Si $a\mathbb{I}$ désigne la matrice identité, notons toutefois que $(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ existe toujours, pour $\lambda > 0$. Cet estimateur est appelé l'estimateur Ridge de niveau λ (introduit dans les années 60 par Hoerl (1962), et associé à une régularisation étudiée par Tikhonov (1963)). Il estimateur apparaît naturellement dans un contexte d'économétrie Bayesienne (nous le reverrons dans la section suivante présentant les techniques de *machine learning*).

2.2 Les résidus

Il n'est pas rare d'introduire le modèle linéaire à partir de la loi des résidus, comme nous l'avions mentionné auparavant. Aussi, l'équation (1) s'écrit aussi souvent :

$$y_i = \beta_0 + \boldsymbol{x}_i^\top\boldsymbol{\beta} + \varepsilon_i \tag{2}$$

où les ε_i sont des réalisations de variables aléatoires i.i.d., de loi $\mathcal{N}(0,\sigma^2)$. On notera parfois $\boldsymbol{\varepsilon} \overset{\mathcal{L}}{\sim} \mathcal{N}(\mathbf{0},\sigma^2\mathbb{I})$, sous une forme vectorielle. Les résidus estimés sont définis par :

$$\widehat{\varepsilon}_i = y_i - \left[\widehat{\beta}_0 + \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right]$$

Ces résidus sont l'outil de base pour diagnostiquer la pertinence du modèle.

Une extension du modèle décrit par l'équation (1) a été proposé pour tenir compte d'un éventuel caractère hétéroscédastique :

$$(Y|\boldsymbol{X}=\boldsymbol{x}) \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu(\boldsymbol{x}),\sigma^2(\boldsymbol{x}))$$

où $\sigma^2(\boldsymbol{x})$ est une fonction positive des variables explicatives. On peut réécrire ce modèle en posant :

$$y_i = \beta_0 + \boldsymbol{x}_i^\top\boldsymbol{\beta} + \sigma^2(\boldsymbol{x}_i)\cdot\varepsilon_i$$

où les résidus sont toujours i.i.d., mais de variance unitaire,

$$\varepsilon_i = \frac{y_i - [\beta_0 + \boldsymbol{x}_i^\top\boldsymbol{\beta}]}{\sigma(\boldsymbol{x}_i)}.$$

Si l'écriture à l'aide des résidus est populaire en économétrie linéaire (lorsque la variable dépendante est continue), elle ne l'est toutefois plus dans les modèles de comptage, ou la régression logistique.

L'écriture à l'aide d'un terme d'erreur (comme dans l'équation (2)) pose toutefois de nombreuses questions quant à la représentation d'une relation économique entre deux grandeurs. Par exemple, on peut supposer qu'il existe une relation (linéaire pour commencer) entre les quantités d'un bien échangé, q et son prix p. On peut ainsi imaginer une équation d'offre

$$q_i = \beta_0 + \beta_1 p_i + u_i$$

où la quantité vendue dépend du prix, mais de manière tout aussi légitime, on peut imaginer que le prix dépend de la quantité produite (ce qu'on pourrait appeler une équation de demande),

$$p_i = \alpha_0 + \alpha_1 q_i + v_i.$$

Historiquement, le terme d'erreur dans l'équation (2) a pu être inteprété comme une erreur idiosyncratique sur la variable y, les variables dites explicatives étant supposées fixées, mais cette interprétation rend souvent le lien entre une relation économique et un modèle économique compliqué, la théorie économique parlant de manière abstraite d'une relation entre grandeur, ma modélisation économétrique imposant une forme spécifique (quelle grandeur est y et quelle grandeur est x) comme le montre plus en détails le chapitre 7 de Morgan (1990).

2.3 Géométrie du modèle linéaire gaussien

Définissons le produit scalaire dans $\mathbb{R}^n$, $\langle \boldsymbol{a},\boldsymbol{b}\rangle = \boldsymbol{a}^\top\boldsymbol{b}$, et notons $\|\cdot\|$ la norme euclidienne associée, $\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}^\top\boldsymbol{a}}$. Notons $\mathcal{E}_{\boldsymbol{X}}$ l'espace engendré par l'ensemble des combinaisons linéaires des composantes $\boldsymbol{x}$ (en rajoutant la constante). Si les variables explicatives sont linéairement indépendantes, $\boldsymbol{X}$ est de plein rang colonne et $\mathcal{E}_{\boldsymbol{X}}$ est un sous-espace de dimension $p+1$ de $\mathbb{R}^n$. Supposons à partir de maintenant que les variables $\boldsymbol{x}$ et la variable y sont ici centrées.

Avec cette notation, notons que le modèle linéaire s'écrit $m(\boldsymbol{x}) = \langle \boldsymbol{x},\boldsymbol{\beta}\rangle$. L'espace $\mathcal{H}_z = \{\boldsymbol{x}\in\mathbb{R}^k : m(\boldsymbol{x}) = z)\}$ est un hyperplan (affine) qui sépare l'espace en deux. Définissons l'opérateur de projection orthogonale sur $\mathcal{E}_{\mathcal{X}} = \mathcal{H}_0$, $\Pi_{\boldsymbol{X}} = \boldsymbol{X}[\boldsymbol{X}^\top\boldsymbol{X}]^{-1}\boldsymbol{X}^\top$. Aussi, la prévision que l'on peut faire pour $\boldsymbol{y}$ est :

$$\widehat{\boldsymbol{y}} = \underbrace{\boldsymbol{X}[\boldsymbol{X}^\top\boldsymbol{X}]^{-1}\boldsymbol{X}^\top}_{\Pi_{\mathcal{X}}}\boldsymbol{y} = \Pi_{\mathcal{X}}\boldsymbol{y}. \tag{3}$$

Comme $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\mathbb{I}-\Pi_{\mathcal{X}})\boldsymbol{y} = \Pi_{\mathcal{X}^\perp}\boldsymbol{y}$, on note que $\widehat{\boldsymbol{\varepsilon}} \perp \boldsymbol{x}$, que l'on interprétera en disant que les résidus sont un terme d'innovation, imprévisible, au sens où $\Pi_{\mathcal{X}}\widehat{\boldsymbol{\varepsilon}} = \mathbf{0}$.

Le théorème de Pythagore s'écrit ici :

$$\|\boldsymbol{y}\|^2 = \|\Pi_{\mathcal{X}}\boldsymbol{y}\|^2 + \|\Pi_{\mathcal{X}^\perp}\boldsymbol{y}\|^2 = \|\Pi_{\mathcal{X}}\boldsymbol{y}\|^2 + \|\boldsymbol{y}-\Pi_{\mathcal{X}}\boldsymbol{y}\|^2 = \|\widehat{\boldsymbol{y}}\|^2 + \|\widehat{\boldsymbol{\varepsilon}}\|^2 \tag{4}$$

qui se traduit classiquement en terme de somme de carrés :

$$\underbrace{\sum_{i=1}^n y_i^2}_{n\times\text{variance totale}} = \underbrace{\sum_{i=1}^n \widehat{y}_i^2}_{n\times\text{variance expliquée}} + \underbrace{\sum_{i=1}^n (y_i-\widehat{y}_i)^2}_{n\times\text{variance résiduelle}}$$

Le coefficient de détermination, R^2 (nous reviendrons sur ce coefficient dans la section 2.7) s'interprête alors comme le carré du cosinus de l'angle θ entre $\boldsymbol{y}$ et $\Pi_{\mathcal{X}}\boldsymbol{y}$:

$$R^2 = \frac{\|\Pi_{\mathcal{X}}\boldsymbol{y}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{\|\Pi_{\mathcal{X}^\perp}\boldsymbol{y}\|^2}{\|\boldsymbol{y}\|^2} = \cos^2(\theta).$$

Une application importante a été obtenue par Frish & Waugh (1933), lorsque l'on partitionne les variables explicatives en deux groupes, $\boldsymbol{X} = [\boldsymbol{X}_1|\boldsymbol{X}_2]$, de telle sorte que la régression devient :

$$\boldsymbol{y} = \beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \varepsilon$$

Frish & Waugh (1933) ont montré qu'on pouvait considérer deux projections successives. En effet, si $\boldsymbol{y}_2^\star = \Pi_{\mathcal{X}_1^\perp}\boldsymbol{y}$ et $\boldsymbol{X}_2^\star = \Pi_{\mathcal{X}_1^\perp}\boldsymbol{X}_2$, on peut montrer que

$$\widehat{\boldsymbol{\beta}}_2 = [\boldsymbol{X}_2^{\star\mathsf{T}}\boldsymbol{X}_2^\star]^{-1}\boldsymbol{X}_2^{\star\mathsf{T}}\boldsymbol{y}_2^\star$$

Autrement dit, l'estimation globale est équivalente à l'estimation indépendante des deux modèles si $\boldsymbol{X}_2^\star = \boldsymbol{X}_2$, c'est à dire $\boldsymbol{X}_2 \in \mathcal{E}_{\boldsymbol{X}_1}^\perp$, que l'on peut noter $\boldsymbol{x}_1 \perp \boldsymbol{x}_2$. On obtient ici le théorème de Frisch-Waugh qui garantie que si les variables explicatives entre les deux groupes sont orthogonales, alors l'estimation globale est équivalente à deux régressions indépendantes, sur chacun des jeux de variables explicatives. Ce qui est un théorème de double projection, sur des espaces orthogonaux. Beaucoup de résultats et d'interprétations sont obtenus par des interprétations géométriques (liées fondamentalement aux liens entre l'espérance conditionnelle et la projection orthogonale dans $\mathcal{L}_2$).

2.4 Du paramétrique au non-paramétrique

La réécriture de l'équation (3) sous la forme

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\boldsymbol{X}^\mathsf{T}}_{S_{\mathcal{X}}}\boldsymbol{y}$$

permet de voir la prévision directement comme une transformation linéaire des observations. De manière plus générale, on peut obtenir un prédicteur linéaire en considérant $m(\boldsymbol{x}) = \boldsymbol{s}_{\boldsymbol{x}}^\mathsf{T}\boldsymbol{y}$, où $\boldsymbol{s}_{\boldsymbol{x}}$ est un vecteur de poids, qui dépendent de $\boldsymbol{x}$, interprété comme un vecteur de lissage. En utilisant les vecteurs $\boldsymbol{s}_{\boldsymbol{x}_i}$, calculés à partir des $\boldsymbol{x}_i$, on obtient une matrice $\boldsymbol{S}$ de taille $n \times n$, et $\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$. Dans le cas de la régression linéaire, $\boldsymbol{s}_{\boldsymbol{x}} = \boldsymbol{X}[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\boldsymbol{x}$, et classiquement, trace($\boldsymbol{S}$) est le nombre de colonnes de la matrice $\boldsymbol{X}$ (le nombre de variables explicatives). Le principe de parcimonie consiste à minimiser cette dimension. Mais dans le cas général, cette dimension est plus complexe à définir. Notons que l'estimateur introduit par Nadaraya (1964) et Watson (1964), dans le cas d'une régression non-paramétrique simple, s'écrit également sous cette forme puisque

$$\widehat{m}_h(x) = \boldsymbol{s}_x^\mathsf{T}\boldsymbol{y} = \sum_{i=1}^n s_{x,i}y_i \quad \text{avec} \quad s_{x,i} = \frac{K_h(x-x_i)}{K_h(x-x_1)+\cdots+K_h(x-x_n)},$$

où $K(\cdot)$ est une fonction noyau, qui attribue une valeur d'autant plus faible que x_i est proche de x, et $h > 0$ est la fenêtre de lissage. Dans ce contexte de prédicteurs linéaires, trace($\boldsymbol{S}$) est souvent vu comme un équivalent au nombre de paramètres (ou complexité du modèle), et $\nu = n - \text{trace}(\boldsymbol{S})$ est alors le nombre de degrés de liberté (comme défini dans Ruppert, Wand & Carroll (2003) et Simonoff (1996)).

2.5 Famille exponentielle et modèles linéaires

Le modèle linéaire Gaussien est un cas particulier d'une vaste famille de modèles linéaires, obtenu lorsque la loi conditionnelle de Y appartient à la famille exponentielle

$$f(y_i|\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right) \quad \text{avec} \quad \theta_i = h(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}).$$

Les fonctions a, b et c sont spécifiées en fonction du type de loi exponentielle (étudiée abondamment en statistique depuis les Darmois (1935), comme le rappelle Brown (1986)) La log-vraisemblance a alors une expression relative simple

$$\log\mathcal{L}(\boldsymbol{\theta},\phi|\boldsymbol{y}) = \prod_{i=1}^n \log f(y_i|\theta_i,\phi) = \frac{\sum_{i=1}^n y_i\theta_i - \sum_{i=1}^n b(\theta_i)}{a(\phi)} + \sum_{i=1}^n c(y_i,\phi)$$

et la condition du premier ordre s'écrit alors

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}, \phi|\boldsymbol{y})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{W}^{-1}[\boldsymbol{y} - \widehat{\boldsymbol{y}}] = \boldsymbol{0}$$

pour reprendre les notations de Müller (2011), où $\boldsymbol{W}$ est une matrice de poids (qui dépend de $\boldsymbol{\beta}$). Compte tenu du lien entre θ et l'espérance de Y, au lieu de spécifier la fonction $h(\cdot)$, on aura plutôt tendance à specifier la fonction de lien $g(\cdot)$ définie par

$$\widehat{y} = m(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = g^{-1}(\boldsymbol{x}^\top \boldsymbol{\beta}).$$

Pour la régression linéaire Gaussienne on prendra un lien Identité, alors que pour la régression de Poisson, le lien naturel (dit canonique) est le lien logarithmique. Ici, comme $\boldsymbol{W}$ dépend de $\boldsymbol{\beta}$ (avec $\boldsymbol{W} = \text{diag}(\nabla g(\widehat{\boldsymbol{y}})\text{Var}[\boldsymbol{y}])$) il n'existe en général par de formule explicite pour l'estimateur du maximum de vraisemblance. Mais un algorithme itératif permet d'obtenir une approximation numérique. En posant

$$\boldsymbol{z} = g(\widehat{\boldsymbol{y}}) + (\boldsymbol{y} - \widehat{\boldsymbol{y}}) \cdot \nabla g(\widehat{\boldsymbol{y}})$$

correspondant au terme d'erreur d'un développement de Taylor à l'ordre 1 de g, on obtient un algorithme de la forme

$$\widehat{\boldsymbol{\beta}}_{k+1} = [\boldsymbol{X}^\top \boldsymbol{W}_k^{-1} \boldsymbol{X}]^{-1} \boldsymbol{X}^\top \boldsymbol{W}_k^{-1} \boldsymbol{z}_k$$

En itérant, on notera $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_\infty$, et on peut montrer que - moyennant quelques hypothèses techniques (cf Müller (2011)) - cet estimateur est asymptotiquement Gaussien, avec

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, I(\boldsymbol{\beta})^{-1}),$$

où numériquement $I(\boldsymbol{\beta}) = \phi \cdot [\boldsymbol{X}^\top \boldsymbol{W}_\infty^{-1} \boldsymbol{X}]$.

D'un point de vue numérique toujours, on résout la condition du premier ordre, et la loi de Y n'intervient pas réellement. Par exemple, on peut estimer une "régression de Poisson" même lorsque $y \in \mathbb{R}_+$, pas nécessairement $y \in \mathbb{N}$. Autrement dit, la loi de Y n'est qu'une interprétation donnée ici, et l'algorithme pourrait être introduit de manière différente (comme nous le verrons dans la section suivante), sans forcément avoir de modèle probabiliste sous-jacent.

2.6 Régression logistique

La régression logistique est le modèle linéaire généralisé obtenu avec une loi de Bernoulli, et une fonction de lien qui est la fonction quantile d'une loi logistique (ce qui correspond au lien canonique au sens de la famille exponentielle). Compte tenu de la forme de la loi de Bernoulli, l'économétrie propose un modèle pour $y_i \in \{0, 1\}$, dans lequel le logarithme de la cote suit un modèle linéaire :

$$\log\left(\frac{\mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}]}{\mathbb{P}[Y \neq 1|\boldsymbol{X} = \boldsymbol{x}]}\right) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta},$$

ou encore :

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}] = \frac{e^{\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}}}{1 + e^{\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}}} = H(\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}), \quad \text{où} \quad H(.) = \frac{exp(.)}{1 + exp(.)},$$

correspondant à la fonction de répartition de la loi logistique. L'estimation de $(\beta_0, \boldsymbol{\beta})$ se fait par maximisation de la vraisemblance :

$$\mathcal{L} = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}\right)^{y_i} \left(\frac{1}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}\right)^{1 - y_i}$$

On continuera à parler des modèles linéaires car les courbes d'isoprobabilités sont ici les hyperplans parallèles $b_0 + \boldsymbol{x}^\top \boldsymbol{\beta}$. À ce modèle, popularisé par Berkson (1944), certains préfèront le modèle probit (comme le raconte Berkson (1951)), introduit par Bliss (1934). Dans ce modèle :

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}] = \Phi(\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}),$$

où Φ désigne la fonction de répartition de la loi normale centrée réduite. Ce modèle présente l'avantage d'avoir un lien direct avec le modèle linéaire Gaussien, puisque

$$y_i = \mathbf{1}(y_i^\star > 0) \quad \text{avec} \quad y_i^\star = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon_i$$

où les résidus sont Gaussiens, de loi $\mathcal{N}(0, \sigma^2)$. Un alternative est d'avoir des résidus centrés de variance unitaire, et de considérer une modélisation latente de la forme $y_i = \mathbf{1}(y_i^\star > \xi)$ (où ξ sera à estimer). On le voit, ces techniques sont fondamentalement liées à un modèle stochastique sous-jacent.

2.7 Qualité d'un ajustement et choix de modèle

Dans le modèle linéaire Gaussien, le coefficient de détermination - noté R^2 - est souvent utilisé comme mesure de la qualité d'ajustement. Compte tenu de la formule de décomposition de la variance

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{variance totale}}=\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{\text{variance résiduelle}}+\underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i-\bar{y})^2}_{\text{variance expliquée}}$$

on définit le R^2 comme le ratio de variance expliquée et de la variance totale, autre interpétation du coefficient que nous avions introduit à partir de la géométrie des moindres carrés.

$$R^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Les sommes des carrés d'erreurs dans cette écriture peut se réécrire comme une log-vraisemblance. Or rappelons qu'à une constante près, dans les modèles linéaires généralisés, la déviance est définie par

$$\text{Deviance}(\boldsymbol{\beta})=-2\log[\mathcal{L}]$$

que l'on peut aussi noter $\text{Deviance}(\widehat{\boldsymbol{y}})$. On peut définir une déviance nulle comme celle obtenue sans utiliser les variables explicatives $\boldsymbol{x}$, de telle sorte que $\widehat{y}_i=\overline{y}$. On peut alors définir, dans un contexte plus général

$$R^2=\frac{\text{Deviance}(\overline{y})-\text{Deviance}(\widehat{\boldsymbol{y}})}{\text{Deviance}(\overline{y})}=1-\frac{\text{Deviance}(\widehat{\boldsymbol{y}})}{\text{Deviance}(\overline{y})}.$$

Toutefois, cette mesure ne peut être utilisée pour choisir un modèle, si on souhaite avoir au final un modèle relativement simple, car elle augmente artificiellement avec l'ajout de variables explicatives sans effet significatif. On aura alors tendance à préférer le R^2 ajusté

$$\bar{R}^2=1-(1-R^2)\frac{n-1}{n-p}=R^2-\underbrace{(1-R^2)\frac{p-1}{n-p}}_{\text{pénalisation}},$$

où p est le nombre de paramètres du modèle (noté plus généralement ν dans la section 2.4). À la mesure de la qualité de l'ajustement, on va pénaliser les modèles trop complexes.

