



HAL
open science

Économétrie & Machine Learning

Arthur Charpentier, Emmanuel Flachaire, Antoine Ly

► **To cite this version:**

Arthur Charpentier, Emmanuel Flachaire, Antoine Ly. Économétrie & Machine Learning. 2018.
hal-01568851v3

HAL Id: hal-01568851

<https://hal.science/hal-01568851v3>

Preprint submitted on 25 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Économétrie & *Machine Learning*

Arthur Charpentier

Université de Rennes 1 & CREM
7 Place Hoche, 35065 Rennes Cedex, France
arthur.charpentier@univ-rennes1.fr

Emmanuel Flachaire

Aix-Marseille Université, AMSE, CNRS & EHESS
5 bd Maurice Bourdet, CS 50498, 13205 Marseille Cedex 01, France
emmanuel.flachaire@univ-amu.fr

et

Antoine Ly

Université Paris-Est
5, boulevard Descartes, 77454 Marne-la-Vallée cedex, France
antoine.ly.pro@gmail.com

- Mai 2018 -

Résumé

L'économétrie et l'apprentissage automatique* 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 phénomènes 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'appropriier 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;

Arthur Charpentier est financé par la chaire *actinfo* de l'Institut Louis Bachelier, Paris.
Les auteurs remercient les participants au cours de l'école doctorale *edge* de l'Université Bretagne-Loire, et deux rapporteurs anonymes pour leurs commentaires et suggestions.

*Dans tout l'article le terme populaire *machine learning* sera traduit sous la forme « *apprentissage automatique* ».

Table des Matières

| | | |
|----------|---|-----------|
| 1 | Introduction | 2 |
| 1.1 | La Modélisation économétrique | 2 |
| 1.2 | Applications | 2 |
| 1.3 | De la grande dimension aux données massives | 4 |
| 1.4 | Statistique computationnelle et non-paramétrique | 4 |
| 1.5 | Plan de l'article | 5 |
| 2 | Économétrie et modèle probabiliste | 6 |
| 2.1 | Fondements de la statistique mathématique | 6 |
| 2.2 | Lois conditionnelles et vraisemblance | 7 |
| 2.3 | Les résidus | 8 |
| 2.4 | Géométrie du modèle linéaire Gaussien | 8 |
| 2.5 | Du paramétrique au non-paramétrique | 9 |
| 2.6 | Famille exponentielle et modèles linéaires | 10 |
| 2.7 | Régression logistique | 12 |
| 2.8 | Régression en grande dimension | 12 |
| 2.9 | Qualité d'un ajustement et choix de modèle | 13 |
| 2.10 | économétrie et tests statistiques | 14 |
| 2.11 | Quitter la corrélation pour quantifier un effet causal | 15 |
| 3 | Philosophie des méthodes d'apprentissage automatique | 15 |
| 3.1 | Apprentissage par une machine | 15 |
| 3.2 | Le tournant des années 80/90 et le formalisme probabiliste | 16 |
| 3.3 | Le choix de l'objectif et la fonction de perte | 18 |
| 3.4 | Boosting et apprentissage séquentiel | 20 |
| 3.5 | Pénalisation et choix de variables | 21 |
| 3.6 | Optimisation et aspects algorithmiques | 24 |
| 3.7 | <i>In-sample</i> , <i>out-of-sample</i> et validation croisée | 25 |
| 4 | Quelques outils d'apprentissage automatique | 27 |
| 4.1 | Réseaux de Neurones | 27 |
| 4.2 | Support Vecteurs Machine | 31 |
| 4.3 | Arbres, Bagging et Forêts Aléatoires | 32 |
| 4.4 | Sélection de modèle de classification | 33 |
| 4.5 | De la classification à la régression | 34 |
| 5 | Applications | 36 |
| 5.1 | Les ventes de sièges auto pour enfants (classification) | 36 |
| 5.2 | L'achat d'une assurance caravane (classification) | 37 |
| 5.3 | Les défauts de remboursement de crédits particuliers (classification) | 39 |
| 5.4 | Les déterminants des salaires (régression) | 39 |
| 5.5 | Les déterminants des prix des logements à Boston (régression) | 41 |
| 6 | Conclusion | 43 |

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 d'apprentissage automatique 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 des moindres carrés, 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 $\{(y_i, \mathbf{x}_i)\}$, avec $i = 1, \dots, n$, $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^p$ et $y_i \in \mathcal{Y}$. Si y_i est qualitative, on parlera d'un problème de classification², 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 θ , dans un modèle paramétrique $m_\theta(\cdot)$, par une valeur $\hat{\theta}$. 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 $\hat{\theta}$ (vu comme une variable aléatoire, grâce à la représentation stochastique sous-jacente), l'apprentissage statistique s'intéresse davantage aux propriétés du modèle optimal $m^*(\cdot)$ (suivant un critère qui reste à définir), voire simplement $m^*(\mathbf{x}_i)$ 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 mediocrity 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

²Nous utiliserons ici le terme « classification » lorsque \mathcal{Y} est un ensemble de classes, typiquement une classification binaire, $\mathcal{Y} = \{0, 1\}$, ce cas correspondant à la réalisation d'une variable indicatrice, $\mathbf{1}_{Y_t \leq 0}$, ou $\mathbf{1}_{Y \in \mathcal{A}}$, par exemple. Ce terme est moins daté que « discrimination » par exemple, et plus général que la constitution d'un « score » (qui est souvent une étape intermédiaire). Il ne doit pas être confondu avec la classification non-supervisée (comme la « classification ascendante hiérarchique ») qui est la constitution de classe homogène à partir d'une mesure de similarité (on utilisera parfois, dans ce cas, le terme de « constitution de classes », ou de « clusters »).

« modèle » économique. Comme le note Debreu (1986), la première étape a été de formuler des affirmations économiques dans un langage mathématique. Les différentes grandeurs sont vues comme des variables, et dans les années 1930, on verra apparaître les « *statistical demand curves* », pour reprendre la terminologie d’Henry Schultz. Cette approche statistique permettra d’aller plus loin que les travaux pionniers 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 macro-économiques. 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 la variation du nombre de demandeurs d’emploi et de la croissance du PIB,

$$\Delta \text{Ch\^omage}_t = \beta_0 + \beta_1 \text{Croissance}_t + \varepsilon_t,$$

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 modélise 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 macro-économiques posaient des problèmes insolubles d’identification. La solution passait forcément par de l’économétrie sur données individuelles (au lieu de données obtenues par agrégation), 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 collectées 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 contentés 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’explique Angrist & Krueger (1991), en affirmant « *research in a structuralist style relies heavily on economic theory to guide empirical work [...] An alternative to structural modeling, [...] the ‘experimentalist’ approach, [...] puts front and center the problem of identifying causal effects from specific events or situations* ».

On peut aussi souligner que si l’approche de la Commission Cowles était très exigeante, en supposant le modèle connu, toute une littérature s’est développée en allégeant cette hypothèse, soit en essayant de choisir le bon modèle (avec les travaux de Hendry & Krolzig (1995) par exemple) ou en proposant de faire des moyennes de modèles (comme développé récemment par Li *et al.* (2017)). Et plus généralement, 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 artificielle) 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 « sparse » ou « parcimonieux »).

1.3 De la grande dimension aux 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 \mathbf{X} , de taille $n \times p$, avec n et p potentiellement grands³. 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 $n \rightarrow \infty$ (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 $p^2/n \rightarrow 0$ lorsque $n, p \rightarrow \infty$. Aussi, il n'est pas rare de parler de grande dimension dès lors que $p > \sqrt{n}$. Un autre concept important est celui de sparsité, 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 $p \rightarrow \infty$. On dit alors que l'espace est « parcimonieux » - 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.

Pour reprendre la terminologie de Bühlmann & van de Geer (2011), les problèmes que nous évoquons ici correspondent à ceux observés en grande dimension, qui est un problème essentiellement statistique. D'un point de vue informatique, on peut aller un peu plus loin, avec des données réellement massives (qui occupent énormément de place en mémoire). Dans ce qui précède, les données étaient stockées dans une matrice \mathbf{X} , de taille $n \times p$. Si cet objet formel est toujours bien défini, il peut y avoir des soucis à stocker une telle matrice, voire manipuler une matrice abondamment utilisée en économétrie, $\mathbf{X}^T \mathbf{X}$ (matrice $n \times n$). La condition du premier ordre (dans le modèle linéaire) est associée à la résolution de $\mathbf{X}^T (\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) = \mathbf{0}$. En dimension raisonnable, on utilise la décomposition QR (c'est à dire la décomposition de Gram-Schmidt). En grande dimension, on peut utiliser des méthodes numériques de descente de gradient, où le gradient est approché sur un sous-échantillon de données (comme décrit par exemple dans Zinkevich *et al.* (2010)) Cet aspect informatique est souvent oublié alors qu'il a été à la base de bon nombre d'avancées méthodologiques, en économétrie. Par exemple, Hoerl & Kennard (1980) reviennent sur l'origine de l'utilisation de la régression Ridge: « *Nous facturions 90\$ par jour pour notre temps, mais avons dû facturer 450\$ par heure d'ordinateur sur un Univac I (...). Avec cette machine, il a fallu 75 minutes de traitement pour inverser une matrice 40 × 40 en passant par une partition 4 × 4 de sous-matrices 10 × 10, en utilisant des bandes magnétiques pour le stockage temporaire. Nous avons noté que les coefficients d'un régression linéaire calculés en utilisant les moindres carrés n'avaient pas toujours de sens. Les coefficients avaient tendance à être trop grands en valeur absolue, certains avaient même le mauvais signe, et ils pouvaient être instables avec de très petits changements dans les données (...). Comme la méthode que nous proposons attaquait l'une des vaches sacrées de la régression linéaire - les moindres carrés - nous avons fait face à une résistance considérable.* ».

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 statis-

³Là 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.

tique 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, une liste exhaustive est présentée dans Shalev-Shwartz & Ben-David (2014)). Mais la frontière entre les deux est très poreuse. à l’intersection se retrouve, par exemple, 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.

L’économétrie non-paramétrique repose sur des décompositions dans des bases fonctionnelles. L’économétrie linéaire consiste à approcher la fonction $m : \mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ 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 :

$$m(\mathbf{x}) = \sum_{j=0}^{\infty} \omega_j g_j(\mathbf{x}) \quad \text{et} \quad \hat{m}(\mathbf{x}) = \sum_{j=0}^{h^*} \hat{\omega}_j g_j(\mathbf{x}),$$

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 \mathbf{x} , obtenu en considérant seulement les observations proches de \mathbf{x} :

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

où $n_{\mathbf{x}}$ est le nombre d’observations au voisinage de \mathbf{x} . En mettant des poids fonctions de la distance à \mathbf{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 à l’aide de théorèmes asymptotiques, soit à l’aide de techniques de validation, comme en apprentissage automatique. 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. Si les économistes ont toujours la culture du tableau présentant la « sortie de régression », les méthodes non-paramétriques sont utiles pour détecter des mauvaises spécifications, des non-prises en compte de nonlinéarité, ou d’effets croisés (et les outils d’apprentissage automatique que nous allons voir peuvent probablement jouer le même rôle).

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 machine, 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 « homoscedastique ». 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 Poissonien, par exemple, donnant naissance aux modèles linéaires généralisés (construits pour des variables dans la famille exponentielle).

Une fois rappelée 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 croisée. 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 d'apprentissage automatique, 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 apprentissage automatique peuvent être utilisés pour mieux détecter la mauvaise spécification des modèles de régression paramétriques, à cause de non-linéarités, et/ou d'interactions manquées.

2 Économétrie et modèle probabiliste

L'importance des modèles probabilistes en économie trouve sa source dans les questionnements de Working (1927) et les tentatives de réponses apportées dans les deux tomes de Tinbergen (1939). Ces derniers ont engendré par la suite énormément de travaux, 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 probabilités 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, \mathbf{x}_i) sont vues comme des réalisations de variables aléatoires (Y_i, \mathbf{X}_i) . En pratique, la loi jointe du couple (Y, \mathbf{X}) nous intéresse toutefois peu : la loi de \mathbf{X} est inconnue, et c'est la loi de Y conditionnelle à \mathbf{X} qui nous intéressera. Dans la suite, nous noterons x une observation, \mathbf{x} un vecteur d'observations, X une variable aléatoire, et \mathbf{X} un vecteur aléatoire et, abusivement, \mathbf{X} pourra aussi désigner la matrice des observations individuelles (les \mathbf{x}_i), suivant le contexte.

2.1 Fondements de la statistique mathématique

Comme le rappelle l'introduction de Vapnik (1998), l'inférence en statistique paramétrique est basée sur la croyance suivante: le statisticien connaît bien le problème à analyser, en particulier, il connaît la loi physique qui génère les propriétés stochastiques des données, et la fonction à trouver s'écrit via un nombre fini de paramètres⁴. Pour trouver ces paramètres, on adopte la méthode du maximum de vraisemblance. Le but de la théorie est de justifier cette approche (en découvrant et en décrivant ses propriétés favorables). On verra qu'en apprentissage, la philosophie est très différente, puisqu'on ne dispose pas d'informations a priori fiables sur la loi statistique sous-jacente au problème, ni-même sur la fonction que l'on voudrait approcher (on va alors proposer des méthodes pour construire une approximation à partir de données à notre disposition, pour reprendre Vapnik (1998)). Un « âge d'or » de l'inférence paramétrique, de 1930 à 1960, a posé les bases de la statistique mathématique, que l'on retrouve dans tous les manuels de statistique, y compris aujourd'hui. Comme le dit Vapnik (1998), le paradigme paramétrique classique est basé sur les trois croyances suivantes:

1. Pour trouver une relation fonctionnelle à partir des données, le statisticien est capable de définir un ensemble de fonctions, linéaires dans leurs paramètres, qui contiennent une bonne approximation de la fonction souhaitée. Le nombre de paramètres décrivant cet ensemble est petit.
2. La loi statistique sous-jacente à la composante stochastique de la plupart des problèmes de la vie réelle est la loi normale. Cette croyance a été soutenue en se référant au théorème de limite centrale, qui stipule que dans de larges conditions la somme d'un grand nombre de variables aléatoires est approximée par la loi normale.

⁴On peut rapprocher cette approche de l'économétrie structurelle, telle que présentée par exemple dans Kean (2010).

3. La méthode du maximum de vraisemblance est un bon outil pour estimer les paramètres.

Nous reviendrons dans cette partie sur la construction du paradigme économétrique, directement inspiré de celui de la statistique inférentielle classique.

2.2 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, \mathbf{X}_i) indépendantes (s'il est possible d'imaginer des observations temporelles - on aurait alors un processus (Y_t, \mathbf{X}_t) - mais nous n'aborderons pas les séries temporelles dans cet article). Plus précisément, on va supposer que conditionnellement aux variables explicatives \mathbf{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 \mathbf{x} . Dans le modèle linéaire Gaussien on suppose que :

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

On parle de modèle linéaire car $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$ est une combinaison linéaire des variables explicatives. C'est un modèle homoscédastique si $\text{Var}[Y|\mathbf{X} = \mathbf{x}] = \sigma^2$, 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 :

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

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 :

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \text{argmax}\{\log \mathcal{L}(\beta_0, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X})\}.$$

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 apprentissage automatique.

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

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} \quad (2)$$

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

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

En fait, l'hypothèse de normalité permet de faire un lien avec la statistique mathématique, mais il est possible de construire cet estimateur donné par l'équation (2). Si on suppose que $Y|\mathbf{X} = \mathbf{x} \stackrel{\mathcal{L}}{\sim} \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon$, avec $\mathbb{E}[\varepsilon] = 0$, $\text{Var}[\varepsilon] = \sigma^2$, $\text{Cov}[\varepsilon, X_j] = 0$ pour tout j , alors $\hat{\boldsymbol{\beta}}$ est un estimateur sans biais de $\boldsymbol{\beta}$ ($\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$) et de variance minimale parmi les estimateurs sans biais linéaires, avec $\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2 [\mathbf{X}^\top \mathbf{X}]^{-1}$. De plus, cet estimateur est asymptotiquement normal

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \quad \text{lorsque} \quad n \rightarrow \infty$$

La condition d'avoir une matrice \mathbf{X} de plein rang peut être (numériquement) forte en grande dimension. Si elle n'est pas vérifiée, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ n'existe pas. Si \mathbb{I} désigne la matrice identité, notons toutefois que $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I})^{-1} \mathbf{X}^\top \mathbf{y}$ 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)). Cette estimateur apparaît naturellement dans un contexte d'économétrie Bayésienne (nous le reverrons dans la section suivante présentant les techniques d'apprentissage automatique).

2.3 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 :

$$y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i \quad (3)$$

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

$$\hat{\varepsilon}_i = y_i - [\hat{\beta}_0 + \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}]$$

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 :

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

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

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

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

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

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 (3)) 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

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

(u_i désignant un terme d'erreur) 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),

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

(v_i désignant un autre terme d'erreur). Historiquement, le terme d'erreur dans l'équation (3) a pu être interprété 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, la 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).

2.4 Géométrie du modèle linéaire Gaussien

Définissons le produit scalaire dans \mathbb{R}^n , $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \mathbf{b}$, et notons $\|\cdot\|$ la norme euclidienne associée, $\|\mathbf{a}\| = \sqrt{\mathbf{a}^\top \mathbf{a}}$ (notée $\|\cdot\|_{\ell_2}$ dans la suite). Notons $\mathcal{E}_{\mathbf{X}}$ l'espace engendré par l'ensemble des combinaisons linéaires des composantes \mathbf{x} (en rajoutant la constante). Si les variables explicatives sont linéairement indépendantes, \mathbf{X} est de plein rang colonne et $\mathcal{E}_{\mathbf{X}}$ est un sous-espace de dimension $p + 1$ de \mathbb{R}^n . Supposons à partir de maintenant que les variables \mathbf{x} et la variable y sont ici centrées. Notons qu'aucune hypothèse de loi n'est faite dans cette section, les propriétés géométriques découlent des propriétés de l'espérance et de la variance dans l'espace des variables de variance finie.