Cette idée va se retrouver dans le critère d'Akaike, où $AIC=\text{Deviance}+2\cdot p$ ou dans le critère de Schwarz, $BIC=\text{Deviance}+\log(n)\cdot p$. En grande dimension (typiquement $p>\sqrt{n}$), on aura tendance à utiliser un AIC corrigé, défini par

$$AICc=\text{Deviance}+2\cdot p\cdot\frac{n}{n-p-1}$$

Ces critères sont utilisés dans les méthodes dites "*stepwise*", introduisant les méthodes ensemblistes. Dans la méthode dite "*forward*", on commence par régresser sur la constante, puis on ajoute une variable à la fois, en retenant celle qui fait le plus baisser le critère AIC, jusqu'à ce que rajouter une variable augmente le critère AIC du modèle. Dans la méthode dite "*backward*", on commence par régresser sur toutes les variables, puis on enlève une variable à la fois, en retirant celle qui fait le plus baisser le critère AIC, jusqu'à ce que retirer une variable augmente le critère AIC du modèle.

Une autre justification de cette notion de pénalisation (nous reviendrons sur cette idée en apprentissage) peut être la suivante. Considérons un prédicteur linéaire (relativement général) $\widehat{\boldsymbol{y}}=\widehat{m}(\boldsymbol{X})=\boldsymbol{S}\boldsymbol{y}$ - où $\boldsymbol{S}$ est la matrice $n\times n$ des $\boldsymbol{s}_{\boldsymbol{x}_i}$ - et supposons que $\boldsymbol{y}=m_0(\boldsymbol{x})+\boldsymbol{\varepsilon}$, avec $\mathbb{E}[\boldsymbol{\varepsilon}]=\boldsymbol{0}$ et $\text{Var}[\boldsymbol{\varepsilon}]=\sigma^2\mathbb{I}$, de telle sorte que $m_0(\boldsymbol{x})=\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]$. Le risque empirique, associé à un modèle m, est ici :

$$\widehat{\mathcal{R}}_n(m)=\frac{1}{n}\sum_{i=1}^{n}(y_i-m(\boldsymbol{x}_i))^2=\frac{1}{n}\|\boldsymbol{y}-m(\boldsymbol{x})\|^2$$

(par convention). On reconnaît ici l'erreur quadratique moyenne, MSE, qui donnera plus généralement le "risque" du modèle m quand on utilise une autre fonction de perte (comme nous le discuterons dans la partie suivante). Notons que:

$$\mathbb{E}[\widehat{\mathcal{R}}_n(m)]=\frac{1}{n}\|m_0(\boldsymbol{x})-m(\boldsymbol{x})\|^2+\frac{1}{n}\mathbb{E}\big(\|\boldsymbol{y}-m_0(\boldsymbol{x})\|^2\big)$$

On peut montrer que :

$$n\mathbb{E}\big[\widehat{\mathcal{R}}_n(\widehat{m})\big] = \mathbb{E}\big(\|\boldsymbol{y} - \widehat{m}(\boldsymbol{x})\|^2\big) = \|(\mathbb{I} - \boldsymbol{S})m_0\|^2 + \sigma^2\|\mathbb{I} - \boldsymbol{S}\|^2$$

de telle sorte que le (vrai) risque de $\widehat{m}$ est :

$$\mathcal{R}_n(\widehat{m}) = \mathbb{E}\big[\widehat{\mathcal{R}}_n(\widehat{m})\big] + 2\frac{\sigma^2}{n}\text{trace}(\boldsymbol{S}).$$

Aussi, si $\text{trace}(\boldsymbol{S}) \geq 0$, le risque empirique sous-estime le vrai risque de l'estimateur. On reconnaît ici le nombre de degrés de liberté du modèle, le terme de droite correspondant au C_p de Mallow, introduit dans Mallows (1973) utilisant non pas la déviance mais le R^2.

2.8 Économétrie et tests statistiques

Le test le plus classique en économétrie est probablement le test de significativité, correspondant à la nullité d'un coefficient dans un modèle de régression linéaire. Formellement, il s'agit du test de $H_0 : \beta_k = 0$ contre $H_1 : \beta_k \neq 0$. Le test de Student, basé sur la statistique $t_k = \widehat{\beta}_k/\text{se}_{\widehat{\beta}_k}$, permet a priori de trancher entre les deux alternatives, à l'aide de la p-value du test, définie par $\mathbb{P}[|T| > |t_k|]$ avec $T \overset{\mathcal{L}}{\sim} t_\nu$, où ν est le nombre de degrés de liberté du modèle ($\nu = p + 1$ pour le modèle linéaire standard). En grande dimension, cette statistique est néanmoins d'un intérêt très limité, compte tenu d'un FDR (False Discovery Ratio) important. Classiquement, avec un niveau de significativité $\alpha = 0.05$, 5% des variables sont faussement significatives. Supposons que nous disposions de $p = 100$ variables explicatives, mais que 5 (seulement) sont réellement significatives. On peut espérer que ces 5 variables passeront le test de Student, mais on peut aussi s'attendre à ce que 5 variables supplémentaires (test faussement positif) ressortent. On aura alors 10 variables perçues comme significatives, alors que seulement la moitié le sont, soit un ratio FDR de 50%. Afin d'éviter cet écueil récurent dans les tests multiples, il est naturel d'utiliser la procédure de Benjamini & Hochberg (1995).

2.9 Sous- et sur-identification

La sous-identification correspond au cas où le vrai modèle serait $y_i = \beta_0 + \boldsymbol{x}_1^\intercal\boldsymbol{\beta}_1 + \boldsymbol{x}_2^\intercal\boldsymbol{\beta}_2 + \varepsilon_i$, mais le modèle estimé est $y_i = \beta_0 + \boldsymbol{x}_1^\intercal\boldsymbol{b}_1 + \eta_i$. L'estimateur du maximum de vraisemblance de $\boldsymbol{b}_1$ est :

$$\begin{aligned}
\widehat{\boldsymbol{b}}_1 &= (\boldsymbol{X}_1^\intercal\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\intercal\boldsymbol{y} \\
&= (\boldsymbol{X}_1^\intercal\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\intercal[\boldsymbol{X}_{1,i}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2,i}\boldsymbol{\beta}_2 + \varepsilon] \\
&= (\boldsymbol{X}_1^\intercal\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\intercal\boldsymbol{X}_1\boldsymbol{\beta}_1 + (\boldsymbol{X}_1^\intercal\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\intercal\boldsymbol{X}_2\boldsymbol{\beta}_2 + (\boldsymbol{X}_1^\intercal\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\intercal\varepsilon \\
&= \boldsymbol{\beta}_1 + \underbrace{(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\intercal\boldsymbol{X}_2\boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}} + \underbrace{(\boldsymbol{X}_1^\intercal\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\intercal\varepsilon}_{\nu_i}
\end{aligned}$$

de telle sorte que $\mathbb{E}[\widehat{\boldsymbol{b}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12}$, le biais étant nul uniquement dans le cas où $\boldsymbol{X}_1^\intercal\boldsymbol{X}_2 = \boldsymbol{0}$ (c'est à dire $\boldsymbol{X}_1 \perp \boldsymbol{X}_2$): on retrouve ici une conséquence du théorème de Frisch-Waugh.

En revanche, la sur-identification correspond au cas où le vrai modèle serait $y_i = \beta_0 + \boldsymbol{x}_1^\intercal\boldsymbol{\beta}_1 + \varepsilon_i$, mais le modèle estimé est $y_i = \beta_0 + \boldsymbol{x}_1^\intercal\boldsymbol{b}_1 + \boldsymbol{x}_2^\intercal\boldsymbol{b}_2 + \eta_i$. Dans ce cas, l'estimation est sans biais, au sens où $\mathbb{E}(\widehat{\boldsymbol{b}}_1) = \boldsymbol{\beta}_1$ mais l'estimateur n'est pas efficient. Et comme nous l'avons vu dans la section précédente, il n'est pas rare d'avoir des valeurs de $\widehat{\boldsymbol{b}}_2$ qui sont considérées comme significativement non-nulles. Nous évoquerons dans la section suivante une méthode efficace de choix de variables (et éviter la sur-identification).

2.10 Quitter la corrélation pour quantifier un effet causal

C'est à Jerry Neyman que l'on doit les premiers travaux sur l'identification de mécanismes causaux, c'est Rubin (1974) qui a formalisé le test, appelé "modèle causal de Rubin" dans Holland (1986).

Les premières approches autour de la notion de causalité en économétrie se sont faites avec l'utilisation des variables instrumentales, des modèles avec discontinuité de régression, l'analyse de différences dans les différences, ainsi que des expériences naturelles ou pas. La causalité est généralement déduite en comparant l'effet d'une politique - ou plus généralement d'un traitement - avec son contrefactuel, idéalement donné par un groupe témoin, aléatoire. L'effet causal du traitement est alors défini comme $\Delta = y_1 - y_0$, c'est à dire la différence entre ce que serait la situation avec traitement (noté $t = 1$) et sans traitement (noté $t = 0$). Le souci est que seul $y = t \cdot y_1 + (1 - t)y_0$ et t sont observés. Autrement dit l'effet causal de la variable t sur y n'est pas observé (puisque seule une des deux variables potentielles - y_0 ou y_1 est observée pour

chaque individu), mais il est aussi individuel, et donc fonction de covariables $\boldsymbol{x}$. Généralement, en faisant des hypothèses sur la distribution du triplet (Y_0, Y_1, T), certains paramètres de la distribution de l'effet causal deviennent identifiables, à partir de la densité des variables observables (Y, T). Classiquement, on sera intéressé par les moments de cette distribution, en particulier l'effet moyen du traitement dans la population, $\mathbb{E}[\Delta]$, voire juste l'effet moyen du traitement en cas de traitement $\mathbb{E}[\Delta|T=1]$. Si le résultat (Y_0, Y_1) est indépendant de la variable d'accès au traitement T, on peut montrer que $\mathbb{E}[\Delta] = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$. Mais si cette hypothèse d'indépendance n'est pas vérifiée, on a un biais de sélection, souvent associé à $\mathbb{E}[Y_0|T=1] - \mathbb{E}[Y_0|T=0]$. Rosenbaum & Rubin (1983) proposent d'utiliser un score de propension à être traité, $p(\boldsymbol{x}) = \mathbb{P}[T=1|\boldsymbol{X}=x]$, en notant que si la variable Y_0 est indépendante de l'accès au traitement T conditionnellement aux variables explicatives $\boldsymbol{X}$, alors elle est indépendante de T conditionnellement au score $p(\boldsymbol{X})$: il suffit de les apparier à l'aide de leur score de propension. Heckman *et al.* (2003) propose ainsi un estimateur à noyau sur le score de propension, ce qui permet d'avoir simplement un estimateur de l'effet du traitement, conditionnellement au fait d'être traité.

3 Philosophie des méthodes de *machine learning*

En apprentissage statistique, on n'a pas besoin de modèle probabiliste. Les probabilités sont un fondement de l'économétrie, alors que les probabilités servent plutôt d'outil pour interpréter un modèle d'apprentissage machine. Cela n'empêche pas de supposer qu'il y a une composante aléatoire dans le modèle, comme dans le "perceptron" de Rosenblatt (1958), où les cellules sont stimulées avec une certaine probabilité, et répondent ensuite à ce stimuli avec une autre probabilité. Nous allons présenter les fondements des techniques de *machine learning* (les exemples d'algorithmes étant présentés dans les sections suivantes). Le point important, comme nous allons le voir, est que la principale préoccupation de l'apprentissage machine est liée aux propriétés de généralisation d'un modèle, c'est-à-dire sa performance - selon un critère choisi a priori - sur des données nouvelles, et donc des tests hors échantillon.

3.1 Apprentissage et fonctions de perte

Commençons par décrire un modèle historiquement important, le "perceptron" de Rosenblatt (1958), introduit dans des problèmes de classification, où $y \in \{0, 1\}$. On dispose de données $\{(y_i, \boldsymbol{x}_i)\}$, et on va construire de manière itérative un ensemble de modèles $m_k(\cdot)$, où à chaque étape, on va apprendre des erreurs du modèle précédent. Dans le perceptron, on considère un modèle linéaire de telle sorte que :

$$m(\boldsymbol{x}) = \mathbf{1}(\beta_0 + \boldsymbol{x}^\intercal \boldsymbol{\beta} \geq 0),$$

où les coefficients β_j sont souvent interprétés comme des "poids" attribués à chacune des variables explicatives. On se donne des poids initiaux $(\beta_0^{(0)}, \boldsymbol{\beta}^{(0)})$, que l'on va mettre à jour en tenant compte de l'erreur de prédiction commise, entre y_i et la prédiction $\widehat{y}_i^{(k)}$:

$$\widehat{y}_i^{(k)} = m^{(k)}(\boldsymbol{x}_i) = \mathbf{1}(\beta_0^{(k)} + \boldsymbol{x}^\intercal \boldsymbol{\beta}^{(k)} \geq 0),$$

avec, dans le cas du perceptron :

$$\beta_j^{(k+1)} = \beta_j^{(k)} + \eta \underbrace{(\boldsymbol{y} - \widehat{\boldsymbol{y}}^{(k)})^\intercal}_{=\ell(\boldsymbol{y}, \widehat{\boldsymbol{y}}^{(k)})} \boldsymbol{x}_j$$

où ici $\ell(y, y') = \mathbf{1}(y \neq y')$ est une fonction de perte, qui permettra de donner un prix à une erreur commise, en prédisant $y' = m(\boldsymbol{x})$ et en observant y. Pour un problème de régression, on peut considérer une erreur quadratique ℓ_2, telle que $\ell(y, m(\boldsymbol{x})) = (y - m(\boldsymbol{x}))^2$ ou en valeur absolue ℓ_1, avec $\ell(y, m(\boldsymbol{x})) = |y - m(\boldsymbol{x})|$. Ici, pour notre problème de classification, nous utilisions une indicatrice de mauvaise qualification (on pourrait argumenter le caractère symétrique de cette fonction de perte, laissant croire qu'un faux positif coûte autant qu'un faux négatif). Une fois spécifiée cette fonction de perte, on voit que l'on cherche à résoudre :

$$m^\star(\boldsymbol{x}) = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, m(\boldsymbol{x})) \right\} \tag{5}$$

pour un ensemble de prédicteurs $\mathcal{M}$ prédéfini. Tout problème d'apprentissage machine est mathématiquement formulé comme un problème d'optimisation, dont la solution détermine un ensemble de paramètres de modèle (si la famille $\mathcal{M}$ est décrite par un ensemble de paramètres - qui peuvent être des coordonnées dans une base fonctionnelle). On pourra noter $\mathcal{M}_0$ l'espace des hyperplans de $\mathbb{R}^p$ au sens où

$$m \in \mathcal{M}_0 \quad \text{signifie} \quad m(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{x} \quad \text{avec} \quad \boldsymbol{\beta} \in \mathbb{R}^p,$$

engendrant la classe des prédicteurs linéaires.

3.2 Apprentissage machine et optimisation

Dans ce dernier cas, quand la fonction de perte est ℓ_2, on retrouve le programme classique dit "des moindres carrés" où l'on cherche à minimiser :

$$S(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i)^2$$

(on retrouve le modèle linéaire et la fonction de perte quadratique comme cas particulier du problème (5). La résolution numérique peut se faire par descente de gradient, en construisant la suite :

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} - \gamma\nabla S(\boldsymbol{\beta}^k) = \widehat{\boldsymbol{\beta}}^{(k)} - \gamma\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{x}_i(\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}^{(k)} - y_i)$$

où γ est vu comme un taux d'apprentissage qui détermine la vitesse de convergence. Parce que ce problème (avec la fonction de perte quadratique) est convexe, chaque synchronisation nous rapproche de la solution optimale $\boldsymbol{\beta}$ (en supposant que le taux d'apprentissage n'a pas été trop mal choisi). Mais pour des fonctions de perte non (strictement) convexe, il convient d'utiliser des algorithmes d'optimisation appropriés. Notons également que dans l'algorithme de Fisher, on utilise la matrice Hessienne pour contrôler la vitesse de la convergence (le gradient donnant la direction). On utilise alors $\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} - H(\boldsymbol{\beta}^k)^{-1}\nabla S(\boldsymbol{\beta}^k)$, où $H(\boldsymbol{\beta})$ désigne la matrice Hessienne associée à $S(\boldsymbol{\beta})$. En grande dimension, le calcul (répété) et l'inversion de cette matrice n'est pas sans causer des problèmes.