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

$$\hat{\mathbf{y}} = \underbrace{\mathbf{X}[\mathbf{X}^\top \mathbf{X}]^{-1} \mathbf{X}^\top}_{\Pi_{\mathcal{X}}} \mathbf{y} = \Pi_{\mathcal{X}} \mathbf{y}. \quad (4)$$

Comme $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbb{I} - \Pi_{\mathcal{X}})\mathbf{y} = \Pi_{\mathcal{X}^\perp}\mathbf{y}$, on note que $\widehat{\boldsymbol{\varepsilon}} \perp \mathbf{x}$, 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 :

$$\|\mathbf{y}\|^2 = \|\Pi_{\mathcal{X}}\mathbf{y}\|^2 + \|\Pi_{\mathcal{X}^\perp}\mathbf{y}\|^2 = \|\Pi_{\mathcal{X}}\mathbf{y}\|^2 + \|\mathbf{y} - \Pi_{\mathcal{X}}\mathbf{y}\|^2 = \|\widehat{\mathbf{y}}\|^2 + \|\widehat{\boldsymbol{\varepsilon}}\|^2 \quad (5)$$

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 2.9) s'interprète alors comme le carré du cosinus de l'angle θ entre \mathbf{y} et $\Pi_{\mathcal{X}}\mathbf{y}$:

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

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

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

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

$$\widehat{\boldsymbol{\beta}}_2 = [\mathbf{X}_2^{*\top} \mathbf{X}_2^*]^{-1} \mathbf{X}_2^{*\top} \mathbf{y}_2^*$$

Autrement dit, l'estimation globale est équivalente à l'estimation indépendante des deux modèles si $\mathbf{X}_2^* = \mathbf{X}_2$, c'est à dire $\mathbf{X}_2 \in \mathcal{E}_{\mathbf{X}_1^\perp}$, que l'on peut noter $\mathbf{x}_1 \perp \mathbf{x}_2$. 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 l'espace des variables de variance finie).

Cette vision géométrique permet de mieux comprendre le problème de la sous-identification, c'est à dire le cas où le vrai modèle serait $y_i = \beta_0 + \mathbf{x}_1^\top \boldsymbol{\beta}_1 + \mathbf{x}_2^\top \boldsymbol{\beta}_2 + \varepsilon_i$, mais le modèle estimé est $y_i = \beta_0 + \mathbf{x}_1^\top \mathbf{b}_1 + \eta_i$. L'estimateur du maximum de vraisemblance de \mathbf{b}_1 est :

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

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

En revanche, la sur-identification correspond au cas où le vrai modèle serait $y_i = \beta_0 + \mathbf{x}_1^\top \boldsymbol{\beta}_1 + \varepsilon_i$, mais le modèle estimé est $y_i = \beta_0 + \mathbf{x}_1^\top \mathbf{b}_1 + \mathbf{x}_2^\top \mathbf{b}_2 + \eta_i$. Dans ce cas, l'estimation est sans biais, au sens où $\mathbb{E}(\widehat{\mathbf{b}}_1) = \boldsymbol{\beta}_1$ 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 $\widehat{\mathbf{b}}_2$ 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).

2.5 Du paramétrique au non-paramétrique

La réécriture de l'équation (4) sous la forme

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \underbrace{\mathbf{X}[\mathbf{X}^\top \mathbf{X}]^{-1} \mathbf{X}^\top}_{\Pi_{\mathcal{X}}} \mathbf{y}$$

permet de voir la prévision directement comme une transformation linéaire des observations. De manière plus générale, on peut obtenir un prédicteur linéaire en considérant $m(\mathbf{x}) = \mathbf{s}_x^\top \mathbf{y}$, où \mathbf{s}_x est un vecteur de poids, qui dépendent de \mathbf{x} , interprété comme un vecteur de lissage. En utilisant les vecteurs \mathbf{s}_{x_i} , calculés à partir des \mathbf{x}_i , on obtient une matrice \mathbf{S} de taille $n \times n$, et $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$. Dans le cas de la régression linéaire décrite auparavant, $\mathbf{s}_x = \mathbf{X}[\mathbf{X}^\top \mathbf{X}]^{-1} \mathbf{x}$, et classiquement, $\text{trace}(\mathbf{S})$ est le nombre de colonnes de la matrice \mathbf{X} (le nombre de variables explicatives). Dans ce contexte de prédicteurs linéaires, $\text{trace}(\mathbf{S})$ est souvent vu comme un équivalent au nombre de paramètres (ou complexité, ou dimension, du modèle), et $\nu = n - \text{trace}(\mathbf{S})$ est alors le nombre de degrés de liberté (comme défini dans Ruppert, Wand & Carroll (2003) et Simonoff (1996)). Le principe de parcimonie⁵ consiste à minimiser cette dimension (la trace de la matrice \mathbf{S}) autant que faire se peut. 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

$$\hat{m}_h(x) = \mathbf{s}_x^\top \mathbf{y} = \sum_{i=1}^n s_{x,i} y_i \quad \text{avec} \quad s_{x,i} = \frac{K_h(x - x_i)}{K_h(x - x_1) + \dots + K_h(x - x_n)},$$

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.

L'introduction de ce méta-paramètre h pose un soucis, car il convient de le choisir judicieusement. En faisant des développements limités, on peut montrer que si X a pour densité f ,

$$\text{biais}[\hat{m}_h(x)] = \mathbb{E}[\hat{m}_h(x)] - m(x) \sim h^2 \left(\frac{C_1}{2} m''(x) + C_2 m'(x) \frac{f'(x)}{f(x)} \right) \quad \text{et} \quad \text{Var}[\hat{m}_h(x)] \sim \frac{C_3}{nh} \frac{\sigma(x)}{f(x)}$$

pour des constantes que l'on peut estimer (voir Simonoff (1996) par exemple). Ces deux fonctions évoluent inversement en fonction de h , comme le rappelle la Figure 1 (où le méta-paramètre est ici h^{-1}). L'idée naturelle est alors de chercher à minimiser l'erreur quadratique moyenne, le MSE, $\text{biais}[\hat{m}_h(x)]^2 + \text{Var}[\hat{m}_h(x)]$, ce qui donne une valeur optimale pour h de la forme $h^* = O(n^{-1/5})$, ce qui rappelle la règle de Silverman (1986). En plus grande dimension, pour des variables \mathbf{x} continues, on peut utiliser un noyau multivarié, de fenêtre matricielle \mathbf{H} ,

$$\mathbb{E}[\hat{m}_{\mathbf{H}}(\mathbf{x})] \sim m(\mathbf{x}) + \frac{C_1}{2} \text{trace}(\mathbf{H}^\top m''(\mathbf{x}) \mathbf{H}) + C_2 \frac{m'(\mathbf{x})^\top \mathbf{H} \mathbf{H}^\top \nabla f(\mathbf{x})}{f(\mathbf{x})} \quad \text{et} \quad \text{Var}[\hat{m}_{\mathbf{H}}(\mathbf{x})] \sim \frac{C_3}{n} \frac{\sigma(\mathbf{x})}{\det(\mathbf{H}) f(\mathbf{x})}.$$

Si \mathbf{H} est une matrice diagonale, avec le même terme h sur la diagonale, alors $h^* = O(n^{-1/(4+\dim(\mathbf{x}))})$. Cela dit, en pratique, on sera davantage intéressé par la version intégrée de l'erreur quadratique,

$$MISE(\hat{m}_h) = \mathbb{E}[MSE(\hat{m}_h(X))] = \int MSE(\hat{m}_h(x)) dF(x),$$

dont on peut montrer que

$$MISE[\hat{m}_h] \sim \overbrace{\frac{h^4}{4} \left(\int x^2 k(x) dx \right)^2 \int [m''(x) + 2m'(x) \frac{f'(x)}{f(x)}]^2 dx}^{\text{biais}^2} + \overbrace{\frac{\sigma^2}{nh} \int k^2(x) dx \cdot \int \frac{dx}{f(x)}}^{\text{variance}},$$

lorsque $n \rightarrow \infty$ et $nh \rightarrow \infty$. On retrouve ici une relation asymptotique qui rappelle l'ordre de grandeur de Silverman (1986),

$$h^* = n^{-\frac{1}{5}} \left(\frac{C_1 \int \frac{dx}{f(x)}}{C_2 \int [m''(x) + 2m'(x) \frac{f'(x)}{f(x)}] dx} \right)^{\frac{1}{5}},$$

sauf que beaucoup de termes ici sont inconnus. On verra, l'apprentissage automatique propose des techniques computationnelles, lorsque l'économètre avait pris l'habitude de chercher des propriétés asymptotiques.

2.6 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

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

⁵ « *pluralitas non est ponenda sine necessitate* » pour reprendre le principe énoncé par Guillaume d'Occam (les multiples ne doivent pas être utilisés sans nécessité).

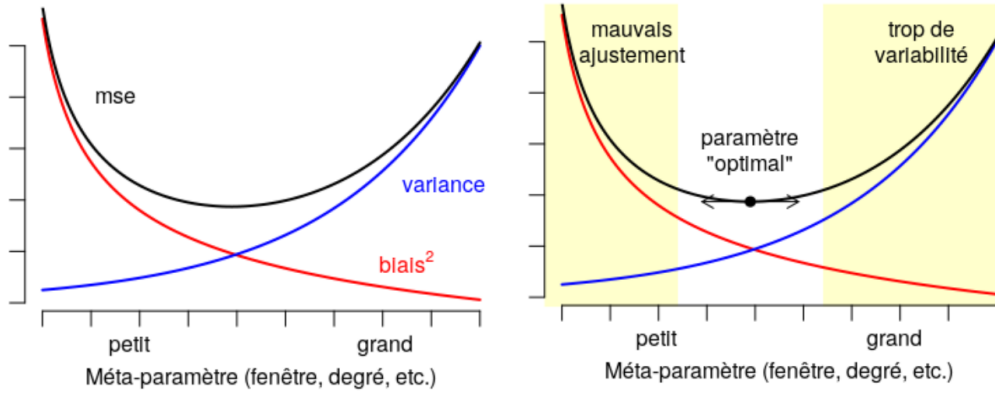


Figure 1: Choix du méta-paramètre et le problème de Boucle d'Or : il ne doit être ni trop grand (sinon il y a trop de variance), ni trop petit (sinon il y a trop de biais) .

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)), et ψ est une fonction bijective que se donne l'utilisateur. La log-vraisemblance a alors une expression relative simple

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

et la condition du premier ordre s'écrit alors

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

pour reprendre les notations de Müller (2011), où \mathbf{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 $\psi(\cdot)$, on aura plutôt tendance à spécifier la fonction de lien $g(\cdot)$ définie par

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

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 \mathbf{W} dépend de $\boldsymbol{\beta}$ (avec $\mathbf{W} = \text{diag}(\nabla g(\hat{\mathbf{y}}) \text{Var}[\mathbf{y}])$) il n'existe en général pas de formule explicite pour l'estimateur du maximum de vraisemblance. Mais un algorithme itératif permet d'obtenir une approximation numérique. En posant

$$\mathbf{z} = g(\hat{\mathbf{y}}) + (\mathbf{y} - \hat{\mathbf{y}}) \cdot \nabla g(\hat{\mathbf{y}})$$

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

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

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

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

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

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.

2.7 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 :

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

ou encore :

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

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

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

On continuera à parler des modèles linéaires car les courbes d'isoprobabilités sont ici les hyperplans parallèles $\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$. à ce modèle, popularisé par Berkson (1944), certains préféreront le modèle probit (comme le raconte Berkson (1951)), introduit par Bliss (1934). Dans ce modèle :

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

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

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

où les résidus sont Gaussiens, de loi $\mathcal{N}(0, \sigma^2)$. Une alternative est d'avoir des résidus centrés de variance unitaire, et de considérer une modélisation latente de la forme $y_i = \mathbf{1}(y_i^* > \xi)$ (où ξ sera à fixer). On le voit, ces techniques sont fondamentalement liées à un modèle stochastique sous-jacent. Mais dans la section 5, nous présenterons plusieurs techniques alternatives - tirées de la littérature en apprentissage - pour ce problème de classification (avec deux classes, ici 0 et 1).

2.8 Régression en grande dimension

Comme nous l'avons mentionné auparavant, la condition du premier ordre $\mathbf{X}^\top (\mathbf{X} \hat{\boldsymbol{\beta}} - \mathbf{y}) = \mathbf{0}$ se résout numériquement en effectuant une décomposition QR, pour un coût en $O(np^2)$ opérations (où p est le rang de $\mathbf{X}^\top \mathbf{X}$). Numériquement, ce calcul peut être long (soit parce que p est grand, soit parce que n est grand), et une stratégie plus simple peut être de faire du sous-échantillonnage. Soit $n_s \ll n$, et considérons un sous-échantillon de taille n_s de $\{1, \dots, n\}$. Alors $\hat{\boldsymbol{\beta}}_s = (\mathbf{X}_s^\top \mathbf{X}_s)^{-1} \mathbf{X}_s^\top \mathbf{y}_s$ est une bonne approximation de $\hat{\boldsymbol{\beta}}$ comme le montre Dhillon *et al.* (2014). Cet algorithme est toutefois dangereux si certains points ont un pouvoir de levier important (i.e. $L_i = \mathbf{x}_i (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i^\top$). Tropp (2011) propose de transformer les données (de manière linéaire), mais une approche plus populaire est de faire du sous-échantillonnage non uniforme, avec une probabilité liée à l'influence des observations (définie par $I_i = \hat{\varepsilon}_i L_i / (1 - L_i)^2$, et qui malheureusement ne peut être calculée qu'une fois le modèle estimé).

De manière générale, on parlera de données massives lorsque la table de données de taille $n \times p$ ne tient pas en mémoire RAM de l'ordinateur. Cette situation est souvent rencontrée en apprentissage statistique de nos jours avec très souvent $p \ll n$. C'est la raison pour laquelle, en pratique de nombreuses bibliothèques d'algorithmes assimilées à de l'apprentissage machine⁶ utilisent des méthodes itératives pour résoudre la condition du premier ordre. Lorsque le modèle paramétrique à calibrer est effectivement convexe et semi-différentiable, il est possible d'utiliser par exemple la méthode de descente de gradient stochastique comme le suggère Bottou (2010). Ce dernier permet de s'affranchir à chaque itération du calcul du gradient sur chaque observation de notre base d'apprentissage. Plutôt que d'effectuer une descente moyenne à chaque itération, on commence par tirer (sans remise) une observation X_i parmi les n disponibles. On corrige ensuite les paramètres du modèle de sorte à ce que la prédiction faite à partir de X_i soit la plus proche possible de la vraie valeur y_i . On réitère ensuite la méthode jusqu'à avoir parcourue l'ensemble des données. Dans cet

⁶comme, par exemple, celles du langage Python.

algorithme il y a donc autant d'itération que d'observations. Contrairement à l'algorithme de descente de gradient (ou méthode de Newton) à chaque itération un seul vecteur de gradient est calculé (et non plus n). Il est néanmoins parfois nécessaire d'exécuter cette algorithme plusieurs fois pour augmenter la convergence des paramètres du modèle.

Si l'objectif est par exemple de minimiser l'erreur quadratique ℓ entre l'estimateur $f_\beta(X)$ et y l'algorithme peut se résumer ainsi :

Étape 0: Mélange des données

Étape d'itérations: Pour $t = 1, \dots, n$, on tire $i \in \{1, \dots, n\}$ sans remise, on pose

$$\beta^{t+1} = \beta^t - \gamma_t \frac{\partial \ell(y_i, f_{\beta^t}(X_i))}{\partial \beta}$$

Cet algorithme peut être réitéré plusieurs fois dans son ensemble selon le besoin de l'utilisateur. L'avantage de cette méthode est qu'à chaque itération, il n'est pas nécessaire de calculer le gradient sur toutes les observations (plus de somme). Elle est donc adaptée aux bases de données volumineuses. Cet algorithme s'appuie sur une convergence en probabilité vers un voisinage de l'optimum (et non pas l'optimum lui même).

2.9 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

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

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

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

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

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

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

$$R^2 = \frac{\text{Deviance}(\bar{y}) - \text{Deviance}(\hat{y})}{\text{Deviance}(\bar{y})} = 1 - \frac{\text{Deviance}(\hat{y})}{\text{Deviance}(\bar{y})}.$$

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 2.5). 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 estimateur dans la classe des prédicteurs linéaires,

$$\mathcal{M} = \{m : m(\mathbf{x}) = s_h(\mathbf{x})^\top \mathbf{y} \text{ où } S = (s(\mathbf{x}_1), \dots, s(\mathbf{x}_n))^\top \text{ est la matrice de lissage}\}$$

et supposons que $\mathbf{y} = m_0(\mathbf{x}) + \varepsilon$, avec $\mathbb{E}[\varepsilon] = \mathbf{0}$ et $\text{Var}[\varepsilon] = \sigma^2 \mathbb{I}$, de telle sorte que $m_0(\mathbf{x}) = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}]$. D'un point de vue théorique, le risque quadratique, associé à un modèle estimé \hat{m} , s'écrit

$$\mathcal{R}(\hat{m}) = \mathbb{E}[(Y - \hat{m}(\mathbf{X}))^2] = \underbrace{\mathbb{E}[(Y - m_0(\mathbf{X}))^2]}_{\text{erreur}} + \underbrace{\mathbb{E}[(m_0(\mathbf{X}) - \mathbb{E}[\hat{m}(\mathbf{X})])^2]}_{\text{biais}} + \underbrace{\mathbb{E}[(\mathbb{E}[\hat{m}(\mathbf{X})] - \hat{m}(\mathbf{X}))^2]}_{\text{variance}}$$

si m_0 désigne le vrai modèle. Le premier terme est parfois appelé « erreur de Bayes », et ne dépend pas de l'estimateur retenu, \hat{m} .

Le risque empirique quadratique, associé à un modèle m , est ici :

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

(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 :

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

On peut montrer que :

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

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

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

Aussi, si $\text{trace}(\mathbf{S}) \geq 0$, 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 Mallows, introduit dans Mallows (1973) utilisant non pas la déviance mais le R^2 .

2.10 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 $t_k = \hat{\beta}_k / \text{se}_{\hat{\beta}_k}$, 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 $T \stackrel{\mathcal{L}}{\sim} \text{Std}_\nu$, 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 récurrent dans les tests multiples, il est naturel d'utiliser la procédure de Benjamini & Hochberg (1995).

2.11 Quitter la corrélation pour quantifier un effet causal

Les modèles économétriques sont utilisés pour mettre en oeuvre des politiques publiques. Il est alors fondamental de bien comprendre les mécanismes sous-jacents pour savoir quelles variables permettent effectivement d'agir sur une variable d'intérêt. Mais on passe alors dans une autre dimension importante de l'économétrie. 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 contre-factuel, 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 \mathbf{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, $p(\mathbf{x}) = \mathbb{P}[T = 1|\mathbf{X} = \mathbf{x}]$, en notant que si la variable Y_0 est indépendante de l'accès au traitement T conditionnellement aux variables explicatives \mathbf{X} , alors elle est indépendante de T conditionnellement au score $p(\mathbf{X})$: 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 d'apprentissage automatique

Parallèlement à ces outils développés par et pour des économistes, toute une littérature a été développée sur des questions similaires, centrées autour de la prévision. Pour Breiman (2001a), une première différence vient du fait que la statistique s'est développée autour du principe d'inférence (ou d'explicitation de la relation liant y aux variables \mathbf{x}) alors qu'une autre culture s'intéresse avant tout à la prédiction. Dans une discussion qui suit l'article, David Cox l'affirme très clairement « *predictive success (...) is not the primary basis for model choice* ». Nous allons présenter les fondements des techniques d'apprentissage automatique (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 par une machine

Aujourd'hui, on parle d'« apprentissage automatique » pour décrire tout un ensemble de techniques, souvent computationnelles, alternatives à l'approche décrite auparavant (correspondant à l'économétrie classique). Avant de les caractériser autant que possible, notons juste qu'historiquement d'autres noms ont pu être donnés. Par exemple, Friedman (1997) propose de faire le lien entre la statistique (qui ressemble beaucoup aux techniques économétriques - test d'hypothèses, ANOVA, régression linéaire, logistique, GLM, etc) et ce qu'il appelait alors « *data mining* » (qui englobait alors les arbres de décisions, les méthodes des plus proches voisins, les réseaux de neurones, etc.). Le pont qu'il contribuera à construire correspond aux techniques d'apprentissages statistiques, décrites dans Hastie *et al.* (2009), mais l'apprentissage machine est un très vaste champ de recherche.