3.3 Autres fonctions de perte et interprétations probabilistes

Dans le cas de la perte quadratique, on notera que l'on peut avoir une interprétation particulière de ce problème, puisque :

$$\overline{y} = \underset{m\in\mathbb{R}}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}\frac{1}{n}[y_i - m]^2\right\}$$

Si l'on suppose - comme on le faisait en économétrie - qu'il existe un modèle probabiliste sous-jacent, et en notant que :

$$\mathbb{E}(Y) = \underset{m\in\mathbb{R}}{\operatorname{argmin}}\left\{\|Y-m\|_{\ell_2}^2\right\} = \underset{m\in\mathbb{R}}{\operatorname{argmin}}\left\{\mathbb{E}\left([Y-m]^2\right)\right\}$$

on notera que ce que l'on essaye d'obtenir ici, en résolvant le problème (5) en prenant pour ℓ la norme ℓ_2, est une approximation (dans un espace fonctionnel donné, $\mathcal{M}$) de l'espérance conditionnelle $\boldsymbol{x}\mapsto\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]$. Une autre fonction de perte particulièrement intéressante est la perte ℓ_1, $\ell_1(y,m) = |y-m|$. Le problème d'optimisation :

$$m^\star(\boldsymbol{x}) = \underset{m\in\mathcal{M}_0}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}|y_i - m(\boldsymbol{x})|\right\}$$

est obtenu en économétrie si on suppose que la loi conditionnelle de Y suit une loi de Laplace centrée sur $m(\boldsymbol{x})$, et en maximisant la (log) vraisemblance. Si on réécrit cette fonction de perte $\ell_1(y,m) = |(y-m)(1/2-\mathbf{1}_{y\leq m})|$, on peut obtenir une généralisation pour $\tau\in(0,1)$, et :

$$m^\star(\boldsymbol{x}) = \underset{m\in\mathcal{M}_0}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}\ell_\tau^{\mathsf{q}}(y_i, m(\boldsymbol{x}))\right\} \quad \text{avec} \quad \ell_\tau^{\mathsf{q}}(x,y) = (x-y)(\tau - \mathbf{1}_{x\leq y})$$

est alors la régression quantile de niveau τ (voir Koenker (2003) et d'Haultefœuille & Givord (2014)). Une autre fonction de perte, introduite par Aigner *et al.* (1977) et analysée dans Waltrup *et al.* (2014), est la fonction associée à la notion d'expectiles :

$$\ell_\tau^{\mathsf{e}}(x,y) = (x-y)^2\cdot\left|\tau - \mathbf{1}_{x\leq y}\right|$$

avec $\tau\in[0,1]$. On voit le parallèle avec la fonction quantile :

$$\ell_\tau^{\mathsf{q}}(x,y) = |x-y|\cdot\left|\tau - \mathbf{1}_{x\leq y}\right|.$$

En lien avec cette approche, Gneiting (2011) a introduit la notion de statistique *ellicitable* - ou de mesure *ellicitable* dans sa version probabiliste (ou distributionnelle) : T sera dite *ellicitable* s'il existe une fonction de perte $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ telle que :

$$T(Y) = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \left\{ \int_{\mathbb{R}} l(x,y) dF(y) \right\} = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \left\{ \mathbb{E}\big[l(x,Y)\big] \text{ où } Y \overset{\mathcal{L}}{\sim} F \right\}$$

La moyenne (espérance mathématique) est ainsi ellicitable par la distance quadratique, ℓ_2 alors que la médiane est ellicitable par la distance ℓ_1. Selon Gneiting (2011), cette propriété est essentielle pour construire des prédictions. Il peut alors exister un lien fort entre des mesures associées à des modèles probabilistes et les fonctions de perte. Enfin, la statistique Bayésienne propose un lien direct entre la forme de la loi a priori et la fonction de perte, comme l'ont étudié Berger (1985) et Bernardo & Smith (2000).

3.4 Boosting et apprentissage séquentiel (Lent)

Résoudre le problème décrit par l'équation (5) est d'autant plus complexe que l'espace fonctionnel $\mathcal{M}$ est volumineux. L'idée du Boosting, tel qu'introduit par Shapire & Freund (2012), est d'apprendre, lentement, à partir des erreurs du modèle, de manière itérative. À la première étape, on estime un modèle m_1 pour $\boldsymbol{y}$, à partir de $\boldsymbol{X}$, qui donnera une erreur $\boldsymbol{\varepsilon}_1$. À la seconde étape, on estime un modèle m_2 pour $\boldsymbol{\varepsilon}_1$, à partir de $\boldsymbol{X}$, qui donnera une erreur $\boldsymbol{\varepsilon}_2$, etc. On va alors retenir comme modèle, au bout de k itération :

$$m^{(k)}(\cdot) = \underbrace{m_1(\cdot)}_{\sim \boldsymbol{y}} + \underbrace{m_2(\cdot)}_{\sim \boldsymbol{\varepsilon}_1} + \underbrace{m_3(\cdot)}_{\sim \boldsymbol{\varepsilon}_2} + \cdots + \underbrace{m_k(\cdot)}_{\sim \boldsymbol{\varepsilon}_{k-1}} = m^{(k-1)}(\cdot) + m_k(\cdot). \tag{6}$$

Ici, l'erreur ε est vue comme la différence entre y et le modèle $m(\boldsymbol{x})$, mais elle peut aussi être vue comme le gradient associé à la fonction de perte quadratique. Formellement, $\boldsymbol{\varepsilon}$ peut être vu comme un $\nabla \ell$ dans un contexte plus géneral (on retrouve ici une interprétation qui fait penser aux résidus dans les modèles linéaires généralisés).

L'équation (6) peut se voir comme une descente du gradient, mais écrit de manière duale. En effet, la descente de gradient permettant d'obtenir le minimum d'une fonction f repose sur une équation de la forme :

$$\underbrace{f(\boldsymbol{x}_k)}_{\langle f, \boldsymbol{x}_k \rangle} \sim \underbrace{f(\boldsymbol{x}_{k-1})}_{\langle f, \boldsymbol{x}_{k-1} \rangle} + \underbrace{(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})}_{\alpha_k} \underbrace{\nabla f(\boldsymbol{x}_{k-1})}_{\langle \nabla f, \boldsymbol{x}_{k-1} \rangle}$$

Le problème (5) est dual dans le sens où $\boldsymbol{x}$ est ici donné, mais la fonction f doit être optimisée. On pourrait alors écrire une descente de gradient de la forme :

$$\underbrace{f_k(\boldsymbol{x})}_{\langle f_k, \boldsymbol{x} \rangle} \sim \underbrace{f_{k-1}(\boldsymbol{x})}_{\langle f_{k-1}, \boldsymbol{x} \rangle} + \underbrace{(f_k - f_{k-1})}_{\beta_k} \underbrace{\star}_{\langle f_{k-1}, \nabla \boldsymbol{x} \rangle}$$

où le terme $\star$ peut être interprété comme un gradient, mais dans un espace fonctionnel, et non plus dans $\mathbb{R}^p$. Le problème (6) va alors se réécrire comme un problème d'optimisation :

$$m^{(k)}(\boldsymbol{x}) = m^{(k-1)}(\boldsymbol{x}) + \underset{f \in \mathcal{M}_L}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, m^{(k-1)}(\boldsymbol{x}) + f(\boldsymbol{x})) \right\} \tag{7}$$

où l'astuce consiste à considérer un espace $\mathcal{M}_L$ relativement simple (on parlera de "*weak learner*"). Classiquement, les fonctions $\mathcal{M}_L$ sont des fonctions en escalier (que l'on retrouvera dans les arbres de classification et de régression) appelés *stumps*. Afin de s'assurer que l'apprentissage est effectivement lent, il n'est pas rare d'utiliser un paramètre de "shrinkage", et au lieu de poser, par exemple, $\varepsilon_1 = y - m_1(\boldsymbol{x})$, on posera $\varepsilon_1 = y - \alpha \cdot m_1(\boldsymbol{x})$ avec $\alpha \in [0,1]$.

On notera que c'est parce qu'on utilise pour $\mathcal{M}_L$ un espace non-linéaire, et que l'apprentissage est lent, que cet algorithme fonctionne bien. Dans le cas du modèle linéaire Gaussien, rappelons en effet que les résidus $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ sont orthogonaux aux variables explicatives, $\boldsymbol{X}$, et il est alors impossible d'apprendre de nos erreurs. La principale difficulté est de s'arrêter à temps, car après trop d'itérations, ce n'est plus la fonction m que l'on approxime, mais le bruit. Ce problème est appelé sur-apprentissage.

3.5 Sur-apprentissage et pénalisation

Le sur-apprentissage signifie que l'on construit un modèle trop complexe, qui aura de faibles qualités prédictives sur un nouvel échantillon (on parlera alors de généralisation). Nous avions évoqué cette idée dans

la section 2.7 au travers du principe de parcimonie, populaire en économétrie. Le critère d'Akaike était basé sur une pénalisation de la vraisemblance en tenant compte de la complexité du modèle (le nombre de variables explicatives retenues). Si en économétrie, il est d'usage de maximiser la vraisemblance (pour construire un estimateur asymptotiquement sans biais), et de juger de la qualité du modèle ex-post en pénalisant la vraisemblance, la stratégie ici sera de pénaliser ex-ante dans la fonction objectif, quitte à construire un estimateur biaisé. Typiquement, on va construire :

$$(\widehat{\beta}_{0,\lambda}, \widehat{\boldsymbol{\beta}}_\lambda) = \operatorname{argmin} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\| \right\}, \tag{8}$$

pour une norme $\|\cdot\|$ choisie a priori, et un paramètre de pénalisation λ (on retrouve en quelque sorte la distinction entre AIC et BIC). Dans cas de la norme ℓ_2, on retrouve l'estimateur Ridge, et pour la norme ℓ_1, on retrouve l'estimateur LASSO (*Least Absolute Shrinkage and Selection Operator*).

La pénalisation utilisée auparavant faisait intervenir le nombre de degrés de liberté du modèle, il peut alors paraître surprenant de faire intervenir $\|\boldsymbol{\beta}\|_{\ell_2}$ comme dans la régression Ridge. On peut toutefois envisager une vision Bayésienne de cette pénalisation. Rappelons que dans un modèle Bayésien :

$$\underbrace{\mathbb{P}[\boldsymbol{\theta}|\boldsymbol{y}]}_{\text{a posteriori}} \propto \underbrace{\mathbb{P}[\boldsymbol{y}|\boldsymbol{\theta}]}_{\text{vraisemblance}} \cdot \underbrace{\mathbb{P}[\boldsymbol{\theta}]}_{\text{a priori}} \quad \text{soit} \quad \log \mathbb{P}[\boldsymbol{\theta}|\boldsymbol{y}] = \underbrace{\log \mathbb{P}[\boldsymbol{y}|\boldsymbol{\theta}]}_{\text{log vraisemblance}} + \underbrace{\log \mathbb{P}[\boldsymbol{\theta}]}_{\text{pénalisation}}.$$

Dans un modèle linéaire Gaussien, si on suppose que la loi *a priori* de $\boldsymbol{\theta}$ suit une loi normale centrée, on retrouve une pénalisation basée sur une forme quadratique des composantes de $\boldsymbol{\theta}$.

Notons que si les estimateurs sans biais sont importants en statistique mathématique, ils ne sont souvent pas optimaux pour un critère de perte quadratique moyenne. Par exemple, si on considère un échantillon i.i.d. $\{y_1, \cdots, y_n\}$ de loi $\mathcal{N}(\theta, \sigma^2)$ et que l'on cherche un estimateur proportionnel à la moyenne, de la forme $\widehat{\theta} = \alpha\overline{y}$, le cas $\alpha = 1$ correspond à l'estimateur du maximum de vraisemblance, et à l'estimateur de la méthode des moments. Pourtant, ce n'est pas l'estimateur qui va minimiser la perte quadratique. Puisque :

$$\text{MSE}[\widehat{\theta}] = \underbrace{(\alpha-1)^2\theta^2}_{\text{biais}[\widehat{\theta}]^2} + \underbrace{\frac{\alpha^2\sigma^2}{n}}_{\text{Var}[\widehat{\theta}]},$$

la valeur optimale est alors : $\alpha^\star = \theta^2 \cdot \left(\theta^2 + \dfrac{\sigma^2}{n}\right)^{-1} < 1$. Autrement dit, pénaliser un estimateur sans biais est naturel, si l'objectif est de minimiser l'erreur quadratique moyenne.

Le problème d'optimisation décrit par l'équation (8) peut se voir comme la minimisation d'un Lagrangien, obtenu à l'aide d'un programme d'optimisation sous contrainte, de la forme :

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|\leq s}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) \right\}.$$

En grande dimension, la sélection de variables peut s'écrire à l'aide de cet algorithme, dans le contexte des modèles sparses : supposons que l'on dispose d'un grand nombre de variables explicatives, p, mais que beaucoup sont juste du bruit, au sens où $\beta_j = 0$ pour un grand nombre de j. Soit s le nombre de covariables pertinentes, $s = \#\mathcal{S}$ avec $\mathcal{S} = \{j = 1, \cdots, p; \beta_j \neq 0\}$. Si on note $\boldsymbol{X}_\mathcal{S}$ la matrice constituée des variables pertinentes, alors on suppose que le vrai modèle est de la forme $y = \boldsymbol{x}_\mathcal{S}^\mathsf{T}\boldsymbol{\beta}_\mathcal{S} + \varepsilon$. En définissant la norme ℓ_0 par $\|\boldsymbol{a}\|_{\ell_0} = \sum \mathbf{1}(a_i \neq 0)$, notons que $\|\boldsymbol{\beta}\|_{\ell_0} = s$, et on cherche ici à résoudre :

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_0}\leq s}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) \right\}$$

Ce problème d'optimisation avec une contrainte discrète peut être difficile à résoudre si p et s sont grands. La solution alternative proposée par Tibshirani (1996) consiste à résoudre :

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_1}\leq s}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) \right\}$$

Ces deux programmes d'optimisation, ainsi que la régression Ridge, sont visualisés sur la Figure 1. Comme le montre la Figure, la solution du problème avec la norme ℓ_0 coincide avec celui de la norme ℓ_1, mais s'avère bien plus simple à résoudre.

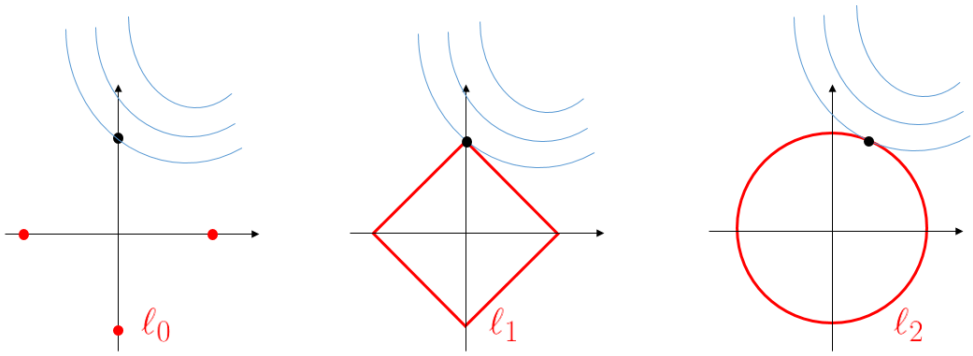


Figure 1: Pénalisation basée sur la norme ℓ_0, ℓ_1 et ℓ_2 de $\boldsymbol{\beta}$, respectivement.

Si on suppose les variables centrées, et si on note $\widehat{\boldsymbol{\beta}}_\lambda^{\mathsf{lasso}}$ l'estimateur LASSO (obtenu avec une pénalisation basée sur la norme ℓ_1), on peut montrer que :

$$\widehat{\boldsymbol{\beta}}_\lambda^{\mathsf{lasso}} \sim \big(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{\Delta}\big)^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$$

où $\boldsymbol{\Delta} = \text{diag}[|\widehat{\boldsymbol{\beta}}_{j,\lambda}^{\mathsf{lasso}}|^{-1}]$ si $\widehat{\boldsymbol{\beta}}_{j,\lambda}^{\mathsf{lasso}} \neq 0$, et 0 sinon. Aussi :

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda^{\mathsf{lasso}}] \sim \big(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{\Delta}\big)^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}$$

et :

$$\text{Var}[\widehat{\boldsymbol{\beta}}_\lambda^{\mathsf{lasso}}] \sim \sigma^2\big(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{\Delta}\big)^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\big(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{\Delta}\big)^{-1}\boldsymbol{X}^\mathsf{T}$$

On a ainsi un estimateur biaisé, mais qui peut être de variance suffisamment faible pour que l'erreur quadratique moyenne soit plus faible qu'en utilisant des moindres carrés.

3.6 *In-sample* et *out-of-sample*

Si ces techniques permettent d'éviter *a priori* le sur-apprentissage, le choix du paramètre de pénalisation λ (ou plus généralement le choix entre deux modèles) doit se faire à l'aide d'un critère (classiquement la somme des pertes ℓ_2 en régression, coïncidant avec la déviance) mais sur des données qui n'ont pas servi à calibrer le modèle.

Considérons un modèle linéaire pour simplifier, et posons ici $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}((\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n))$. Alors le risque (ou ici la déviance) *in sample* (IS) s'écrit :

$$\widehat{\mathcal{R}}^{\mathsf{IS}} = \text{Deviance}_{\mathsf{IS}}(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}[y_i - \boldsymbol{x}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}((\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n))]^2$$

Il n'est alors pas possible d'utiliser le théorème central limite, car les composantes ne sont pas indépendantes, et il n'est pas possible de garantir que :

$$\frac{\text{Deviance}_{\mathsf{IS}}(\widehat{\boldsymbol{\beta}})}{n} \to \mathbb{E}\left([Y - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}]^2\right) \text{ lorsque } n \to \infty.$$

La stratégie naturelle est alors de calculer une déviance *out-of-sample* (OS) :

$$\widehat{\mathcal{R}}^{\mathsf{OS}} = \text{Deviance}_{\mathsf{OS}}(\widehat{\boldsymbol{\beta}}) = \sum_{i=n+1}^{m+n}[y_i - \boldsymbol{x}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}((\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)]^2$$

On peut noter que :

$$\text{deviance}_{\mathsf{IS}} - \text{deviance}_{\mathsf{OS}} \approx 2 \cdot \nu$$

où ν représente le nombre de degrés de libertés, qui n'est pas sans rappeler la pénalisation utilisée dans le critère d'Akaike.

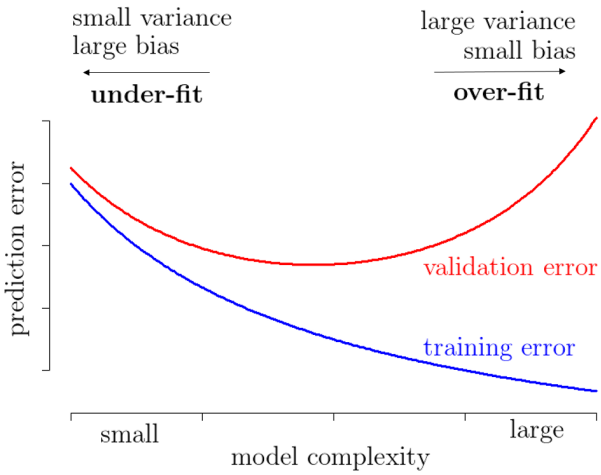


Figure 2: Généralisation et Sur-Apprentissage

Plus généralement, on peut noter, à partir d'une population de taille n,

$$\widehat{\mathcal{R}}^{\mathsf{IS}} = \frac{1}{n_1}\sum_{i=1}^{n_1} \ell(y_i, \widehat{m}(\boldsymbol{x}_i)) \quad \text{et} \quad \widehat{\mathcal{R}}^{\mathsf{OS}} = \frac{1}{n-n_1}\sum_{i=n_1+1}^{n} \ell(y_i, \widehat{m}(\boldsymbol{x}_i))$$

où les premières n_1 observations sont utilisées pour calibrer le modèle $m(\cdot)$.

En pratique, il n'est pas rare de séparer (aléatoirement) l'échantillon de départ en deux, en gardant $2/3$ des observations pour calibrer, et $1/3$ pour valider. La Figure 2 montre qu'en complexifiant le modèle (en augmentant ν), les erreurs diminuent sur la base d'apprentissage, la base qui sert à calibrer le modèle, mais le comportement est bien différent sur une base d'autres observations (données de validation). Complexifier le modèle (augmenter ν) permet - au début - de construire un meilleur modèle, mais assez vite, on atteint une phase de sur-apprentissage, le modèle commençant à expliquer le bruit, et non plus la tendance.