L'apprentissage dit « naturel » (par opposition à celui d'une machine) est celui des enfants, qui apprennent à parler, à lire, à jouer. Apprendre à parler signifie segmenter et catégoriser des sons, et les associer à des significations. Un enfant apprend aussi simultanément la structure de sa langue maternelle et acquiert un ensemble de mots décrivant le monde qui l'entoure. Plusieurs techniques sont possibles, allant d'un apprentissage par coeur, par généralisation, par découverte, apprentissage plus ou moins supervisé ou autonome, etc. L'idée en intelligence artificielle est de s'inspirer du fonctionnement du cerveau pour apprendre, pour

permettre un apprentissage « artificiel » ou « automatique », par une machine. Une première application a été d'apprendre à une machine à jouer à un jeu (*tic-tac-toe*, échecs, go, etc). Une étape indispensable est d'expliquer l'objectif qu'il doit atteindre pour gagner. Une approche historique a été de lui apprendre les règles du jeu. Si cela permet de jouer, cela ne permettra pas à la machine de *bien* jouer. En supposant que la machine connaisse les règles du jeu, et qu'elle a le choix entre plusieurs dizaines de coups possible, lequel doit-elle choisir ? L'approche classique en intelligence artificielle utilise l'algorithme dit *min-max* utilisant une fonction d'évaluation : dans cet algorithme, la machine effectue une recherche en avant dans l'arbre des coups possibles, aussi loin que les ressources de calcul le lui permettent (une dizaine de coups aux échecs, par exemple). Ensuite, elle calcule différents critères (qui lui ont été indiqués au préalable) pour toutes les positions (nombre de pièces prises, ou perdues, occupation du centre, etc. dans notre exemple du jeu d'échec), et finalement, la machine joue le coup qui lui permet de maximiser son gain. Un autre exemple peut être celui de la classification et de la reconnaissance d'images ou de formes. Par exemple, la machine doit identifier un chiffre dans une écriture manuscrite (chèque, code postal). Il s'agit de prédire la valeur d'une variable y , en sachant qu'a priori $y \in \{0, 1, 2, \dots, 8, 9\}$. Une stratégie classique est de fournir à la machine des bases d'apprentissage, autrement dit ici des millions d'images labélisées (identifiées) de chiffres manuscrits. Une stratégie simple et naturelle est d'utiliser un critère de décision basé sur les plus proches voisins dont on connaît l'étiquette (à l'aide d'une métrique prédéfinie).

La méthode des plus proches voisins (« *k-nearest neighbors* ») peut être décrit de la manière suivante : on considère (comme dans la partie précédente) un ensemble de n observations, c'est à dire des paires (y_i, \mathbf{x}_i) avec $\mathbf{x}_i \in \mathbb{R}^p$. Considérons une distance Δ sur \mathbb{R}^p (la distance Euclidienne ou la distance de Mahalanobis, par exemple). étant donnée une nouvelle observation $\mathbf{x} \in \mathbb{R}^p$, supposons les observations ordonnées en fonction de la distance entre les \mathbf{x}_i et \mathbf{x} , au sens où

$$\Delta(\mathbf{x}_1, \mathbf{x}) \leq \Delta(\mathbf{x}_2, \mathbf{x}) \leq \dots \leq \Delta(\mathbf{x}_n, \mathbf{x})$$

alors on peut considérer comme prédiction pour y la moyenne des k plus proches voisins,

$$m_k(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^k y_i.$$

L'apprentissage fonctionne ici par induction, à partir d'un échantillon (appelé base d'apprentissage).

L'apprentissage automatique englobe ces algorithmes qui donnent aux ordinateurs la capacité d'apprendre sans être explicitement programmé (comme l'avait défini Arthur Samuel en 1959). La machine va alors explorer les données avec un objectif précis (comme chercher les plus proches voisins dans l'exemple que nous venons de décrire). Tom Mitchell a proposé une définition plus précise en 1998 : on dit qu'un programme d'ordinateur apprend de l'expérience E par rapport à une tâche T et une mesure de performance P , si sa performance sur T , mesurée par P , s'améliore avec l'expérience E . La tâche T peut être un score de défaut par exemple, et la performance P peut être le pourcentage d'erreurs commise. Le système apprend si le pourcentage de défauts prédit augmente avec l'expérience.

On le voit, l'apprentissage machine est fondamentalement un problème d'optimisation d'un critère à partir de données (dites d'apprentissage). Nombreux sont les ouvrages de programmation qui proposent des algorithmes, sans jamais faire mention d'un quelconque modèle probabiliste. Dans Watt *et al.* (2016) par exemple, il n'est fait mention du mot « probabilité » qu'une seule fois, avec cette note de bas de page qui surprendra et fera sourire les économètres, « *logistic regression can also be interpreted from a probabilistic perspective* » (page 86). Mais beaucoup d'ouvrages récents proposent une relecture des approches d'apprentissage machine à l'aide de théories probabilistes, suite aux travaux de Vaillant et Vapnik. En proposant le paradigme de l'apprentissage « *probablement à peu près correct* » (PAC), une saveur probabiliste a été rajouté à l'approche jusqu'alors très computationnelle, en quantifiant l'erreur de l'algorithme d'apprentissage (dans un problème de classification).

3.2 Le tournant des années 80/90 et le formalisme probabiliste

On dispose d'un échantillon d'apprentissage, avec des observations (\mathbf{x}_i, y_i) où les variables y sont dans un ensemble \mathcal{Y} . Dans le cas de la classification, $\mathcal{Y} = \{-1, +1\}$, mais on peut imaginer un ensemble relativement général. Un prédicteur est une fonction m à valeurs dans \mathcal{Y} , permettant d'étiqueter (ou de classer) les nouvelles observations à venir. On suppose que les étiquettes sont produites par un classifieur f appelé cible. Pour un statisticien, cette fonction serait le vrai modèle. Naturellement, on veut construire m le plus proche possible de f . Soit \mathbb{P} une distribution (inconnue) sur \mathcal{X} . L'erreur de m relativement à la cible f est définie par

$$\mathcal{R}_{\mathbb{P}, f}(m) = \mathbb{P}[m(\mathbf{X}) \neq f(\mathbf{X})] \text{ où } \mathbf{X} \sim \mathbb{P},$$

ou écrit de manière équivalente,

$$\mathcal{R}_{\mathbb{P},f}(m) = \mathbb{P}[\{\mathbf{x} \in \mathcal{X} : m(\mathbf{x}) \neq f(\mathbf{x})\}].$$

Pour trouver notre classifieur, il devient nécessaire de supposer qu'il existe un lien entre les données de notre échantillon et le couple (\mathbb{P}, f) , c'est à dire un modèle de génération des données. On va alors supposer que les \mathbf{x}_i sont obtenus par des tirages indépendants suivant \mathbb{P} , et qu'ensuite $y_i = f(\mathbf{x}_i)$.

On peut ici définir le risque empirique d'un modèle m ,

$$\widehat{\mathcal{R}}(m) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(m(\mathbf{x}_i) \neq y_i).$$

Il est alors important d'admettre qu'on ne peut pas trouver un modèle parfait, au sens où $\mathcal{R}_{\mathbb{P},f}(m) = 0$. En effet, si on considère le cas le plus simple qui soit, avec $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$ et que \mathbb{P} soit telle que $\mathbb{P}(\{\mathbf{x}_1\}) = p$ et $\mathbb{P}(\{\mathbf{x}_2\}) = 1 - p$. Il est aussi possible d'observer x_1 et x_2 , et malgré tout, de se tromper sur les étiquettes. Aussi, au lieu de chercher un modèle parfait, on peut tenter d'avoir un modèle approximativement correct. On va alors chercher à trouver m tel que $\mathcal{R}_{\mathbb{P},f}(m) \leq \epsilon$, où ϵ est un seuil spécifié a priori.

Une fois admis ce premier écueil, qui fait penser à l'erreur de modèle, notons aussi un second problème. Sur notre exemple à deux valeurs, la probabilité de ne jamais observer x_2 parmi n tirages suivant \mathbb{P} est p^n . Il sera alors impossible de trouver $m(x_2)$ car cette valeur n'aura jamais été observée. Autrement dit, aucun algorithme ne peut nous assurer d'avoir avec certitude, avec n observations, $\mathcal{R}_{\mathbb{P},f}(m) \leq \epsilon$. On va alors chercher à être probablement approximativement correct (PAC). Pour se faire, on autorise l'algorithme à se tromper avec une probabilité δ , là aussi fixée a priori.