À la fin des années 60, Vapnik et Chernovenkis utilisent un modèle probabiliste pour montrer qu'un "bon" modèle reposait sur un compromis entre le biais, et la complexité. Cette complexité peut être vue comme la dimension ν utilisée en économétrie, mais avec un formalisme plus difficile à interpréter, liée au cardinal du plus grand ensemble de points que l'algorithme de classification peut "pulvériser" (notion de *shattered sets* introduite par Vapnik & Chervonenkis (1971)). En *machine learning*, on parlera de dimension de Vapnik-Chernovenkis, dite VC, pour décrire la complexité d'un modèle de classification. Par exemple pour un modèle de régression polynômiale simple, $m(x_i) = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \varepsilon_i$, la dimension VC est $p+1$, ce qui correspond à la dimension de la matrice $\boldsymbol{X}$ utilisée pour la régression, évoquée dans la section 2.4.

3.7 Techniques de validation croisée

Si la validation croisée est une technique classique en apprentissage, il convient de se souvenir qu'une approche similaire avait été introduite par Quenouille (1949) et formalisé par Quenouille (1956) et Tukey (1958) en statistique, pour réduire le biais. En effet, si on suppose que $\{y_1, \cdots, y_n\}$ est un échantillon tiré suivant une loi F_θ, et que l'on dispose d'un estimateur $T_n(\boldsymbol{y}) = T_n(y_1, \cdots, y_n)$, mais que cet estimateur est biaisé, avec $\mathbb{E}[T_n(\boldsymbol{Y})] = \theta + O\left(n^{-1}\right)$, il est possible de réduire le biais en considérant :

$$\widetilde{T}_n(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} T_{n-1}(\boldsymbol{y}_{(i)}) \quad \text{avec} \quad \boldsymbol{y}_{(i)} = (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n).$$

On peut alors montrer que $\mathbb{E}[\widetilde{T}_n(\boldsymbol{Y})] = \theta + O\left(n^{-2}\right)$.

L'idée de la validation croisée repose sur l'idée de construire un estimateur en enlevant une observation. Comme on souhaite construire un modèle prédictif, on va comparer la prévision obtenu avec le modèle estimé, et l'observation manquante :

$$\widehat{\mathcal{R}}^{\mathsf{CV}} = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \widehat{m}_{(i)}(\boldsymbol{x}_i))$$

On parlera ici de méthode *leave-one-out* (LOOCV).

On utilise classiquement cette technique pour trouver le paramètre optimal dans les méthodes de lissage exponentiel, pour des séries chronologiques. Dans le lissage simple, on va construire une prédiction de la forme ${}_t\widehat{y}_{t+1} = \alpha \cdot {}_{t-1}\widehat{y}_t + (1-\alpha)\cdot y_t$, avec $\alpha \in [0,1]$, et on va considérer :

$$\alpha^\star = \underset{\alpha\in[0,1]}{\operatorname{argmin}} \left\{ \sum_{t=2}^{T} \ell({}_{t-1}\widehat{y}_t, y_t) \right\},$$

comme le décrit Hyndman *et al.* (2009).

Le principal problème de la méthode *leave-one-out* est qu'elle nécessite de calibrer n modèles, ce qui peut être problématique en grande dimension.

Une méthode alternative est la validation croisée par k-blocs (dit *k-fold cross validation*) qui consiste à utiliser une partition de $\{1,\cdots,n\}$ en k groupes (ou blocs) de même taille, $\mathcal{I}_1,\cdots,\mathcal{I}_k$, et notons $\mathcal{I}_{\overline{j}} = \{1,\cdots,n\}\backslash\mathcal{I}_j$. En notant $\widehat{m}_{(j)}$ construit sur l'échantillon $\mathcal{I}_{\overline{j}}$, on pose alors :

$$\widehat{\mathcal{R}}^{k-\mathsf{CV}} = \frac{1}{k}\sum_{j=1}^{k}\mathcal{R}_j \quad \text{où} \quad \mathcal{R}_j = \frac{k}{n}\sum_{i\in\mathcal{I}_j}\ell(y_i, \widehat{m}_{(j)}(\boldsymbol{x}_i)).$$

La validation croisée standard, où une seule observation est enlevée à chaque fois (LOOCV), est un cas particulier, avec $k = n$. Utiliser $k = 5, 10$ a un double avantage par rapport à $k = n$: (1) le nombre d'estimations à effectuer est beaucoup plus faible, 5 ou 10 plutôt que n ; (2) les échantillons utilisés pour l'estimation sont moins similaires et donc, moins corrélés les uns aux autres, ce qui tend à éviter les excès de variance, commme le rappelle James *et al.* (2013).

Une autre alternative consiste à utiliser des échantillons boostrappés. Soit $\mathcal{I}_b$ un échantillon de taille n obtenu en tirant avec remise dans $\{1,\cdots,n\}$ pour savoir quelles observations $(y_i, \boldsymbol{x}_i)$ seront gardées dans la population d'apprentissage (à chaque tirage). Notons $\mathcal{I}_{\overline{b}} = \{1,\cdots,n\}\backslash\mathcal{I}_b$. En notant $\widehat{m}_{(b)}$ construit sur l'échantillon $\mathcal{I}_b$, on pose alors :

$$\widehat{\mathcal{R}}^{\mathsf{B}} = \frac{1}{B}\sum_{b=1}^{B}\mathcal{R}_b \quad \text{où} \quad \mathcal{R}_b = \frac{n_{\overline{b}}}{n}\sum_{i\in\mathcal{I}_{\overline{b}}}\ell(y_i, \widehat{m}_{(b)}(\boldsymbol{x}_i)),$$

où $n_{\overline{b}}$ est le nombre d'observations qui n'ont pas été conservées dans $\mathcal{I}_b$. On notera qu'avec cette technique, en moyenne $e^{-1} \sim 36.7\%$ des observations ne figurent pas dans l'échantillon boostrappé, et on retrouve un ordre de grandeur des proportions utilisées en créant un échantillon de calibration, et un échantillon de test. En fait, comme l'avait montré Stone (1977), la minimization du AIC est à rapprocher du critère de validation croisée, et Shao (1997) a montré que la minimisation du BIC correspond à de la valisation croisée de type k-fold, avec $k = n/\log n$.

4 Quelques algorithmes de *machine learning*

4.1 Réseaux de Neurones

Le premier neurone artificiel fonctionnel fut introduit par Franck Rosenblatt au milieu du XXième siècle, dans Rosenblatt (1958). Ce neurone qualifié de nos jours d'élémentaire porte le nom de Perceptron. Il a permis dans ses premières utilisations à déterminer le sexe d'un individu présenté aux travers d'une photo. Si ce premier neurone est important c'est qu'il introduit le premier formalisme mathématique d'un neurone biologique. On peut décrire un neurone artificiel par analogie avec une cellule nerveuse :

- les synapses apportant l'information à la cellule sont formalisés par un vecteur réel. La dimension du vecteur d'entrée du neurone (qui n'est d'autre qu'une fonction) correspond biologiquement au nombre de connections synaptiques;

- chaque signal apporté par un synapse est ensuite analysé par la cellule. Mathématiquement, ce schéma est transcrit par la pondération des différents éléments constitutifs du vecteur d'entrée;

- en fonction de l'information acquise, le neurone décide de retransmettre ou non un signal aux travers l'axome. Ce phénomène est répliqué par l'introduction d'une fonction d'activation. Le signal de sortie est modélisé par un nombre réel calculé comme image par la fonction d'activation du vecteur d'entrée pondéré.

Ainsi, un neurone artificiel est un modèle semi-paramétrique. Le choix de la fonction d'activation est en effet laissé à l'utilisateur. Nous introduisons dans le paragraphe qui suit une formalisation rigoureuse qui nous permettra de poser le modèle. On peut alors définir un neurone élémentaire formellement par :

- un espace d'entrée $\mathcal{X}$ généralement $\mathbb{R}^k$ avec $k \in \mathbb{N}^*$;
- un espace de sortie $\mathcal{Y}$ généralement $\mathbb{R}$ ou un ensemble dénombrable fini (classiquement $\{0,1\}$, mais on préférera ici $\{-1,1\}$);
- un vecteur de paramètres $\boldsymbol{w} \in \mathbb{R}^p$
- une fonction d'activation $\phi : \mathbb{R} \to \mathbb{R}$. Cette fonction doit être dans l'idéal monotone, dérivable et saturante afin de s'assurer de certaines propriétés de convergences

Cette dernière fonction ϕ fait penser à la transformation logistique ou probit, en économétrie (qui sont des fonctions de répartition puisque $\mathcal{Y}$ était l'ensemble $\{0,1\}$). Pour les réseaux de neurones, on utilisera plutôt la tangente hyperbolique, la fonction arctangente ou les fonctions sigmoïdes pour des problèmes de classification. On appellera neurone toute application f_w de $\mathcal{X}$ dans $\mathcal{Y}$ définie par :

$$f_{\boldsymbol{w}}(x) = \phi(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X}$$

Pour le perceptron introduit par Rosenblatt (1958), on assimile un neurone élémentaire à la fonction :

$$y = f_{\boldsymbol{w}}(\boldsymbol{x}) = \text{signe}(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathcal{X}$$

On remarque que selon cette formalisation, beaucoup de modèles statistiques comme par exemple les régressions logistiques pourraient être vus comme des neurones. En effet si l'on regarde d'un peu plus près, tout modèle GLM (*Generalized Linear Model*) pourrait s'interpréter comme un neurone formel où la fonction d'activation ϕ n'est d'autre que l'inverse de la fonction de lien canonique (par exemple). Si g désigne la fonction de lien du GLM , $\boldsymbol{w}$ le vecteur de paramètres, y la variable à expliquer et $\boldsymbol{x}$ le vecteur des variables explicatives de même dimension que $\boldsymbol{w}$:

$$g(\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]) = \boldsymbol{w}^\mathsf{T}\boldsymbol{x}$$

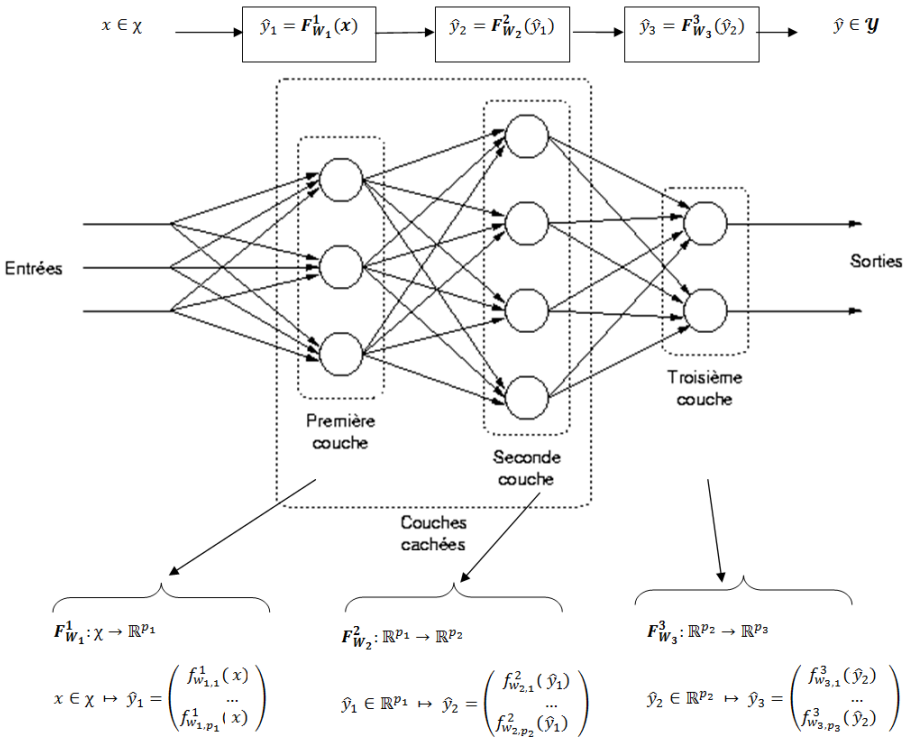
On retrouve la modélisation neuronale en prenant $\phi = g^{-1}$. Cependant, là où réside la différence majeure entre les GLM et le modèle neuronale est que ce dernier n'introduit aucune hypothèse de distribution sur $Y|\boldsymbol{X}$ (on n'a d'ailleurs pas besoin d'introduire ici de modèle probabiliste). D'autre part, lorsque le nombre de neurones par couche augmente, la convergence n'est pas nécessairement garantie si la fonction d'activation ne vérifie pas certaines propriétés (qu'on ne retrouve pas dans la majorité des fonctions de liens canoniques des GLM). Cependant, comme énoncé précédemment, la théorie des réseaux de neurones introduit des contraintes mathématiques supplémentaires sur la fonction g (Cybenko (1989)). Ainsi par exemple, une régression logistique peut être perçue comme un neurone. Mais les régressions linéraires ne vérifient pas toutes les hypothèses nécessaires.

Toujours par analogie avec le fonctionnement du système nerveux, il est alors possible de connecter différents neurones entre eux. Par soucis de calibrage, il est courant d'observer des structures de réseaux de neurones par couche. Chaque couche de neurones recevant à chaque fois le même vecteur d'observation. Ainsi un réseau de neurones à une couche par exemple sera constitué de plusieurs neurones recevant tous le même signal d'entrée. De façon plus formelle, chaque couche peut être associée à une fonction de transfert qui prend en entrée un vecteur et fournit en sortie un autre vecteur. Dans le cadre des réseaux de neurones, chaque fonction de transfert est une fonction paramétrique avec un nombre de paramètres proportionnel aux nombres de neurones de chaque couche. On introduit les notations suivantes :

- $K \in \mathbb{N}^*$: nombre de couches;
- $\forall k \in \{1, \cdots K\}$, p_k représente le nombre de neurones dans la couche k;
- $\forall k \in \{1, \cdots K\}$,W_k désigne la matrice des paramètres associés à la couche k. Plus précisément, W_k est une matrice $p_k \times p_{k-1}$ et pour tout $l \in \{1, \cdots p_k\}$, $w_{k,l} \in \mathbb{R}^{p_{k-1}}$ désigne le vecteur de poids associé au neurone élémentaire l de la couche k;
- on appellera $W = \{W_1, .., W_K\}$, l'ensemble des paramètres associés au réseau de neurones.
- $F^k_{W_k} : \mathbb{R}^{p_{k-1}} \to \mathbb{R}^{p_k}$ désigne la fonction de transfert associé à la couche k. Pour des raisons de simplification, on pourra également écrire F^k;

- $\hat{y}_k \in \mathbb{R}^{p_k}$ représentera le vecteur image de la couche $k \in \{1, \cdots, K\}$;
- on appellera $F = F_W = F^1 \circ \cdots \circ F^K$ la fonction de transfert associée au réseau global. A ce titre, si $\boldsymbol{x} \in \mathcal{X}$, on pourra noter $\widehat{\boldsymbol{y}} = F_W(\boldsymbol{x})$.

Figure 3: Exemple de notations associées aux réseaux de neurones multicouche.



La Figure 3 permet d'illustrer les notations présentées. Chaque cercle représente un neurone élémentaire. Chaque rectangle englobant plusieurs cercles représente une couche. On parle de couche d'entrée pour la première couche prenant en "input" les observation $\boldsymbol{x} \in \mathcal{X}$, de couche de sortie pour la couche fournissant en "output" la prédiction $\hat{\boldsymbol{y}} \in \mathcal{Y}$. Les autres couches sont couramment appelées couches cachées.

Un réseau de neurones multicouches est donc également un modèle semi-paramétrique dont les paramètres sont l'ensemble des composantes des matrices W_k pour tout entier k de $\{1, \cdots, K\}$. Chaque fonction d'activation associée à chaque neurone (chaque cercle de la Figure 3, tiré de http://intelligenceartificielle.org) est à déterminer par l'utilisateur.

Une fois que les paramètres à calibrer du modèle sont identifiés (ici les réels constituant les matrices W_k pour chaque couche $k \in \{1, \cdots, K\}$) , il est nécessaire de fixer une fonction de perte ℓ. En effet, on rappelle que l'objectif de l'apprentissage supervisé sur une base d'apprentissage de $n \in \mathbb{N}^*$ couples $(y_i, \boldsymbol{x}_i) \in \mathcal{Y} \times \mathcal{X}$ est de minimiser le risque empirique :

$$\widehat{\mathcal{R}}_n(F_W) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, F_W(\boldsymbol{x}_i))$$

Pour une structure de réseau de neurones fixée (c'est-à-dire nombre de couches, nombre de neurones par couches et fonctions d'activation fixés), le programme revient donc à déterminer l'ensemble de paramètres $W^* = (W_1, ..., W_K)$ de sorte que :

$$W^* \in \underset{W=(W_1,\ldots,W_K)}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, F_W(\boldsymbol{x}_i)) \right\}$$

De cette formule apparaît l'importance du choix de la fonction ℓ. Appelée fonction de perte, elle quantifie l'erreur moyenne commise par notre modèle F_W sur la base d'apprentissage. *A priori* ℓ peut être choisie

arbitrairement. Cependant, dans l'optique de résoudre un programme d'optimisation, on préféra des fonctions de coût sous-différentiables et convexes afin de garantir la convergence des algorithmes d'optimisation. Parmi les fonctions de perte classiques, en plus de la fonction de perte quadratique ℓ_2 on retiendra la fonction dite *Hinge* - $\ell(y, \hat{y}) = \max(0, 1 - y\hat{y})$ - ou la fonction dite logistique - $\ell(y, \hat{y}) = \log(1 - e^{-y\hat{y}})$.

En définitive les réseaux de neurones sont des modèles semi-paramétriques dont le nombre de paramètres est croissant avec le nombre de couches et de neurones par couche. Il est laissé à l'utilisateur de choisir les fonctions d'activation et la structure du réseau.

Les réseaux de neurones ont été utilisés très tôt en économie et en finance, en particulier sur les défauts d'entreprises - Tam & Kiang (1992) ou Altman *et al.* (1994) - ou plus récemment la notation de crédit - Blanco *et al.* (2013) ou Khashman (2011). Cependant les structures telles que présentées précédemment sont généralement limitées. L'apprentissage profond (ou *deep learning*) caractérise plus particulièrement des réseaux de neurones plus complexes (parfois plus d'une dizaine de couches avec parfois des centaines de neurones par couche). Si aujourd'hui ces structures sont très populaires en analyse du signal (image, texte, son) c'est qu'elles sont capables à partir d'une quantité d'observations très importante d'extraire des informations que l'humain ne peut percevoir et de faire face à des problèmes non linéaires, comme le rappelle LeCun *et al.* (2015).

L'extraction d'informations peut, par exemple, se faire grâce à la convolution. Procédé non supervisé, il a permis notamment d'obtenir d'excellente performance dans l'analyse d'image. Techniquement, cela peut s'apparenter à une transformation à noyaux (comme utilisé dans les techniques SVM). Si une image peut être perçue comme une matrice dont chaque coordonnée représente un pixel, une convolution reviendrait à appliquer une transformation sur un point (ou une zone) de cette matrice générant ainsi une nouvelle donnée. Le procédé peut ainsi être répété en appliquant des transformations différentes (d'où la notion de couches convolutives). Le vecteur final obtenu peut alors enfin alimenter un modèle neuronal comme introduit dans le paragraphe précédant. En fait, plus généralement, une couche de convolution peut être perçue comme un filtre qui permet de transformer la donnée initiale.

Une explication intuitive pour laquelle l'apprentissage approfondi, en particulier les réseaux nerveux profonds, est si puissant pour décrire des relations complexes dans les données, c'est leur construction autour de l'approximation fonctionnelle simple et l'exploitation d'une forme de hiérarchie, comme le note Lin *et al.* (2016). Néanmoins les modèles de *deep learning* sont plus difficiles à appréhender car ils nécessitent beaucoup de jugement empirique. En effet, si aujourd'hui les librairies open sources (keras, torch etc.) permettent de paralléliser plus facilement les calculs en utilisant par exemple les GPU (*Graphical Processor Units*), il reste néanmoins à l'utilisateur de déterminer la structure du réseau de neurones le plus approprié.

4.2 Support Vecteurs Machine

Les modèles de classification binaire en économétrie sont associés à des observations dans l'ensemble $\{0, 1\}$, correspondant aux valeurs prises par une variable indicatrice, dont la distribution suit une loi de Bernoulli. En apprentissage statistique, comme en traitement du signal, on préférera avoir des observations dans l'ensemble $\{-1, +1\}$.

Cortes & Vapnik (1995) ont posé les bases théorique des modèles dit SVM, proposant une alternative aux réseaux de neurones alors très populaires comme algorithme de classification dans la communauté de l'apprentissage machine. L'idée initiale des méthodes de *Support Vectors Machine* (SVM) consiste à trouver un hyperplan séparateur divisant l'espace en deux ensembles de points le plus homogène possible (i.e. contenant des labels identiques). En dimension deux, l'algorithme consiste à déterminer une droite séparant l'espace en deux zones les plus homogènes possibles. La résolution de ce problème possédant parfois une infinité de solution (il peut en effet exister une infinité de droites qui séparent l'espace en deux zones distinctes et homogènes), on rajoute généralement une contrainte supplémentaire. L'hyperplan séparateur doit se trouver le plus éloigné possible des deux sous-ensembles homogènes qu'il engendre. On parlera ainsi de marge. L'algorithme ainsi décrit est alors un SVM linéaire à marge.

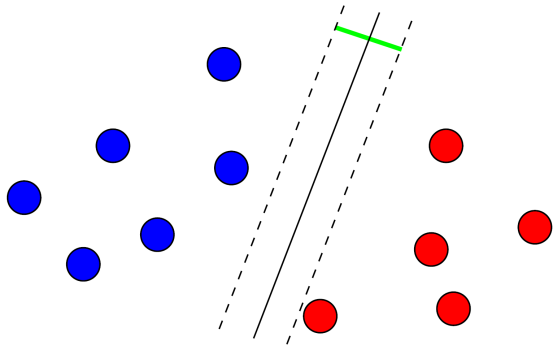
Si un plan peut être caractérisé entièrement par un vecteur directeur $\boldsymbol{w}$ orthogonal à ce dernier et une constante b, appliquer un algorithme SVM à un ensemble de $n \in \mathbb{N}^*$ points $\boldsymbol{x}_i$ de $\mathbb{R}^p$ labellisés par $y_i \in \{-1, 1\}$ revient alors à résoudre un programme d'optimisation sous contrainte similaire à celui d'un Ridge ou LASSO. Plus particulièrement, on sera amené à résoudre :

$$(\boldsymbol{w}^\star, b^\star) = \underset{\boldsymbol{w}, b}{\operatorname{argmin}} \|\boldsymbol{w}\|$$

$$\text{sous contrainte } \ \forall i \in \{1, \cdots, n\}, \ y_i(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b) - 1 \geq 0$$

La contrainte peut être relâchée en autorisant que dans un sous-ensemble, un point puisse ne pas être du

Figure 4: Schéma d'illustration d'un SVM à marge, source: cours d'apprentissage statistique en génomique de Jean-Philippe Vert



même label que la majeure partie des points de ce sous-ensemble à condition de ne pas être trop loin de la frontière. C'est ce qu'on appelle les SVM linéaire à marge légère (*soft margin*).

S'il n'est pas possible de séparer les points, une astuce consiste à les transformer dans une dimension supérieure, de sorte que les données deviennent alors linéairement séparables. Trouver la bonne transformation qui sépare les données est toutefois très difficile. Cependant, il existe une astuce mathématique pour résoudre ce problème avec élégance, en définisant les transformations $T(\cdot)$ et les produits scalaires via un noyau $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle T(\boldsymbol{x}_1), T(\boldsymbol{x}_2) \rangle$. L'un des choix les plus courants pour une fonction de noyau est la fonction de base radiale (noyau gaussien) $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\big(-\|\boldsymbol{x}_1 - \boldsymbol{x}_2)\|^2\big)$. Il n'existe néanmoins pas de règles à ce jour permettant de choisir le "meilleur" noyau.

4.3 Arbres, Bagging et Forêts Aléatoires

Les arbres de classification ont été introduits dans Quinlan (1986) mais c'est surtout Breiman (2001a) qui a assuré la popularité de l'algorithme. On parle de modèle CART pour *Classification And Regression Tree*. L'idée est de diviser consécutivement (par une notion de branchement) les données d'entrée jusqu'à ce qu'un critère d'affectation (par rapport à la variable cible) soit atteint, selon une règle prédéfinie.

L'intuition de la construction des arbres de classification est la suivante. L'entropie $H(\boldsymbol{x})$ est associée à la quantité de désordre dans les données $\boldsymbol{x}$ par rapport aux modalités prises par la variable de classification y, et chaque partition vise à réduire ce désordre. L'interprétation probabiliste est de créer les groupes les plus homogènes possible, en réduisant la variance par groupe (variance intra), ou de manière équivalente en créant deux groupes aussi différents que possible, en augmentant la variance entre les groupe (variance inter). À chaque étape, nous choisissons la partition qui donne la plus forte réduction de désordre (ou de variance). L'arbre de décision complet se développe en répétant cette procédure sur tous les sous-groupes, où chaque étape k aboutit à une nouvelle partition en 2 branches, qui subdivise notre ensemble de données en 2. Enfin, on décide quand mettre fin à cette constitution de nouvelles branches, en procédant à des affectations finales (nœuds dits foliaires). Il existe plusieurs options pour mettre fin à cette croissance. L'une est de construire un arbre jusqu'à ce toutes les feuilles soient pures, c'est à dire composé d'une seule observation. Une autre option est de définir une règle d'arrêt liée à la taille, ou à la décomposition, des feuilles. Les exemples de règles d'arrêt peuvent être d'une taille minimale (au moins 5 éléments par feuille), ou une entropie minimale. On parlera alors d'élagage de l'arbre.

À un nœud donné, constitué de n_0 observations $(\boldsymbol{x}_i, y_i)$ avec $i \in \mathcal{I}_0$, on va couper en deux branches (une à gauche et une à droite), partitionnant ainsi $\mathcal{I}_0$ en $\mathcal{I}_\text{g}$ et $\mathcal{I}_\text{d}$. Soit I le critère d'intérêt, comme l'entropie du nœud (ou plutôt du nœud vu en tant que feuille):

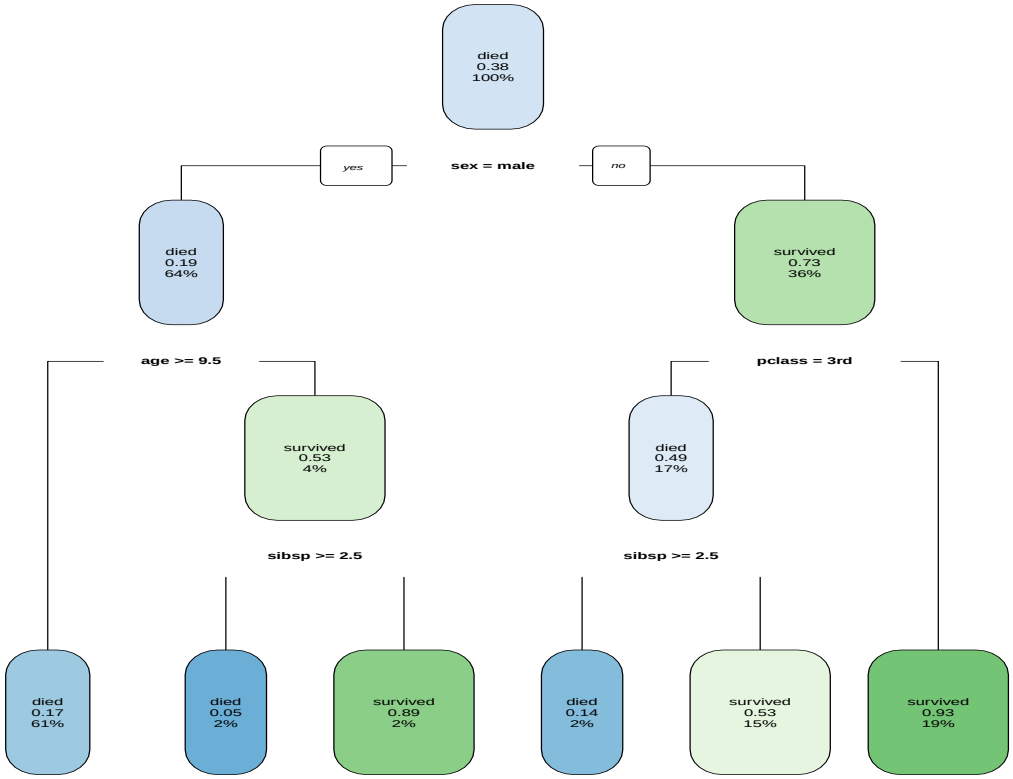
$$I(\boldsymbol{y}_0) = -n_0 p_0 \log p_0 \quad \text{où} \quad p_0 = \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} y_i,$$

ou la variance du nœud:

$$I(\boldsymbol{y}_0) = n_0 p_0 (1 - p_0) \quad \text{où} \quad p_0 = \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} y_i,$$

ce dernier étant également l'indice de Gini.

Figure 5: Schéma d'illustration d'un arbre de décision permettant de prédire le taux de survie d'un individu du Titanic.



On partitionnera entre la branche gauche et la branche droite si le gain $I(\boldsymbol{y}_0) - \big[I(\boldsymbol{y}_{\mathrm{g}}) + I(\boldsymbol{y}_{\mathrm{d}})\big]$ est suffisamment important. Lors de la construction des arbres, on va chercher la partition qui donne le gain le plus important possible. Ce problème combinatoire étant complexe, le critère suggéré par Breiman (2001a) est de considérer un découpage suivant une des variables, avec $\mathcal{I}_{\mathrm{g}} = \{i \in \mathcal{I}_0 : x_{k,i} < s\}$ et $\mathcal{I}_{\mathrm{d}} = \{i \in \mathcal{I}_0 : x_{k,i} > s\}$, pour une variable k et un seuil s (si la variable est continue, sinon on considère des regroupements de modalités pour des variables qualitatives).

Les arbres de décision ainsi décrits sont simples à obtenir et faciles à interpréter (comme le montre la Figure 5 sur les données du Titanic[2]), mais ils sont peu robustes, et leur pouvoir prédictif est souvent très faible, en particulier si l'arbre est très profond. Une idée naturelle est de développer un ensemble de modèles d'arbres à peu près indépendants, qui prédisent conjointement mieux qu'un modèle d'arbre unique. On va utiliser le bootstrap, en tirant (avec remise) n observations parmi $\{(\boldsymbol{x}_i, y_i)\}$. Ces forêts, une fois agrégées donnent souvent de bien meilleurs résultats que les arbres isolés, mais elles sont difficiles à interpréter. Ces techniques ressemblent toutefois beaucoup à ce qui est fait lorsque l'on utilise les techniques de bootrap en régression (par exemple pour construire des tubes de confiance dans une régression fonctionnelle).

Le principe du *bagging*, pour *bootstrap aggregating*, consiste à générer des échantillons aléatoires, en tirant avec remise dans l'échantillon d'origine, comme pour le bootstrap. Chaque échantillon ainsi généré permet d'estimer un nouvel arbre de classification, formant ainsi une forêt. C'est l'aggrégation de tous ces arbres qui conduit à la prévision. Le résultat global est moins sensible à l'échantillon initial et donne souvent de meilleurs résultats de prévisions.

Les forêts aléatoires, ou *random forests*, sont une amélioration du *bagging* lorsque les variables explicatives (ou covariables) sont corrélées ou colinéaires. C'est le même principe que le *bagging*, mais en plus, lors de la construction d'un arbre de classification, à chaque branche, un ensemble de m covariables est tiré aléatoirement. Autrement dit, chaque branche d'un arbre ne s'appuie pas sur le même ensemble de covariables. Cela permet d'atténuer les problèmes éventuels générés par la présence de covariables fortement colinéaires,

[2]Ce jeu de données, contenant des informations sur tous les passagers (et membres d'équipage) du Titanic, dont la variable y indiquant si la personne a survécu a été abondamment utilisé pour illustrer les techniques de classification, voir https://www.kaggle.com/c/titanic/data.

	$y=0$	$y=1$	
$\widehat{y}_s=0$	TN_s	FN_s	TN_s+FN_s
$\widehat{y}_s=1$	FP_s	TP_s	FP_s+TP_s
	TN_s+FP_s	FN_s+TP_s	n

Table 1: Matrice de confusion, ou tableau de contingence pour un seuil s donné.

qui peuvent être très génants lors de l'estimation d'un modèle de régression paramétrique. Finalement, l'idée du bootstrap est appliquée en faisant des tirages aléatoires dans les observations, mais aussi dans les covariables. Le premier permet de construire une forêt d'arbres, afin de stabiliser l'estimation, le second permet d'atténuer les problèmes éventuels issus d'une forte colinéarité entre les covariables.

4.4 Sélection de modèle de classification

Étant donné un modèle $m(\cdot)$ approchant $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]$, et un seuil $s\in[0,1]$, posons

$$\widehat{y}^{(s)}=\mathbf{1}[m(\boldsymbol{x})>s]=\begin{cases}1 \text{ si } m(\boldsymbol{x})>s\\ 0 \text{ si } m(\boldsymbol{x})\leq s\end{cases}$$

La matrice de confusion est alors le tableau de contingence associé aux comptages $\boldsymbol{N}=[N_{u,v}]$ avec

$$N_{u,v}^{(s)}=\sum_{i=1}^{n}\mathbf{1}(\widehat{y}_i^{(s)}=u,y_j=v)$$

pour $(u,v)\in\{0,1\}$. La Table 1 présente un tel tableau, avec le nom de chacun des éléments : TP (*true positive*) sont les vrais positifs, correspondant aux 1 prédit en 1, TN (*true negative*) sont les vrais négatifs, correspondant aux 0 prédit en 0, FP (*false positive*) sont les faux positifs, correspondant aux 0 prédit en 1, et enfin FN (*false negative*) sont les faux négatifs, correspondant aux 1 prédit en 0).

Plusieurs quantités sont dérivées de ce tableau. La sensibilité correspond à la probabilité de prédire 1 dans la population des 1, ou taux de vrais positifs. Elle est donnée par le ratio FP/(FP+TN). La spécificité est la probabilité de prédire 0 dans la population des 0 ou taux de vrais négatifs. On s'intéressera toutefois davantage au taux de faux négatifs, c'est à dire la probabilité de prédire 1 dans la population des 0. La représentation de ces deux valeurs lorsque s varie donne la courbe ROC (*receiver operating characteristic*) :

$$\text{ROC}_s=\left(\frac{\text{FP}_s}{\text{FP}_s+\text{TN}_s},\frac{\text{TP}_s}{\text{TP}_s+\text{FN}_s}\right)=(\mathsf{sensibility}_s,1-\mathsf{specificity}_s) \text{ pour } s\in[0,1].$$

Une telle courbe est présentée dans la partie suivante, sur des données réelles.

Les deux grandeurs intensivement utilisées en *machine learning* sont l'indice κ, qui compare la précision observée avec celle espérée, avec un modèle aléatoire (tel que décrit dans Landis & Koch (1977)) et l'AUC correspondant à l'aire sous la courbe ROC. Pour le premier indice, une fois choisi s, notons $\boldsymbol{N}^{\perp}$ le tableau de contingence correspond aux cas indépendants (défini à partir de $\boldsymbol{N}$ dans le test d'indépendance du chi-deux). On pose alors

$$\text{précision totale}=\frac{\text{TP+TN}}{n}$$

alors que

$$\text{précision aléatoire}=\frac{[\text{TN+FP}]\cdot[\text{TP+FN}]+[\text{TP+FP}]\cdot[\text{TN+FN}]}{n^2}$$

On peut alors définir

$$\kappa=\frac{\text{précision totale}-\text{précision aléatoire}}{1-\text{précision aléatoire}}$$

Classiquement s sera fixé égal à 0.5, comme dans une classification bayésienne naïve, mais d'autres valeurs peuvent être retenues, en particulier si les deux erreurs ne sont pas symmétriques (nous reviendrons sur ce point dans un exemple par la suite).

Il existe des compromis entre des modèles simples et complexes mesurés par leur nombre de paramètres (ou plus généralement les degrés de liberté) en matière de performance et de coût. Les modèles simples sont généralement plus faciles à calculer, mais peuvent conduire à des ajustements plus mauvais (avec un biais élevé par exemple). Au contraire, les modèles complexes peuvent fournir des ajustements plus précis,

mais risquent d'être coûteux en termes de calcul. En outre, ils peuvent surpasser les données ou avoir une grande variance et, tout aussi que des modèles trop simples, ont de grandes erreurs de test. Comme nous l'avons rappelé auparavant, dans l'apprentissage machine, la complexité optimale du modèle est déterminée en utilisant le compromis de biais-variance.

4.5 De la classification à la régression

Comme nous l'avons rappelé en introduction, historiquement, les méthodes de *machine learning* se sont orientées autour des problèmes de classification (avec éventuellement plus de 2 modalités [3]), et assez peu dans le cas où la variable d'intérêt y est continue. Néanmoins, il est possible d'adapter quelques techniques, comme les arbres et les forêts aléatoires, le boosting, ou les réseaux de neurones.