Aussi, quand on construit un classifieur, on ne connaît ni \mathbb{P} , ni f , mais on se donne un critère de précision ϵ , et un paramètre de confiance δ , et on dispose de n observations. Notons que n , ϵ et δ peuvent être liés. On cherche alors un modèle m tel que $\mathcal{R}_{\mathbb{P},f}(m) \leq \epsilon$ avec probabilité (au moins) $1 - \delta$, de manière à être probablement approximativement correct.

Wolpert (1996) a montré (détaillé dans Wolpert & Macready (1997)) qu'il n'existe pas d'algorithme d'apprentissage universel. En particulier, on peut montrer qu'il existe \mathbb{P} telle que $\mathcal{R}_{\mathbb{P},f}(m)$ soit relativement grande, avec une probabilité (elle aussi) relativement grande.

L'interprétation qui en est faite est qu'il est nécessaire d'avoir un biais pour apprendre. Comme on ne peut pas apprendre (au sens PAC) sur l'ensemble des fonctions m , on va alors contraindre m à appartenir une classe particulière, notée \mathcal{M} . Supposons pour commencer que \mathcal{M} contienne un nombre fini de modèles possibles. On peut alors montrer que pour tout ϵ et δ , que pour tout \mathbb{P} et f , si on dispose d'assez d'observations (plus précisément $n \geq \epsilon^{-1} \log[\delta^{-1}|\mathcal{M}|]$), alors avec une probabilité plus grande que $1 - \delta$, $\mathcal{R}_{\mathbb{P},f}(m^*) \leq \epsilon$ où

$$m^* \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbf{1}(m(\mathbf{x}_i) \neq y_i) \right\}$$

autrement dit m^* est un modèle dans \mathcal{M} qui minimise le risque empirique.

Un peut aller un peu plus loin, en restant dans le cas où $\mathcal{Y} = \{-1, +1\}$. Une classe \mathcal{M} de classifieurs sera dite PAC-apprenable s'il existe $n_{\mathcal{M}} : [0, 1]^2 \rightarrow \mathbb{N}$ tel que, pour tout ϵ , δ , \mathbb{P} et si on suppose que la cible f appartient à \mathcal{M} , alors en utilisant $n > n_{\mathcal{M}}(\epsilon, \delta)$ tirages d'observations \mathbf{x}_i suivant \mathbb{P} , étiquetés y_i par f , alors il existe $m \in \mathcal{M}$ tel que, avec probabilité $1 - \delta$, $\mathcal{R}_{\mathbb{P},f}(m) \leq \epsilon$. La fonction $n_{\mathcal{M}}$ est alors appelée complexité d'échantillon pour apprendre. En particulier, nous avons vu que si \mathcal{M} contient un nombre fini de classifieurs, alors \mathcal{M} est PAC-apprenable avec la complexité $n_{\mathcal{M}}(\epsilon, \delta) = \epsilon^{-1} \log[\delta^{-1}|\mathcal{M}|]$.

Naturellement, on souhaiterait avoir un résultat plus général, en particulier si \mathcal{M} n'est pas fini. Pour cela, il faut utiliser la dimension VC de Vapnik-Chervonenkis, qui repose sur l'idée de pulvérisation de nuages de points (pour une classification binaire). Considérons k points $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$, et considérons l'ensemble

$$\mathcal{E}_k = \{(m(\mathbf{x}_1), \dots, m(\mathbf{x}_k)) \text{ pour } m \in \mathcal{M}\}.$$

Notons que les éléments de \mathcal{E}_k appartiennent à $\{-1, +1\}^k$. Autrement dit $|\mathcal{E}_k| \leq 2^k$. On dira que \mathcal{M} pulvérise l'ensemble des points si toutes les combinaisons sont possibles, c'est à dire $|\mathcal{E}_k| = 2^k$. Intuitivement, les étiquettes de l'ensemble de points ne procurent pas assez d'information sur la cible f , car tout est possible. La dimension VC de \mathcal{M} est alors

$$VC(\mathcal{M}) = \sup \{k \text{ tel que } \mathcal{M} \text{ pulvérise } \{\mathbf{x}_1, \dots, \mathbf{x}_k\}\}.$$

Par exemple si $\mathcal{X} = \mathbb{R}$ et que l'on considère l'ensemble des modèles (simples) de la forme $m_{a,b} = \mathbf{1}_{\pm}(x \in [a, b])$. Aucun ensemble de points $\{x_1, x_2, x_3\}$ ordonnés ne peut être pulvérisé car il suffit d'assigner respectivement $+1, -1$ et $+1$ à x_1, x_2 et x_3 respectivement, donc $VC < 3$. En revanche $\{0, 1\}$ est pulvérisé donc $VC \geq 2$. La dimension de cet ensemble de prédicteur est 2. Si on augmente d'une dimension, $\mathcal{X} = \mathbb{R}^2$ et que l'on considère l'ensemble des modèles (simples) de la forme $m_{\mathbf{a}, \mathbf{b}} = \mathbf{1}_{\pm}(x \in [\mathbf{a}, \mathbf{b}])$ (où $[\mathbf{a}, \mathbf{b}]$ désigne le rectangle), alors la dimension de \mathcal{M} est ici 4.

Pour introduire les SVM, plaçons nous dans le cas où $\mathcal{X} = \mathbb{R}^k$, et considérons des séparations par des hyperplans passant par l'origine (on dira homogènes), au sens où $m_{\mathbf{w}}(\mathbf{x}) = \mathbf{1}_{\pm}(\mathbf{w}^T \mathbf{x} \geq 0)$. On peut montrer qu'aucun ensemble de $k + 1$ points ne peut être pulvérisé par ces deux espaces homogènes dans \mathbb{R}^k , et donc $VC(\mathcal{M}) = k$. Si on rajoute une constante, au sens où $m_{\mathbf{w}, b}(\mathbf{x}) = \mathbf{1}_{\pm}(\mathbf{w}^T \mathbf{x} + b \geq 0)$, on peut montrer qu'aucun ensemble de $k + 2$ points ne peut être pulvérisé par ces deux espaces (non homogènes) dans \mathbb{R}^k , et donc $VC(\mathcal{M}) = k + 1$.

De cette dimension VC, on déduit le théorème dit fondamental de l'apprentissage : si \mathcal{M} est une classe de dimension $d = VC(\mathcal{M})$, alors il existe des constantes positives \underline{C} et \overline{C} telles que la complexité d'échantillon pour que \mathcal{M} soit PAC-apprenable vérifie

$$\underline{C}\epsilon^{-1}(d + \log[\delta^{-1}]) \leq n_{\mathcal{M}}(\epsilon, \delta) \leq \overline{C}\epsilon^{-1}(d \log[\epsilon^{-1}] + \log[\delta^{-1}]).$$

Le lien entre la notion d'apprentissage (tel que défini dans Vapnik (1984)) et la dimension VC a été clairement établi dans Blumer *et al.* (1989).

Néanmoins, si les travaux de Vapnik et Chervonenkis sont considérés comme fondateurs de l'apprentissage statistique, il convient de citer aussi les travaux de Thomas Cover dans les années 60 et 70, en particulier Cover (1965) sur les capacités des modèles linéaires et Cover & Hart (1967) sur l'apprentissage dans le contexte de l'algorithme des k plus proches voisins. Ces travaux ont fait le lien entre l'apprentissage, la théorie de l'information (avec l'ouvrage de référence Cover & Thomas (1991)), la complexité et la statistique. D'autres auteurs ont rapprochés les deux communautés, de l'apprentissage et de la statistique par la suite. Par exemple Halbert White proposait de voir les réseaux de neurones dans un contexte statistique dans White (1989), allant jusqu'à affirmer que « *learning procedures used to train artificial neural networks are inherently statistical techniques. It follows that statistical theory can provide considerable insight into the properties, advantages, and disadvantages of different network learning methods* ». Ce tournant à la fin des années 80 ancrera la théorie de l'apprentissage dans un contexte probabiliste.

3.3 Le choix de l'objectif et la fonction de perte

Ces choix (de l'objectif et de la fonction de perte) sont essentiels, et très dépendants du problème considéré. Commençons par décrire un modèle historiquement important, le « perceptron » de Rosenblatt (1958), introduit dans des problèmes de classification, où $y \in \{-1, +1\}$, inspiré par McCulloch & Pitts (1943). On dispose de données $\{(y_i, \mathbf{x}_i)\}$, 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 :

$$m(\mathbf{x}) = \mathbf{1}_{\pm}(\beta_0 + \mathbf{x}^T \boldsymbol{\beta} \geq 0) = \begin{cases} +1 & \text{si } \beta_0 + \mathbf{x}^T \boldsymbol{\beta} \geq 0 \\ -1 & \text{si } \beta_0 + \mathbf{x}^T \boldsymbol{\beta} < 0 \end{cases},$$

où les coefficients $\boldsymbol{\beta}$ 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 $\hat{y}_i^{(k)}$:

$$\hat{y}_i^{(k)} = m^{(k)}(\mathbf{x}_i) = \mathbf{1}_{\pm}(\beta_0^{(k)} + \mathbf{x}_i^T \boldsymbol{\beta}^{(k)} \geq 0),$$

avec, dans le cas du perceptron :

$$\beta_j^{(k+1)} = \beta_j^{(k)} + \eta \underbrace{(\mathbf{y} - \hat{\mathbf{y}}^{(k)})^T}_{=\ell(\mathbf{y}, \hat{\mathbf{y}}^{(k)})} \mathbf{x}_j$$

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 $y' = m(\mathbf{x})$ et en observant y . Pour un problème de régression, on peut considérer une erreur quadratique ℓ_2 , telle que $\ell(y, m(\mathbf{x})) = (y - m(\mathbf{x}))^2$ ou en valeur absolue ℓ_1 , avec $\ell(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$. Ici, pour notre problème de classification, nous utilisons une indicatrice de mauvaise qualification (on pourrait discuter 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 reconnaît dans le problème décrit auparavant une descente de gradient, et on voit que l'on cherche à résoudre :

$$m^*(\mathbf{x}) = \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \sum_{i=1}^n \ell(y_i, m(\mathbf{x}_i)) \right\} \quad (6)$$

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

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

engendrant la classe des prédicteurs linéaires. On aura alors l'estimateur qui minimise le risque empirique. Une partie des travaux récents en apprentissage statistique vise à étudier les propriétés de l'estimateur \widehat{m}^* , dit « oracle », dans une famille d'estimateurs \mathcal{M} ,

$$\widehat{m}^* = \operatorname{argmin}_{\widehat{m} \in \mathcal{M}} \{ \mathcal{R}(\widehat{m}, m) \}.$$

Cet estimateur est, bien entendu, impossible à définir car il dépend de m , le vrai modèle, inconnu.