Pour les arbres de régression, Morgan & Sonquist (1963) ont proposé la méthode AID, basée sur la formule de décomposition de la variance de l'équation (4), avec un algorithme proche de celui de la méthode CART décrite auparavant. Dans le contexte de la classification, on calculait, à chaque nœud (dans le cas de l'indice de Gini) en sommant sur la feuille de gauche $\{x_{k,i} < s\}$ et celle de droite $\{x_{k,i} > s\}$

$$I = \sum_{i:x_{k,i}<s} \overline{y}_{\mathrm{g}}(1-\overline{y}_{\mathrm{g}}) + \sum_{i:x_{k,i}>s} \overline{y}_{\mathrm{d}}(1-\overline{y}_{\mathrm{d}})$$

où $\overline{y}_{\mathrm{g}}$ et $\overline{y}_{\mathrm{d}}$ désignent les fréquences de 1 dans la feuille de gauche et de droite, respectivement. Dans le cas d'un arbre de régression, on utilisera

$$I = \sum_{i:x_{k,i}<s} \big(y_i - \overline{y}_{\mathrm{g}}\big)^2 + \sum_{i:x_{k,i}>s} \big(y_i - \overline{y}_{\mathrm{d}}\big)^2$$

qui va correspondre à la somme (pondérée) des variances intra. Le partage optimal sera celui qui aura le plus de variance intra (on veut les feuilles les plus homogènes possibles) ou de manière équivalente, on veut maximiser la variance intra.

Dans le contexte des forêts aléatoires, on utilise souvent un critère majoritaire en classification (la classe prédite sera la classe majoritaire dans une feuille), alors que pour les régression, on utilise la moyenne des prédictions, sur tous les arbres.

Dans la partie précédente, nous avons présenté la dimension "apprentissage" du *machine learning* en présentant le *boosting*. Dans un contexte de régression (variable y continue), l'idée est de créer une succession de modèles en écrivant l'équation (7) sous la forme :

$$m^{(k)}(\boldsymbol{x}) = m^{(k-1)}(\boldsymbol{x}) + \alpha_k \underset{f\in\mathcal{F}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i, m^{(k-1)}(\boldsymbol{x}) + f(\boldsymbol{x}))^2 \right\}$$

où α_k est un paramètre de *shrinkage*, où le second terme correspond à un arbre de régression, sur les résidus, $y_i - m^{(k-1)}(\boldsymbol{x}_i)$.

Mais il existe existe d'autres techniques permettant d'apprendre de manière séquentielle. Dans un modèle additif (GAM) on va chercher une écriture de la forme

$$m(\boldsymbol{x}) = \sum_{j=1}^{p} m_j(x_j) = m_1(x_1) + \cdots + m_p(x_p)$$

L'idée de la poursuite de projection repose sur une décomposition non pas sur les variables explicatives, mais sur des combinaisons linéaires. On va ainsi considérer un modèle

$$m(\boldsymbol{x}) = \sum_{j=1}^{k} g_j(\boldsymbol{\omega}_j^{\mathsf{T}}\boldsymbol{x}) = g_1(\boldsymbol{\omega}_1^{\mathsf{T}}\boldsymbol{x}) + \cdots + g_k(\boldsymbol{\omega}_k^{\mathsf{T}}\boldsymbol{x}).$$

Tout comme les modèles additifs, les fonctions $g_1, \cdots, g_k$ sont à estimer, tout comme les directions $\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_k$. Cette écriture est relativement générale, et permet de tenir compte d'intéractions et d'effets croisés (ce que nous ne pouvions pas faire avec les modèles additifs qui ne tiennent compte que de non-linéarités). Par exemple en dimension 2, un effet multiplicatif $m(x_1, x_2) = x_1 \cdot x_2$ s'écrit

$$m(x_1, x_2) = x_1 \cdot x_2 = \frac{(x_1+x_2)^2}{4} - \frac{(x_1-x_2)^2}{4}$$

[3] Par exemple dans le cas de reconnaissance de lettres ou de chiffres

autrement dit $g_1(x) = x^2/4$, $g_1(x) = -x^2/4$, $\boldsymbol{\omega}_1 = (1,1)^\mathsf{T}$ et $\boldsymbol{\omega}_1 = (1,-1)^\mathsf{T}$. Dans la version simple, avec $k = 1$, avec une fonction de perte quadratique, on peut utiliser un développement de Taylor pour approcher $[y_i - g(\boldsymbol{\omega}^\mathsf{T}\boldsymbol{x}_i)]^2$, et construire classiquement un algorithme itératif. Si on dispose d'une valeur initiale $\boldsymbol{\omega}_0$, notons que

$$\sum_{i=1}^{n}[y_i - g(\boldsymbol{\omega}^\mathsf{T}\boldsymbol{x}_i)]^2 \approx \sum_{i=1}^{n} g'(\boldsymbol{\omega}_0^\mathsf{T}\boldsymbol{x}_i)^2 \left[\boldsymbol{\omega}^\mathsf{T}\boldsymbol{x}_i + \frac{y_i - g(\boldsymbol{\omega}_0^\mathsf{T}\boldsymbol{x}_i)}{g'(\boldsymbol{\omega}_0^\mathsf{T}\boldsymbol{x}_i)} - \boldsymbol{\omega}^\mathsf{T}\boldsymbol{x}_i\right]^2$$

qui correspondrait à l'approximation dans les modèles linéaires généralisés sur la fonction $g(\cdot)$ qui était la fonction de lien (supposée connue). On reconnait un problème de moindres carrés pondérés. La difficulté ici est que les fonctions $g_j(\cdot)$ sont inconnues.

5 Applications

Les données massives ont rendu nécessaire le développement de techniques d'estimation permettant de pallier les limites des modèles paramétriques, jugés trop restrictifs, et des modèles non-paramétriques classiques, dont l'estimation peut être difficile en présence d'un nombre élevé de variables. L'*apprentissage statistique*, ou *apprentissage machine*, propose de nouvelles méthodes d'estimation non-paramétriques, performantes dans un cadre général et en présence d'un grand nombre de variables.[4] En outre, certaines de ces méthodes permettent d'atténuer le problème délicat posé à l'estimation classique par la présence simultanée de variables fortement corrélées. Toutefois, l'obtention d'une plus grande flexibilité s'obtient au prix d'un manque d'interprétation qui peut être important.

En pratique, une question importante est de savoir quel est le meilleur modèle ? La réponse à cette question dépend du problème sous-jacent. Si la relation entre les variables est correctement approximée par un modèle linéaire, un modèle paramétrique correctement spécifié devrait être performant. Par contre, si le modèle paramétrique n'est pas correctement spécifié, car la relation est fortement non-linéaire et/ou fait intervenir des effets croisés non-négligeables, alors les méthodes statistiques issues du *machine learning* devraient être plus performantes.

La bonne spécification d'un modèle de régression est une hypothèse souvent posée, elle est rarement vérifiée et justifiée. Dans les applications qui suivent, nous montrons comment les méthodes statistiques issues du *machine learning* peuvent être utilisées pour justifier la bonne spécification d'un modèle de régression paramétrique, ou pour détecter une mauvaise spécification. Des applications en classification sont présentées dans un premier temps, sections 5.1, 5.2 et 5.3. D'autres applications sont ensuite présentées dans le contexte de régression classique, sections 5.4 et 5.5.

5.1 Les ventes de sièges auto pour enfants (classification)

Nous reprenons ici un exemple utilisé dans James *et al.* (2013). Le fichier de données contient les ventes de sièges auto pour enfants dans 400 magasins (Sales), ainsi que plusieurs variables, dont la qualité de présentation en rayonnage (Shelveloc, égal à "mauvais", "moyen", "bon") et le prix (Price).[5] Une variable dépendante binaire est artificiellement crée, pour qualifier une forte vente ou non (High="oui" si Sales > 8 et à "non" sinon). Dans cette application, on cherche à évaluer les déterminants d'un bon niveau de vente.

Dans un premier temps, on considère un modèle de régression linéaire latent:

$$y^\star = \gamma + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta} + \varepsilon, \qquad \varepsilon \sim G(0,1), \tag{9}$$

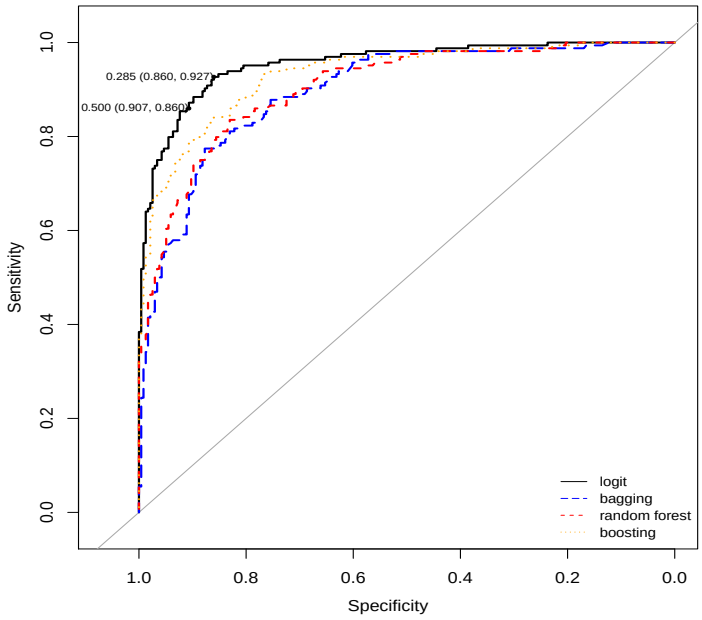
où $\boldsymbol{x}$ est composé de k variables explicatives, $\boldsymbol{\beta}$ est un vecteur de k paramètres inconnus et ε est un terme d'erreur *i.i.d.* avec une fonction de répartition G d'espérance nulle et de variance égale à un. La variable dépendante y° n'est pas observé, mais seulement y, avec:

$$y = \begin{cases} 1 & \text{si } y^\star > \xi, \\ 0 & \text{si } y^\star \leq \xi. \end{cases} \tag{10}$$

On peut alors exprimer la probabilité d'avoir y égal à 1, comme suit :

$$\mathbb{P}(Y = 1) = G(\beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) \tag{11}$$

[4] Entre autres, voir Hastie *et al.* (2009) et James *et al.* (2013).
[5] C'est le fichier de données Carseats de la librairie ISLR.



	AUC
logit	0.9544
bagging	0.8973
random forest	0.9050
boosting	0.9313

Figure 6: Ventes de sièges auto: courbes ROC et aires sous la courbe (AUC).

où $\beta_0 = \gamma - \xi$.[6] L'estimation de ce modèle se fait par maximum de vraisemblance en sélectionnant *a priori* une loi paramétrique G. Si on suppose que G est la loi Normale, c'est un modèle probit, si on suppose que G est la loi logistique, c'est un modèle logit. Dans un modèle logit/probit, il y a deux sources potentielles de mauvaise spécification :

(i) la relation linéaire $\beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$ est mal spécifiée

(ii) la loi paramétrique utilisée G n'est pas la bonne

En cas de mauvaise spécification, de l'une ou l'autre sorte, l'estimation n'est plus valide. Le modèle le plus flexible est le suivant :

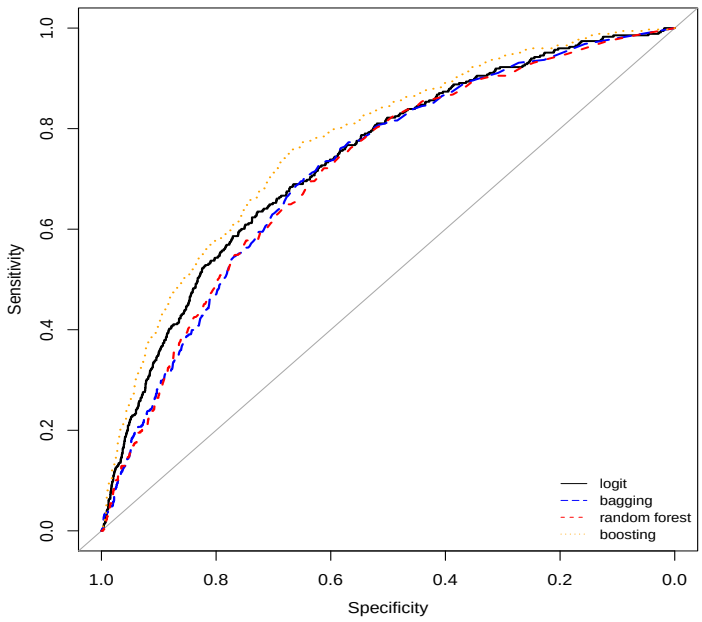
$$\mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}] = G(h(\boldsymbol{x})) \tag{12}$$

où h est une fonction inconnue et G une fonction de répartition inconnue. Les modèles de *bagging*, de forêt aléatoire et de *boosting* permettent d'estimer ce modèle général sans faire de choix *à priori* sur la fonction h et sur la distribution G. L'estimation du modèle logit/probit est néanmoins plus performante si h et G sont correctement spécifiés.

Nous estimons le modèle (11) avec la loi logistique pour G, et le modèle (12) avec les méthodes de *bagging*, de forêt aléatoire et de *boosting*. Nous faisons une analyse de validation croisée par 10 blocs. Les probabilités individuelles des données *out-of-sample*, c'est à-dire de chacun des blocs non-utilisée pour l'estimation, sont utilisées pour évaluer la qualité de la classification.

La figure 6 présente la courbe ROC, ainsi que l'aire sous la courbe (AUC), pour les estimations logit, bagging, random forest et boosting. La *courbe ROC* est un graphique qui représente simultanément la qualité de la prévision dans les deux classes, pour des valeurs différentes du seuil utilisé pour classer les individus (*cutoff*). Une manière naturelle de classer les individus consiste à les attribuer dans la classe pour laquelle ils ont la plus grande probabilité estimée. Dans le cas d'une variable binaire, cela revient à prédire la classe d'appartenance pour laquelle la probabilité estimée est supérieure à 0.5. Mais un autre seuil pourrait être utilisé. Par exemple, dans la figure 6, un point de la courbe ROC du modèle logit indique qu'en prenant un seuil égal à 0.5, la réponse "non" est correctement prédite à 90.7% (specificity), et la réponse "oui" à 86% (sensitivity). Un autre point indique qu'en prenant un seuil égal à 0.285, la réponse "non" est correctement prédite à 86% (specificity), et la réponse "oui" à 92.7% (sensitivity). Comme décrit auparavant, un modèle de classification idéal aurait une courbe ROC de la forme Γ. Autrement dit, le meilleur modèle est celui dont

[6] $\mathbb{P}[Y = 1] = \mathbb{P}[Y^\star > \xi] = \mathbb{P}[\gamma + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta} + \varepsilon > \xi] = \mathbb{P}[\varepsilon > \xi - \gamma - \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}] = \mathbb{P}[\varepsilon < \gamma - \xi + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}]$. En posant $\gamma - \xi = \beta_0$, on obtient $\mathbb{P}[Y = 1] = G(\beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta})$. En général, on suppose que le terme d'erreur est de variance σ^2, auquel cas les paramètres du modèle (11) deviennent β_0/σ et $\boldsymbol{\beta}/\sigma$, ce qui veut dire que les paramètres du modèle latent (9) ne sont pas identifiables, ils sont estimés à un paramètre d'échelle près.



	AUC
logit	0.7372
bagging	0.7198
random forest	0.7154
boosting	0.7691

Figure 7: Achat d'assurance: courbes ROC et aires sous la courbe (AUC).

la courbe est au-dessus des autres. Un critère souvent utilisé pour sélectionner le meilleur modèle est celui dont l'aire sous la courbe ROC est la plus grande (AUC). L'avantage d'un tel critère est qu'il est simple à comparer et qu'il ne dépend pas du choix du seuil de classification.

Dans notre exemple, la courbe ROC du modèle logit domine les autres courbes, et son aire sous la courbe est la plus grande (AUC=0.9544). Ces résultats indiquent que ce modèle fournit les meilleures prévisions de classification. N'étant dominé par aucun autre modèle, ce constat suggère que le modèle linéaire logit est correctement spécifié et qu'il n'est pas utile d'utiliser un modèle plus général et plus complexe.

5.2 L'achat d'une assurance caravane (classification)

Nous reprenons à nouveau un exemple utilisé dans James *et al.* (2013). Le fichier de données contient 85 variables sur les charactéristiques démographiques de 5822 individus.[7] La variable dépendante (Purchase) indique si l'individu a acheté une assurance caravane, c'est une variable binaire, égale à "oui" ou "non". Dans le fichier de données, seulement 6% des individus ont pris une telle assurance. Les classes sont donc fortement désequilibrées.

Nous estimons le modèle (11) avec la loi logistique et le modèle (12) avec les méthodes *bagging*, forêt aléatoire et *boosting* (les paramètres de *tuning* sont ceux de James *et al.* (2013), n.trees=1000 et shrinkage=0.01). Nous faisons une analyse de validation croisée par 10 blocs. Les probabilités individuelles des données *out-of-sample*, c'est à-dire de chacun des blocs non-utilisée pour l'estimation, sont utilisées pour évaluer la qualité de la classification.

La figure 7 présente la courbe ROC, ainsi que l'aire sous la courbe (AUC), pour les estimations logit, bagging, random forest et boosting. La courbe du modèle boosting domine les autres courbes, son aire sous la courbe est la plus grande (AUC=0.7691). Ces résultats indiquent que le boosting fournit les meilleures prévisions de classification. Notons que, comparées à l'exemple précédent, les courbes sont assez éloignées de la forme en coude, ce qui suggère que la classification ne sera pas aussi bonne.

Il faut faire attention aux résultats d'une classification standard, c'est-à-dire avec un seuil de classification égal à 0.5, qui est souvent pris par défaut dans les logiciels (la prédiction de la réponse de l'individu i est "non" 1 si la probabilité estimée qu'il réponde "non" est supérieure à 0.5, sinon c'est "oui"). La partie gauche du tableau 2 présente les taux de classifications correctes avec ce seuil (0.5 cutoff), pour les différentes méthodes. Avec le meilleur modèle et le seuil standard (boosting et seuil à 0.5), les réponses "non" sont correctes à 99.87% (spécificité, specificity) et les réponses "oui" sont toutes fausses (sensitivité, sensitivity). Autrement dit, cela équivaut à utiliser un modèle qui prédit que personne n'achète d'assurance caravane. Sélectionner un tel modèle est absurde pour l'analyste, qui est surtout intéressé par les 6% des individus qui en ont pris

[7] C'est le fichier de données Caravan de la librairie ISLR sous R.

	0.5 cutoff		cutoff optimal	
	spécificité	sensitivité	spécificité	sensitivité
logit	0.9967	0.0057	0.7278	0.6351
bagging	0.9779	0.0661	0.6443	0.7069
random forest	0.9892	0.0316	0.6345	0.6954
boosting	0.9987	0.0000	0.6860	0.7385

Table 2: Achat d'assurance: sensibilité au choix du seuil de classification.

une. Ce résultat s'explique par la présence de classes fortement déséquilibrées. En effet, dans notre exemple, en prévoyant que personne n'achète d'assurance, on fait "seulement" 6% d'erreur. Mais ce sont des erreurs qui conduisent à ne rien expliquer.