Mais revenons un peu davantage sur ces fonctions de perte. Une fonction de perte ℓ est une fonction $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+$, symétrique, qui vérifie l'inégalité triangulaire, et telle que $\ell(\mathbf{x}, \mathbf{y}) = 0$ si et seulement si $\mathbf{x} = \mathbf{y}$. La norme associée est $\| \cdot \|$, telle que $\ell(\mathbf{x}, \mathbf{y}) = \| \mathbf{x} - \mathbf{y} \| = \ell(\mathbf{x} - \mathbf{y}, \mathbf{0})$ (en utilisant le fait que $\ell(\mathbf{x}, \mathbf{y} + \mathbf{z}) = \ell(\mathbf{x} - \mathbf{y}, \mathbf{z})$ - nous reverrons cette propriété fondamentale par la suite).

Pour une fonction de perte quadratique, on notera que l'on peut avoir une interprétation particulière de ce problème, puisque :

$$\bar{y} = \operatorname{argmin}_{m \in \mathbb{R}} \left\{ \sum_{i=1}^n \frac{1}{n} [y_i - m]^2 \right\} = \operatorname{argmin}_{m \in \mathbb{R}} \left\{ \sum_{i=1}^n \ell_2(y_i, m) \right\},$$

où ℓ_2 est la distance quadratique usuelle. Si l'on suppose - comme on le faisait en économétrie - qu'il existe un modèle probabiliste sous-jacent, et en notant que :

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

on notera que ce que l'on essaye d'obtenir ici, en résolvant le problème (6) en prenant pour ℓ la norme ℓ_2 , est une approximation (dans un espace fonctionnel donné, \mathcal{M}) de l'espérance conditionnelle $\mathbf{x} \mapsto \mathbb{E}[Y | \mathbf{X} = \mathbf{x}]$. Une autre fonction de perte particulièrement intéressante est la perte ℓ_1 , $\ell_1(y, m) = |y - m|$. Rappelons que

$$\text{médiane}(\mathbf{y}) = \operatorname{argmin}_{m \in \mathbb{R}} \left\{ \sum_{i=1}^n \ell_1(y_i, m) \right\}.$$

Le problème d'optimisation :

$$\widehat{m}^* = \operatorname{argmin}_{m \in \mathcal{M}_0} \left\{ \sum_{i=1}^n |y_i - m(\mathbf{x}_i)| \right\}$$

est obtenu en économétrie si on suppose que la loi conditionnelle de Y suit une loi de Laplace centrée sur $m(\mathbf{x})$, et en maximisant la (log) vraisemblance (la somme des valeurs absolues des erreurs correspond à la log-vraisemblance d'une loi de Laplace). On pourra noter d'ailleurs que si la loi conditionnelle de Y est symétrique par rapport à 0, la médiane et la moyenne coïncident. 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 $\tau \in (0, 1)$:

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

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 :

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

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

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

Koenker & Machado (1999) et Yu & Moyeed (2001) ont d'ailleurs noté un lien entre cette condition et la recherche du maximum de vraisemblance lorsque la loi conditionnelle de Y suit une loi de Laplace asymétrique.

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

$$T(Y) = \operatorname{argmin}_{x \in \mathbb{R}} \left\{ \int_{\mathbb{R}} \ell(x, y) dF(y) \right\} = \operatorname{argmin}_{x \in \mathbb{R}} \left\{ \mathbb{E}[\ell(x, Y)] \text{ où } Y \stackrel{\mathcal{L}}{\sim} F \right\}.$$

La moyenne (espérance mathématique) est ainsi elicitable par la distance quadratique, ℓ_2 , alors que la médiane est elicitable 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). Nous reviendrons sur l'utilisation de ces différentes normes dans la section sur la pénalisation.

3.4 Boosting et apprentissage séquentiel

Nous l'avons vu auparavant: la modélisation repose ici sur la résolution d'un problème d'optimisation, et résoudre le problème décrit par l'équation (6) 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 \mathbf{y} , à partir de \mathbf{X} , qui donnera une erreur ε_1 . à la seconde étape, on estime un modèle m_2 pour ε_1 , à partir de \mathbf{X} , qui donnera une erreur ε_2 , etc. On va alors retenir comme modèle, au bout de k itération :

$$m^{(k)}(\cdot) = \underbrace{m_1(\cdot)}_{\sim \mathbf{y}} + \underbrace{m_2(\cdot)}_{\sim \varepsilon_1} + \underbrace{m_3(\cdot)}_{\sim \varepsilon_2} + \cdots + \underbrace{m_k(\cdot)}_{\sim \varepsilon_{k-1}} = m^{(k-1)}(\cdot) + m_k(\cdot). \quad (7)$$

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

L'équation (7) 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

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

Le problème (6) est dual dans le sens où c'est la fonction f qui doit être optimisée. On pourrait alors écrire une descente de gradient de la forme :

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

où le terme \star peut être interprété comme un gradient, mais dans un espace fonctionnel, et non plus dans \mathbb{R}^p . Le problème (7) va alors se réécrire comme un problème d'optimisation :

$$m^{(k)} = m^{(k-1)} + \operatorname{argmin}_{h \in \mathcal{H}} \left\{ \sum_{i=1}^n \underbrace{\ell(y_i - m^{(k-1)}(\mathbf{x}_i), h(\mathbf{x}_i))}_{\varepsilon_{k,i}} \right\} \quad (8)$$

où l'astuce consiste à considérer un espace \mathcal{H} relativement simple (on parlera de « *weak learner* »). Classiquement, les fonctions \mathcal{H} sont des fonctions en escalier (que l'on retrouvera dans les arbres de classification et de régression) appelées « *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, $\varepsilon_1 = y - m_1(\mathbf{x})$, on

posera $\varepsilon_1 = y - \alpha \cdot m_1(\mathbf{x})$ avec $\alpha \in [0, 1]$. On notera que c'est parce qu'on utilise pour \mathcal{H} 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 $\widehat{\varepsilon} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ sont orthogonaux aux variables explicatives, \mathbf{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.

Cette présentation a l'avantage d'avoir une heuristique faisant penser à un modèle économétrique, en modélisant de manière itérative les résidus par un modèle (très) simple. Mais ce n'est souvent pas la présentation retenue dans la littérature en apprentissage, qui insiste davantage sur une heuristique d'algorithme d'optimisation (et d'approximation du gradient). La fonction est apprise de manière itérative, en partant d'une valeur constante,

$$m^{(0)} = \operatorname{argmin}_{m \in \mathbb{R}} \left\{ \sum_{i=1}^n \ell(y_i, m) \right\}.$$

puis on considère l'apprentissage suivant

$$m^{(k)} = m^{(k-1)} + \operatorname{argmin}_{h \in \mathcal{H}} \sum_{i=1}^n \ell(y_i, m^{(k-1)}(\mathbf{x}_i) + h(\mathbf{x}_i)), \quad (9)$$

qui peut s'écrire, si \mathcal{H} est un ensemble de fonctions différentiables,

$$m^{(k)} = m^{(k-1)} - \gamma_k \sum_{i=1}^n \nabla_{m^{(k-1)}} \ell(y_i, m^{(k-1)}(\mathbf{x}_i)), \quad (10)$$

où

$$\gamma_k = \operatorname{argmin}_{\gamma} \sum_{i=1}^n \ell \left(y_i, m^{(k-1)}(\mathbf{x}_i) - \gamma \nabla_{m^{(k-1)}} \ell(y_i, m^{(k-1)}(\mathbf{x}_i)) \right).$$

Pour mieux comprendre le lien avec l'approche décrite auparavant, à l'étape k , on définit des pseudo-résidus en posant

$$r_{i,k} = - \left. \frac{\partial \ell(y_i, m(\mathbf{x}_i))}{\partial m(\mathbf{x}_i)} \right|_{m(\mathbf{x})=m^{(k-1)}(\mathbf{x})} \quad \text{pour } i = 1, \dots, n.$$

On cherche alors un modèle simple pour expliquer ces pseudo-résidus en fonction des variables explicatives \mathbf{x}_i , i.e. $r_{i,k} = h^*(\mathbf{x}_i)$, où $h^* \in \mathcal{H}$. Dans un second temps, on cherche un multiplicateur optimal en résolvant

$$\gamma_k = \operatorname{argmin}_{\gamma \in \mathbb{R}} \left\{ \sum_{i=1}^n \ell(y_i, m^{(k-1)}(\mathbf{x}_i) + \gamma h^*(\mathbf{x}_i)) \right\}$$

puis on met à jour le modèle en posant $m_k(\mathbf{x}) = m_{k-1}(\mathbf{x}) + \gamma_k h^*(\mathbf{x})$. Plus formellement, on passe de l'équation (8) - qui montre clairement qu'on construit un modèle sur les résidus - à l'équation (9) - qui sera ensuite retraduit comme une problème de calcul de gradient - en notant que $\ell(y, m+h) = \ell(y-m, h)$. Classiquement, les fonctions \mathcal{H} sont construites avec des arbres de régression. Il est aussi possible d'utiliser une forme de pénalisation en posant $m_k(\mathbf{x}) = m_{k-1}(\mathbf{x}) + \nu \gamma_k h^*(\mathbf{x})$, avec $\nu \in (0, 1)$. Mais revenons un peu plus longuement sur l'importance de la pénalisation avant de discuter les aspects numériques de l'optimisation.