Plusieurs méthodes peuvent être utiles pour pallier à ce problème, lié aux classes fortement désequilibrées (pour plus d'informations, voir Kuhn & Johnson (2013), chapitre 16). Une solution simple consiste à utiliser un seuil de classification différent. La courbe ROC présente les résultats en fonction de plusieurs seuils de classification, où la classification parfaite est illustrée par le couple (specificity, sensitivity)=(1,1), c'est à-dire par le coin supérieur gauche dans le graphique. Aussi, on choisit comme seuil de classification optimal celui qui correspond au point de la courbe ROC qui est le plus proche du point (1,1), ou du coin supérieur gauche. La partie droite du tableau 2 présente les taux de classifications correctes avec les seuils optimaux (*optimal cutoff*), pour les différentes méthodes (les seuils optimaux des méthodes logit, *bagging*, forêt aléatoire et *boosting* sont, respectivement, égaux à $0.0655, 0.0365, 0.0395, 0.0596$). Avec le boosting et un seuil optimal, les réponses "non" sont correctes à 68.6% (specificity) et les réponses "oui" à 73.85% (sensitivity). L'objet de l'analyse étant de prévoir correctement les individus susceptibles d'acheter une assurance caravane (classe "oui"), et les distinguer suffisamment des autres (classe "non"), le choix du seuil optimal est beaucoup plus performant que le seuil standard 0.5. Notons qu'avec un modèle logit et un seuil optimal, le taux de classifications correctes de la classe "non" est de 72.78%, celui de la classe "oui" est de 63.51%. Par rapport au boosting, le logit prédit un peu mieux la classe "non", mais nettement moins bien la classe "oui".

5.3 Les défauts de remboursement de crédits particuliers (classification)

Considérons la base allemande de crédits particuliers, utilisée dans Nisbet, Elder & Miner (2001) et Tufféry (2001), avec 1000 observations, et 19 variables explicatives, dont 12 qualitatives c'est à dire, en les disjonctant (en créant une variable indicatrice pour chaque modalité), 48 variables explicatives potentielles.

Une question récurrente en modélisation est de savoir quelles sont les variables qui mériteraient d'être utilisées. La réponse la plus naturelle pour un économètre pourrait être une méthode de type *stepwise* (parcourir toutes les combinaisons possibles de variables étant a priori un problème trop complexe en grande dimension). La suite des variables dans une approche *forward* est présentée dans la première colonne du tableau 3. Une approche mentionnée avant qui peut être utile est le LASSO, qui, en pénalisant convenablement la norme ℓ_1 du vecteur de paramètres $\boldsymbol{\beta}$. On peut ainsi, séquentiellement, trouver les valeurs du paramètre de pénalisation λ, qui permet d'avoir une variable explicative supplémentaire, non nulle. Ces variables sont présentées dans la dernière colonne. On note que les deux première variable considéré comme non nulle (pour un λ assez grand) sont les deux première à ressortir lors d'une procédure *forward*. Enfin, une dernière méthode a été proposée par Breiman (2001b), en utilisant tous les arbres créé lors de la construction d'une forêt aléatoire : l'importance de la variable x_k dans une forêt de T arbres est donnée par:

$$\text{Importance}(x_k) = \frac{1}{T}\sum_{t=1}^{n}\sum_{j\in N_{t,k}} p_t(j)\Delta\mathcal{I}(j)$$

où $N_{t,k}$ désigne l'ensemble des noeuds de l'arbre t utilisant la variable x_k comme variable de séparation, $p_t(j)$ désigne la proportion des observations au nœud j, et $\Delta(j)$ est la variation d'indice au nœud j (entre le nœud précédant, la feuille de gauche et celle de droite). Dans la colonne centrale du tableau 3 sont présentées les variables par ordre d'importance décroissante, lorsque l'indice utilisé est l'indice de Gini.

Avec l'approche *stepwise* et l'approche LASSO, on reste sur des modèles logistiques linéaires. Dans le cas des forêtes aléatoires (et des arbres), des intéractions entre variables peuvent être prises en compte, lorsque 2 variables sont présentes. Par exemple la variable residence_since est présente très haut parmi les variables prédictive (troisième variable la plus importante) sans rester dans une procédure *stepwise*, et devenant non nulle très tard.

Stepwise	AIC	Random Forest	Gini	Lasso
checking_statusA14	1112.1730	checking_statusA14	30.818197	checking_statusA14
credit_amount(4e+03,Inf]	1090.3467	installment_rate	20.786313	credit_amount(4e+03,Inf]
credit_historyA34	1071.8062	residence_since	19.853029	credit_historyA34
installment_rate	1056.3428	duration(15,36]	11.377471	duration(36,Inf]
purposeA41	1044.1580	credit_historyA34	10.966407	credit_historyA31
savingsA65	1033.7521	credit_amount	10.964186	savingsA65
purposeA43	1023.4673	existing_credits	10.482961	housingA152
housingA152	1015.3619	other_payment_plansA143	10.469886	duration(15,36]
other_payment_plansA143	1008.8532	telephoneA192	10.217750	purposeA41
personal_statusA93	1001.6574	age	10.071736	installment_rate
savingsA64	996.0108	savingsA65	9.547362	property_magnitudeA124
other_partiesA103	991.0377	checking_statusA12	9.502445	age(25,Inf]
checking_statusA13	985.9720	housingA152	8.757095	checking_statusA13
checking_statusA12	982.9530	jobA173	8.734460	purposeA43
employmentA74	980.2228	personal_statusA93	8.715932	other_partiesA103
age(25,Inf]	977.9145	property_magnitudeA123	8.634527	employmentA72
purposeA42	975.2365	personal_statusA92	8.438480	savingsA64
duration(15,36]	972.5094	purposeA43	8.362432	employmentA74
duration(36,Inf]	966.7004	employmentA73	8.225416	purposeA46
purposeA49	965.1470	employmentA75	8.089682	personal_statusA93
purposeA410	963.2713	duration(36,Inf]	8.029945	personal_statusA92
credit_historyA31	962.1370	purposeA42	8.025749	savingsA63
purposeA48	961.1567	property_magnitudeA122	7.908813	telephoneA192

Table 3: Crédit: choix de variables, tri séquentiel, par approche *stepwise*, par fonction d'importance dans une forêt aléatoire et par LASSO.

5.4 Les déterminants des salaires (régression)

Afin d'expliquer les salaires (individuels) en fonction du niveau d'étude, de l'expérience de la personne, et son genre, il est classique d'utiliser l'équation de salaire de Mincer - décrite dans Mincer (1974) - tel que le rappelle Lemieux (2006):

$$\log(\mathsf{wage}) = \beta_0 + \beta_1\,\mathsf{ed} + \beta_2\,\mathsf{exp} + \beta_3\,\mathsf{exp}^2 + \beta_4\,\mathsf{fe} + \varepsilon \tag{13}$$

où ed est le niveau d'études, ex l'expérience professionnelle et fe une variable indicatrice, égale à 1 si l'individu est une femme et à 0 sinon. D'après la théorie du capital humain, le salaire espéré augmente avec l'expérience, de moins en moins vite, pour atteindre un maximum avant de diminuer. L'introduction du carré de exp permet de prendre en compte une telle relation. La présence de la variable fe permet quand à elle de mesurer une éventuelle discrimination salariale entre les hommes et les femmes.

Le modèle (13) impose une relation linéaire entre le salaire et le niveau d'étude, et une relation quadratique entre le salaire et l'expérience professionnelle. Ces relations peuvent paraître trop restrictives. Plusieurs études montrent notamment que le salaire ne diminue pas après un certain age, et qu'une relation quadratique ou un polynôme de degré plus élevé est plus adapté (comme décrit dans Murphy & Welch (1990) et Bazen & Charni (2015)).

Le modèle (13) impose également que la différence salariale entre les hommes et les femmes est indépendente du niveau d'étude et de l'expérience. Il est trop restrictif si, par exemple, on suspecte que l'écart de salaire moyen entre les hommes et les femmes est faible pour les postes non-qualifiés et fort pour les postes qualifiés, ou faible en début de carrière et fort en fin de carrière (*effets d'intéractions*).

Le modèle le plus flexible est le modèle entièrement non-paramétrique :

$$\log(\mathsf{wage}) = m(\mathsf{ed},\ \mathsf{exp},\ \mathsf{fe}) + \varepsilon \tag{14}$$

où $m(\cdot)$ est une fonction quelconque. Il a l'avantage de pouvoir tenir compte de relations non-linéaires quelconques et d'intéractions complexes entre les variables. Mais, sa grande flexibilité se fait au détriment d'une interprétation plus difficile du modèle. En effet, il faudrait un graphique en 4-dimensions pour représenter la fonction m. Une solution consiste à représenter la fonction m en 3 dimensions, en fixant la valeur de l'une des variables, mais la fonction représentée peut être très différente avec une valeur fixée différente.

Nous utilisons les données d'une enquête de l'US Census Bureau daté de mai 1985, issues de l'ouvrage de Berndt (1990) et disponibles sour R.[8] Nous estimons les deux modèles et utilisons une analyse de validation

[8] C'est le fichier de données CPS1985 de la librairie AER.

	Modèle (13)	Modèle (14)			
$\widehat{\mathcal{R}}^{10-\mathsf{CV}}$	OLS	Splines	Bagging	R.Forest	Boosting
out-of-sample	0.2006	0.2004	0.2762	0.2160	0.2173

Table 4: Salaire: analyse de validation croisée par blocs ($K = 10$) : performances de l'estimation des modèles linéaire (13) et entièrement non-paramétrique (14).

	Modèle (15)	Modèle (16)		
$\widehat{\mathcal{R}}^{10-\mathsf{CV}}$	OLS	Bagging	R.Forest	Boosting
in-sample	21.782	1.867	1.849	7.012
out-of-sample	24.082	9.590	9.407	11.789

Table 5: Prix des logements à Boston: analyse de validation croisée par blocs ($K = 10$): performances de l'estimation des modèles linéaire (15) et entièrement non-paramétrique (16).

croisées par 10 blocs pour sélectionner la meilleure approche. Le modèle paramétrique (13) est estimé par Moindres Carrés Ordinaires (OLS). Le modèle entièrement non-paramétrique (14) est estimé par la méthode des *splines*, car il en comprend peu de variables, ainsi que par les méthodes *bagging*, *random forest* et *boosting*.

Le tableau 4 présente les résultats de la validation croisée en 10 blocs (*10-fold cross-validation*). Le meilleur modèle est celui qui minimise le critère $\widehat{\mathcal{R}}^{10-\mathsf{CV}}$. Les résultats montrent que le modèle (13) est au moins aussi performant que le modèle (14), ce qui suggère que le modèle paramétrique (13) est correctement spécifié.

5.5 Les déterminants des prix des logements à Boston (régression)

Nous reprenons ici l'un des exemples utilisé dans James *et al.* (2013), dont les données sont disponibles sous R. Le fichier de données contient les valeurs médianes des prix des maisons (medv) dans $n = 506$ quartiers autour de Boston, ainsi que 13 autres variables, dont le nombre moyen de pièces par maison (rm), l'age moyen des maisons (age) et le pourcentage de ménages dont la catégorie socio-professionnelle est peu élevée (lstat).[9]

Considérons le modèle de régression linéaire suivant :

$$\mathsf{medv} = \alpha + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon \tag{15}$$

où $\boldsymbol{x} = [\mathsf{chas},\mathsf{nox},\mathsf{age},\mathsf{tax},\mathsf{indus},\mathsf{rad},\mathsf{dis},\mathsf{lstat},\mathsf{crim},\mathsf{black},\mathsf{rm},\mathsf{zn},\mathsf{ptratio}]$ est un vecteur en dimension 13, $\boldsymbol{X}$ une matrice $n \times 13$ et $\boldsymbol{\beta}$ est un vecteur de 13 paramètres. Ce modèle spécifie une relation linéaire entre la valeur des maisons et chacune des variable explicatives.

Le modèle le plus flexible est le modèle entièrement non-paramétrique :

$$\mathsf{medv} = m(\boldsymbol{x}) + \varepsilon. \tag{16}$$

L'estimation de ce modèle avec les méthodes du noyau ou les splines peut être problématique, car le nombre de variables est relativement élevé ($\boldsymbol{X}$ est une matrice qui contient 13 variables), ou au moins trop élevé pour envisager estimer une surface en dimension 13. Nous estimons les deux modèles et utilisons une analyse de validation croisée par 10-blocs pour sélectionner la meilleure approche. Le modèle paramétrique (15) est estimé par Moindres Carrés Ordinaires (OLS) et le modèle entièrement non-paramétrique (16) est estimé par trois méthodes différentes: *bagging*, forêt aléatoire et *boosting* (nous utilisons ici les valeurs par défaut utilisées dans James *et al.* (2013), pp. 328-331).

Le tableau 5 présente les résultats de la validation croisée en 10 blocs (*10-fold cross-validation*). La première ligne (*in-sample*) présente la qualité de l'ajustement des modèles en utilisant seulement les données d'apprentissage, c'est-à-dire celles qui ont servi à estimer le modèle, pour calculer le MSE. La deuxième ligne (*out-of-sample*) présente la qualité de l'ajustement en utilisant d'autres données que celles ayant servies à estimer le modèle, pour calculer l'erreur quadratique. À partir des résultats in-sample, les méthodes de *bagging* et de *random forest* paraissent incroyablement plus performantes que l'estimation OLS du modèle

[9] C'est le fichier de données `Boston` de la librairie `MASS`. Pour une description complète des données, voir: https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Boston.html.

	%IncMSE	IncNodePurity
rm	61.35	18345.41
lstat	36.20	15618.22
dis	29.37	2601.72
nox	24.91	1034.71
age	17.86	554.50
ptratio	17.43	626.58
tax	16.60	611.37
crim	16.26	1701.73
indus	9.45	237.35
black	8.72	457.58
rad	4.53	166.72
zn	3.10	35.73
chas	0.87	39.05

Table 6: Prix des logements à Boston: mesures de l'importance de chacune des variables dans l'estimation *random forest* du modèle (16), en considérant tout l'échantillon.

linéaire (15), le critère $\widehat{\mathcal{R}}^{10-\mathsf{CV}}$ passant de 21.782 à 1.867 et 1.849. Les résultats *out-of-sample* vont dans le même sens, mais la différence est moins importante, le critère $\widehat{\mathcal{R}}^{10-\mathsf{CV}}$ passant de 24.082 à 9.59 et 9.407. Ces résultats illustrent un phénomène classique des méthodes non-linéaires, comme le *bagging* et la forêt aléatoire, qui peuvent être très performantes pour prédire les données utilisées pour l'estimation, mais moins performantes pour prédire des données hors-échantillon. C'est pourquoi la sélection de la meilleure estimation est habituellement basée sur une analyse *out-of-sample*, telle que présentée dans la deuxième ligne.

La différence entre l'estimation du modèle linéaire (15) et du modèle entièrement non-paramétrique (16) est importante (24.082 vs 9.590, 9.407 et 11.789). Un tel écart suggère que le modèle linéaire est mal spécifié, et que des relations non-linéaire et/ou des effets d'intéractions sont présentes dans la relation entre le prix des logements, medv, et les variables explicatives $\boldsymbol{x}$. Ce résultat nous conduit à chercher une meilleure spécification paramétrique.

À partir du modèle paramétrique (15), et afin de prendre en compte d'éventuelles non-linéarités, le modèle additif généralisé (GAM) suivant peut être considéré :

$$\mathsf{medv} = m_1(x_1) + m_2(x_2) + \cdots + m_{13}(x_{13}) + \varepsilon, \tag{17}$$

où $m_1, m_2, \ldots m_{13}$ sont des fonctions inconnues. L'avantage de ce modèle est qu'il permet de considérer n'importe quelle relation non-linéaire entre la variable dépendante et chacune des variables explicatives. De plus, il ne souffre pas du problème du fléau de la dimension, car chacune des fonction est de dimension 1, et il est facilement interprétable. Toutefois, il ne prend pas en compte d'éventuels effets d'intéractions.

L'estimation du modèle additif généralisé (17) par la méthode des *splines*, dans le cadre d'une analyse de validation croisée par 10-blocs, donne une valeur $\widehat{\mathcal{R}}^{10-\mathsf{CV}} = 13.643$. Par rapport au modèle paramétrique (15), il y a un gain important (13.643 vs. 24.082). Mais la différence avec le modèle entièrement non-paramétrique (16) reste conséquente (13.643 vs 9.590, 9.407, 11.789). Une telle différence suggère que la prise en compte de relations individuelles pouvant être fortement non-linéaires n'est pas suffisante, et que des effets d'intéractions entre les variables sont présents. Nous pourrions inclure dans le modèle les variables d'intéractions les plus simples entre toutes les paires de variables $(x_i \times x_j)$, mais cela impliquerait de rajouter un très grand nombre de variables au modèle initial (78 dans notre cas), qui ne serait pas sans conséquence sur la qualité de l'estimation du modèle. Quoi qu'il en soit, nous pouvons dire pour le moment que le modèle linéaire est mal spécifié et qu'il existe des effets d'intéractions pouvant être forts dans la relation entre medv et X, l'identification de tels effets restant délicat.

Afin d'aller plus loin, les outils développés en apprentissage statistique peuvent être à nouveau d'un grand recours. Par exemple, l'estimation *random forest* s'accompagne de mesures de l'importance de chacune des variables dans l'estimation du modèle (décrit dans la section précédante). Le tableau 6 présente ces mesures dans le cadre du modèle (16), estimé sur l'échantillon complet. Les résultats suggèrent que les variables rm et lstat sont les variables les plus importantes pour expliquer les variations des prix des logements medv. Ce constat nous conduit à enrichir la relation initiale, en rajoutant les intéractions liées à ces deux variables seulement, qui sont les plus importantes.

Nous estimons le modèle additif généralisé incluant les variables d'intéractions, sur l'échantillon complet:

$$\mathsf{medv} = m_1(x_1) + \cdots + m_{13}(x_{13}) + (\mathsf{rm}{:}x)\,\gamma + (\mathsf{lstat}{:}x)\,\delta + \varepsilon, \tag{18}$$

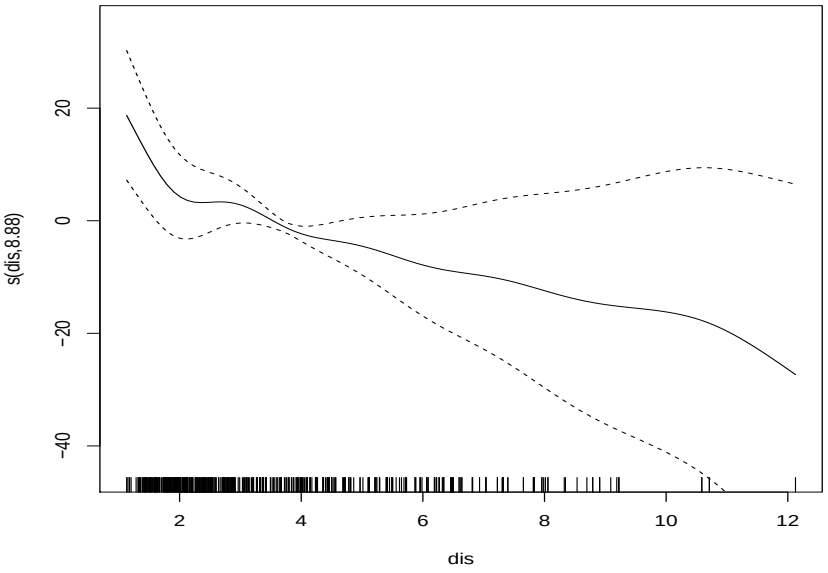


Figure 8: Estimation de la relation $m_7(x_7)$ dans le modèle additif généralisé (18), où $x_7 = \mathsf{dis}$.

où (rm:x) représente les variables d'intéractions de rm avec toutes les autres variables de x et (lstat:x) représente les variables d'intéractions de lstat avec toutes les autres variables de x.[10] L'analyse des résultats de cette estimation suggère que les fonctions $\hat{m}_i$ sont linéaires pour toutes les variables, sauf pour la variable dis, dont la relation estimée est présentée dans la figure 8. Cette variable mesure la distance moyenne à cinq centres d'emplois de la région. L'effet semble diminuer plus rapidement avec la distance, lorsque celle-ci n'est pas très élevée. Au delà d'une certaine distance (au delà de 2, en log), l'effet est réduit, il continue à diminuer mais plus doucement. Cette relation non-linéaire peut être approchée par une régression linéaire par morceaux en considérant un nœud.

Finalement, l'analyse précédente nous conduit à considérer le modèle linéaire suivant:

$$\mathsf{medv} = \alpha + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta} + (\mathsf{dis} - 2)_+ \theta + (\mathsf{rm}{:}x)\,\gamma + (\mathsf{lstat}{:}x)\,\delta + \varepsilon \tag{19}$$

où $(\mathsf{dis}-2)_+$ est égal à la valeur de son argument si ce dernier est positif. Par rapport au modèle linéaire initial, ce modèle inclut une relation linéaire par morceaux avec la variable dis, ainsi que des effets d'intéractions entre rm, lstat et chacune des autres variables de $\boldsymbol{x}$.

Le tableau 7 présente les résultats de la validation croisée en 10 blocs (*10-fold cross-validation*) de l'estimation des modèles paramétriques (15) et (19), estimés par Moindres Carrés Ordinaires (OLS), et du modèle additif généralisé (17) estimé par les splines. Il montre que l'ajout des variables d'intéractions et de la relation linéaire par morceaux dans le modèle (19) donne des résultat beaucoup plus performants que le modèle initial (15): le critère $\widehat{\mathcal{R}}^{10-\mathsf{CV}}$ est divisé par plus de deux, il passe de 24.082 à 11.759. En comparant ces résultats avec ceux du tableau 5, on constate également que le modèle paramétrique (19), estimé par OLS, est aussi performant que le modèle général (16) estimé par *boosting* ($\widehat{\mathcal{R}}^{10-\mathsf{CV}} = 11.789$). La différence avec les méthodes *bagging* et forêt aléatoire n'est quant à elle pas très importante ($\widehat{\mathcal{R}}^{10-\mathsf{CV}} = 9.59, 9.407$)

Finalement, les méthodes *bagging*, forêt aléatoire et *boosting* ont permis de mettre en évidence une mauvaise spécification du modèle paramétrique initial, puis de trouver un modèle paramétrique beaucoup plus performant, en prenant compte des effets de non-linéarités et d'intéractions appropriées.

6 Conclusion

Si les "deux cultures" (ou les deux communautés) de l'économétrie et du *machine learning* se sont développées en parallèle, le nombre de passerelles entre les deux ne cesse d'augmenter. Alors que Varian (2014) présentait

[10] On a (rm:x) = [$\mathsf{rm}\times\mathsf{chas}$, $\mathsf{rm}\times\mathsf{nox}$, $\mathsf{rm}\times\mathsf{age}$, $\mathsf{rm}\times\mathsf{tax}$, $\mathsf{rm}\times\mathsf{indus}$, $\mathsf{rm}\times\mathsf{rad}$, $\mathsf{rm}\times\mathsf{dis}$, $\mathsf{rm}\times\mathsf{lstat}$, $\mathsf{rm}\times\mathsf{crim}$, $\mathsf{rm}\times\mathsf{black}$, $\mathsf{rm}\times\mathsf{zn}$, $\mathsf{rm}\times\mathsf{ptratio}$] et ($\mathsf{lstat}$:$x$) = [$\mathsf{lstat}\times\mathsf{chas}$, $\mathsf{lstat}\times\mathsf{nox}$, $\mathsf{lstat}\times\mathsf{age}$, $\mathsf{lstat}\times\mathsf{tax}$, $\mathsf{lstat}\times\mathsf{indus}$, $\mathsf{lstat}\times\mathsf{rad}$, $\mathsf{lstat}\times\mathsf{dis}$, $\mathsf{lstat}\times\mathsf{crim}$, $\mathsf{lstat}\times\mathsf{black}$, $\mathsf{lstat}\times\mathsf{zn}$, $\mathsf{lstat}\times\mathsf{ptratio}$].

$\widehat{\mathcal{R}}^{10-\mathsf{CV}}$	Modèle (15) OLS	Modèle (17) Splines	Modèle (19) OLS
out-of-sample	24.082	13.643	11.759

Table 7: Prix des logements à Boston: analyse de validation croisée par blocs ($K = 10$) : performances de l'estimation du modèle linéaire (15) et du modèle linéaire (19) incluant les effets d'intéractions et une non-linéarité par morceaux.

les apports importants de l'économétrie à la communauté du *machine learning*, nous avons tenté ici de présenter des concepts et des outils développés au fil du temps par ces derniers, qui pourraient être utiles aux économètres, dans un contexte d'explosion du volume de données. Si nous avons commencé par opposer ces deux mondes, c'est aussi pour mieux comprendre leurs forces et leurs faiblesses. Les fondements probabilistes de l'économétrie sont incontestablement sa force, avec non seulement une interprétabilité des modèles, mais aussi une quantification de l'incertitude. Néanmoins, nous l'avons vu à plusieurs reprises sur des données réelles, les performances prédictives des modèles de *machine learning* sont intéressantes, car elles permettent de mettre en avant une mauvaise spécification d'un modèle économétrique. De la même manière que les techniques non-paramétriques permettent d'avoir un point de référence pour juger de la pertinence d'un modèle paramétrique, les outils de *machine learning* permettent d'améliorer un modèle économétrique, en détectant un effet non-linéaire ou un effet croisé oublié.

Une illustration des interactions possibles entre les deux communautés se trouve par exemple dans Belloni *et al.* (2010, 2012), dans un contexte de choix d'instrument dans une régression. Reprenant les données de Angrist & Krueger (1991) sur un problème de réussite scolaire, ils montrent comment mettre en œuvre efficacement les techniques d'économétrie instrumentale quand on peut choisir parmi 1530 instruments disponibles (problème qui deviendra récurrent avec l'augmentation du volume de données). Comme nous l'avons vu tout au long de cet article, même si les approches peuvent être fondamentalement différentes dans les deux communautés, bon nombre d'outils développés par la communauté du *machine learning* méritent d'être utilisés par les économètres.

References

Ahamada, I. & E. Flachaire (2011). Non-Parametric Econometrics. Oxford University Press.

Aigner, D., Lovell, C.A.J & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, **6**, 21–37.

Aldrich, J. (2010). The Econometricians' Statisticians, 1895-1945. *History of Political Economy*, **42** 111–154.

Altman, E., Marco, G. & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience).*Journal of Banking & Finance* **18**, 505–529.

Angrist, J.D. & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics*, **114**, 533–575.

Angrist, J.D. & Pischke, J.S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspective*, **24**, 3–30.

Angrist, J.D. & Pischke, J.S. (2015). Mastering Metrics. Princeton University Press.

Angrist, J.D. & Krueger, A.B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics*, **106**, 979–1014.

Bajari, P., Nekipelov, D., Ryan, S.P. & Yang, M. 2015. Machine learning methods for demand estimation. *American Economic Review*, **105** 481–485.

Bazen, S. & K. Charni (2015). Do earnings really decline for older workers? AMSE 2015-11 Discussion Paper, Aix-Marseille University.

Bellman, R.E. (1957). Dynamic programming. Princeton University Press.

Belloni, A., Chernozhukov, V. & Hansen, C. (2010). Inference Methods for High-Dimensional Sparse Econometric Models. *Advances in Economics and Econometrics*, 245–295

Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica*, **80**, 2369–2429.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**:289–300.

Berger, J.O. (1985). Statistical decision theory and Bayesian Analysis (2nd ed.). Springer-Verlag.

Berk, R.A. (2008). Statistical Learning from a Regression Perspective. Springer Verlag.

Berkson, J. (1944). Applications of the logistic function to bioassay. *Journal of the American Statistical Association*, **9**, 357–365.

Berkson, J. (1951). Why I prefer logits to probits. *Biometrics*, **7** (4), 327–339.

Bernardo, J.M. & Smith, A.F.M. (2000). Bayesian Theory. John Wiley.

Berndt, E. R. (1990). The Practice of Econometrics: Classic and Contemporary. Addison Wesley.

Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer Verlag.

Blanco, A. Pino-Mejias, M., Lara, J. & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from peru. *Expert Systems with Applications*, **40**, 356–364.

Bliss, C.I. (1934). The method of probits. *Science*, **79**, 38–39.

Breiman, L. (2001a). Statistical Modeling: The Two Cultures. *Statistical Science*, **16**:3, 199–231.

Breiman, L. (2001b). Random forests. *Machine learning*, **45**:1, 5–32.

Brown, L.D. (1986) Fundamentals of statistical exponential families: with applications in statistical decision theory. Institute of Mathematical Statistics, Hayworth, CA, USA.

Bühlmann, P. & van de Geer, S. (2011). Statistics for High Dimensional Data: Methods, Theory and Applications. Springer Verlag.

Clarke, B.S., Fokoué, E. & Zhang, H.H. (2009). Principles and Theory for Data Mining and Machine Learning. Springer Verlag.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20** 273–297.

Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function 1989 *Mathematics of Control, Signals, and Systems*, **2**, 303–314.

Darmois, G. (1935). Sur les lois de probabilites a estimation exhaustive. *Comptes Rendus de l'Académie des Sciencs, Paris*, **200** 1265–1266.

Davison, A.C. (1997). Bootstrap. Cambridge University Press.

Davidson, R. & MacKinnon, J.G. (1993). Estimation and Inference in Econometrics. Oxford University Press.

Davidson, R. & MacKinnon, J.G. (2003). Econometric Theory and Methods. Oxford University Press.

Duo, Q. (1993). The Formation of Econometrics. Oxford University Press.

Debreu, G. 1986. Theoretic Models: Mathematical Form and Economic Content. *Econometrica*, **54**, 1259–1270.

Efron, B. & Tibshirani, R. (1993). Bootstrap. Chapman Hall CRC.

Engel, E. (1857). Die Productions- und Consumtionsverhältnisse des Königreichs Sachsen. *Statistisches Bureau des Königlich Sächsischen Ministeriums des Innern.*

Feldstein, M. & Horioka, C. (1980). Domestic Saving and International Capital Flows. *Economic Journal*, **90**, 314–329.

Flach, P. (2012). Machine Learning. Cambridge University Press.

Frisch, R. & Waugh, F.V. (1933). Partial Time Regressions as Compared with Individual Trends. *Econometrica*. **1**, 387–401.

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, **106**, 746–762.

Givord, P. (2010). Méthodes économétriques pour l'évaluation de politiques publiques. INSEE Document de Travail, **08**

Grandvalet, Y., Mariéthoz, J., & Bengio, S. 2005. Interpretation of SVMs with an application to unbalanced classification. *Advances in Neural Information Processing Systems* **18**.

Haavelmo, T. (1944). The probability approach in econometrics, *Econometrica*, **12**:iii-vi and 1–115.

Hastie, T. & Tibshirani, R. (1990). Generalized Additive Models. Chapman & Hall/CRC.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning. Springer Verlag.

Hastie, T., Tibshirani, W. & Wainwright, M. (2015). Statistical Learning with Sparsity. Chapman CRC.

d'Haultefœuille, X. & Givord, P. (2014) La régression quantile en pratique. *Économie & Statistiques*, **471**, 85–111.

Hebb, D.O. (1949). The organization of behavior, New York, Wiley.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

Heckman, J.J., Tobias, J.L. & Vytlacil, E. (2003). Simple Estimators for Treatment Parameters in a Latent-Variable Framework. *The Review of Economics and Statistics*, **85**, 748–755.

Herbrich, R., Keilbach, M., Graepel, T. Bollmann-Sdorra, P. & Obermayer, K. (1999). Neural Networks in Economics. *in* Computational Techniques for Modelling in Economics, T. Brenner Eds. Springer Verlag, 169–196.

Hoerl, A.E. (1962). Applications of ridge analysis to regression problems. *Chemical Engineering Progress*, **58**:3, 54–59.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945–960.

Hyndman, R. , Koehler, A.B., Ord, J.K. & Snyder, R.D. (2009). Forecasting with Exponential Smoothing. Springer Verlag.

James, G., D. Witten, T. Hastie, & R. Tibshirani (2013). An introduction to Statistical Learning. Springer Series in Statistics.

Khashman, A. (2011). Credit risk evaluation using neural networks: Emotional versus conven- tional models. *Applied Soft Computing*, **11**, 5477–5484.

Keen, M.P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, **156**, 3–20.

Koch, I. (2013). Analysis of Multivariate and High-Dimensional Data. Cambridge University Press.

Koenker, R. (1998). Galton, Edgeworth, Frish, and prospects for quantile regression in Econometrics. Conference on Principles of Econometrics, Madison.

Koenker, R. (2003). Quantile Regression. Cambridge University Press.

Kolda, T. G. & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review* **51**, 455–500.

Koopmans, T.C. (1957). Three Essays on the State of Economic Science. McGraw-Hill.

Kuhn, M. & Johnson, K. (2013). Applied Predictive Modeling. Springer Verlag.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature* **521** 436–444.

Lemieux, T. (2006). The "Mincer Equation" Thirty Years After Schooling, Experience, and Earnings. *in* Jacob Mincer A Pioneer of Modern Labor Economics, Grossbard Eds, 127–145, Springer Verlag.

Li, J. & J. S. Racine (2006). Nonparametric Econometrics. Princeton University Press.

Lin, H.W., Tegmark, M. & Rolnick, D. (2016). Why does deep and cheap learning work so well? *ArXiv e-prints*.

Lucas, R.E. (1976). Econometric Policy Evaluation: A Critique. *Carnegie-Rochester Conference Series on Public Policy*, 19–46.

Mallows, C.L. (1973). Some Comments on C_p. *Technometrics*, **15**, 661–675.

Mincer, J. (1974). Schooling, experience and earnings. Columbia University Press.

Mitchell, T. (1997). Machine Learning. McGraw-Hill.

Morgan, J.N. & Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistcal Association*, **58**, 415–434.

Morgan, M.S. (1990). The history of econometric ideas. Cambridge University Press.

Mohri, M., Rostamizadeh, A. & Talwalker, A. (2012) Foundations of Machine Learning. MIT Press.

Mullainathan, S. & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, **31** 87–106.

Müller, M. (2011). Generalized Linear Models *in* Handbook of Computational Statistics, J.E Gentle, W.K. Härdle & Y. Mori Eds. Springer Verlag.

Murphy, K.R. (2012). Machine Learning: a Probabilistic Perspective. MIT Press.

Murphy, K. M. & F. Welch (1990). Empirical age-earnings profiles. *Journal of Labor Economics* **8**, 202–229.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and its Applications*, **9**:1, 141–2.

Nevo, A. & Whinston, M.D. (2010). Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference. *Journal of Economic Perspective*, **24**, 69–82.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux expériences agricoles : Essai des principes. Mémoire de master, republibé dans *Statistical Science*, **5**, 463–472.

Nisbet, R., Elder, J. & Miner, G. (2011). Handbook of Statistical Analysis and Data Mining Applications. Academic Press, New York.

Okun, A. (1962). Potential GNP: Its measurement and significance. *Proceedings of the Business and Economics Section of the American Statistical Association*, 98–103.

Orcutt, G.H. (1952). Toward a partial redirection of econometrics. *Review of Economics and Statistics*, **34** 195–213.

Pagan, A. & A. Ullah (1999). Nonparametric Econometrics. Themes in Modern Econometrics. Cambridge: Cambridge University Press.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**, 559-–572.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*. **10**, 61–74.

Portnoy, S. (1988). Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity. *Annals of Statistics*, **16**:356–366.

Quenouille, M. H. (1949). Problems in Plane Sampling. *The Annals of Mathematical Statistics* **20**(3):355–375.

Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika* **43**(3-4), 353–360.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning* **1** 81–106.

Reiersøol, O. (1945). Confluence analysis of means of instrumental sets of variables. *Arkiv. for Mathematik, Astronomi Och Fysik*, **32**.

Rosenbaum, P. & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41–55.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408.

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**, 688–701.

Ruppert, D., Wand, M. P. & Carroll, R.J. (2003). Semiparametric Regression. Cambridge University Press.

Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, **44**:1.

Schultz, H. (1930). The Meaning of Statistical Demand Curves. University of Chicago.

Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association* **88**:(422), 486–494.

Shao, J. (1997). An Asymptotic Theory for Linear Model Selection. *Statistica Sinica*, **7**, 221–264.

Shapire, R.E. & Freund, Y. (2012). Boosting. MIT Press.

Simonoff, J. S. (1996). Smoothing Methods in Statistics. Springer.

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B* , **39**:1, 44–47.

Tam, K.Y. & Kiang, M.Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, **38**, 926–947.

Tan, H. (1995). Neural-Network model for stock forecasting. MSc Thesis, Texas Tech. University.

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

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics*, **4**: 1035–1038.

Tinbergen, J. (1939). Statistical Testing of Business Cycle Theories. Vol. 1: A Method and its Application to Investment activity; Vol. 2: Business Cycles in the United States of America, 1919—1932. Geneva: League of Nations.

Tobin, J. (1958). Estimation of Relationship for Limited Dependent Variables. *Econometrica*, **26**, 24–36.

Tufféry, S. (2001). Data Mining and Statistics for Decision Making. Wiley Interscience.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, **29**:614–623.

Vapnik, V. (1998). Statistical Learning Theory. Wiley.

Vapnik, C, & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **16**:264–280.

Varian, H.R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, **28**(2):3–28.

Waltrup, L.S., Sobotka, F., Kneib, T. & Kauermann, G. (2014). Expectile and quantile regression—David and Goliath? *Statistical Modelling*, **15**, 433 – 456.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, **26**:4, 359–372.

Widrow, B. & Hoff, M.E. Jr. (1960). Adaptive Switching Circuits. IRE WESCON Convention Record, **4**:96–104.

Working, E. J. (1927). What do statistical 'demand curves' show? *Quarterly Journal of Economics*, **41**:212–35.