3.5 Pénalisation et choix de variables

Dans la section 2.9, nous avons évoqué le 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 :

$$(\widehat{\beta}_{0,\lambda}, \widehat{\beta}_{\lambda}) = \operatorname{argmin} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \mathbf{x}^T \boldsymbol{\beta}) + \lambda \text{ pénalisation}(\boldsymbol{\beta}) \right\}, \quad (11)$$

où la fonction de pénalisation sera souvent une norme $\|\cdot\|$ choisie a priori, et un paramètre de pénalisation λ (on retrouve en quelque sorte la distinction entre AIC et BIC si la fonction de pénalisation est la complexité du

| optimisation contrainte | | pénalisation | |
|-------------------------|---|--|----------|
| meilleur groupe | $\operatorname{argmin}_{\beta: \ \beta\ _{\ell_0} \leq s} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \mathbf{x}^\top \beta) \right\}$ | $\operatorname{argmin}_{\beta, \lambda} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \mathbf{x}^\top \beta) + \lambda \ \beta\ _{\ell_0} \right\}$ | (\ell 0) |
| LASSO | $\operatorname{argmin}_{\beta: \ \beta\ _{\ell_1} \leq s} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \mathbf{x}^\top \beta) \right\}$ | $\operatorname{argmin}_{\beta, \lambda} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \mathbf{x}^\top \beta) + \lambda \ \beta\ _{\ell_1} \right\}$ | (\ell 1) |
| Ridge | $\operatorname{argmin}_{\beta: \ \beta\ _{\ell_2} \leq s} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \mathbf{x}^\top \beta) \right\}$ | $\operatorname{argmin}_{\beta, \lambda} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \mathbf{x}^\top \beta) + \lambda \ \beta\ _{\ell_2} \right\}$ | (\ell 2) |

Table 1: Optimisation contrainte et régularisation.

modèle - le nombre de variables explicatives retenues). Dans le 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 $\|\beta\|_{\ell_2}$ comme dans la régression RIDGE. On peut toutefois envisager une vision Bayésienne de cette pénalisation. Rappelons que dans un modèle Bayésien :

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

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

Avant de revenir en détails sur ces deux estimateurs, obtenus en utilisant la norme ℓ_1 ou la norme ℓ_2 , revenons un instant sur un problème très proche : celui du meilleur choix de variables explicatives. Classiquement (et ça sera encore plus vrai en grande dimension), on peut disposer d'un grand nombre de variables explicatives, p , mais beaucoup sont juste du bruit, au sens où $\beta_j = 0$ pour un grand nombre de j . Soit s le nombre de covariables (réellement) pertinentes, $s = \#\mathcal{S}$ avec $\mathcal{S}(\beta) = \{j = 1, \dots, p; \beta_j \neq 0\}$. Si on note $\mathbf{X}_{\mathcal{S}}$ la matrice constituée des variables pertinentes (en colonnes), alors on suppose que le vrai modèle est de la forme $y = \mathbf{x}_{\mathcal{S}}^\top \beta_{\mathcal{S}} + \varepsilon$. Intuitivement, un estimateur intéressant serait alors $\hat{\beta}_{\mathcal{S}} = [\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}]^{-1} \mathbf{X}_{\mathcal{S}}^\top \mathbf{y}$, mais cet estimateur n'est que théorique car \mathcal{S} est ici inconnue. Cet estimateur est l'estimateur oracle évoqué auparavant. On peut alors être tenté de résoudre

$$(\hat{\beta}_{0,s}, \hat{\beta}_s) = \operatorname{argmin} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \mathbf{x}^\top \beta) \right\}, \text{ sous la contrainte } \#\mathcal{S}(\beta) = s.$$

Ce problème a été introduit par Foster & George (1994) en introduisant la norme ℓ_0 . Plus précisément, définissons ici les trois normes suivantes

$$\|\mathbf{a}\|_{\ell_0} = \sum_{i=1}^d \mathbf{1}(a_i \neq 0), \quad \|\mathbf{a}\|_{\ell_1} = \sum_{i=1}^d |a_i| \quad \text{et} \quad \|\mathbf{a}\|_{\ell_2} = \left(\sum_{i=1}^d a_i^2 \right)^{1/2}, \quad \text{pour } \mathbf{a} \in \mathbb{R}^d.$$

Considérons les problèmes d'optimisation de la Table 1. Si on considère le problème classique où ℓ est la norme quadratique, les deux problèmes de l'équation (\ell 1) de la Table 1 sont équivalents, au sens où, pour toute solution (β^*, s^*) au problème de gauche, il existe λ^* tel que (β^*, λ^*) soit solution du problème de droite; et inversement. Le résultat est également vrai pour les problèmes (\ell 2)⁷. Il s'agit en effet de problèmes convexes. En revanche, les deux problèmes (\ell 0) ne sont pas équivalents : si pour (β^*, λ^*) solution du problème de droite, il existe s^* tel que β^* soit solution du problème de gauche, la réciproque n'est pas vraie. Plus généralement, si on veut utiliser une norme ℓ_p , la sparsité est obtenue si $p \leq 1$ alors qu'il faut avoir $p \geq 1$ pour avoir la convexité du programme d'optimisation.

On peut être tenté de résoudre le programme pénalisé (\ell 0) directement, comme le suggère Foster & George (1994). Numériquement, c'est un problème combinatoire complexe en grande dimension (Natarajan (1995) note que c'est un problème NP-difficile), mais il est possible de montrer que si $\lambda \sim \sigma^2 \log(p)$, alors

$$\mathbb{E}([\mathbf{x}^\top \hat{\beta} - \mathbf{x}^\top \beta_0]^2) \leq \underbrace{\mathbb{E}(\mathbf{x}_{\mathcal{S}}^\top \hat{\beta}_{\mathcal{S}} - \mathbf{x}^\top \beta_0)^2}_{=\sigma^2 \#\mathcal{S}} \cdot (4 \log p + 2 + o(1)).$$

⁷Pour (\ell 1), s'il y a équivalence au niveau théorique, il peut exister des soucis numériques car il n'y a pas forcément unicité de la solution.

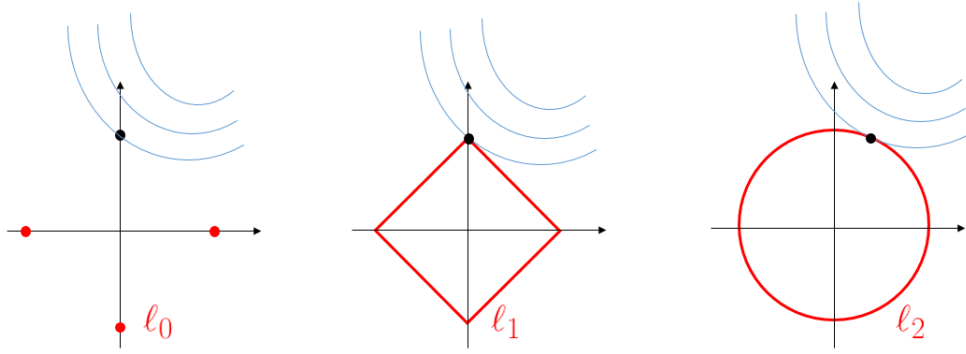


Figure 2: Pénalisation basée sur la norme ℓ_0 , ℓ_1 et ℓ_2 de β , respectivement (inspiré de Hastie *et al.* (2016)).

Notons que dans ce cas

$$\widehat{\beta}_{\lambda,j}^{\text{sub}} = \begin{cases} 0 & \text{si } j \notin \mathcal{S}_\lambda(\beta) \\ \widehat{\beta}_j^{\text{ols}} & \text{si } j \in \mathcal{S}_\lambda(\beta), \end{cases}$$

où $\mathcal{S}_\lambda(\beta)$ désigne l'ensemble des coordonnées non nulle lors de la résolution de (ℓ_0) .

Le problème (ℓ_2) est strictement convexe si ℓ est la norme quadratique, autrement dit, l'estimateur Ridge est toujours bien défini, avec en plus une forme explicite pour l'estimateur,

$$\widehat{\beta}_\lambda^{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I})^{-1} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I})^{-1} (\mathbf{X}^\top \mathbf{X}) \widehat{\beta}^{\text{ols}}.$$

Aussi, on peut en déduire que

$$\text{biais}[\widehat{\beta}_\lambda^{\text{ridge}}] = -\lambda [\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I}]^{-1} \widehat{\beta}^{\text{ols}} \quad \text{et} \quad \text{Var}[\widehat{\beta}_\lambda^{\text{ridge}}] = \sigma^2 [\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I}]^{-1} \mathbf{X}^\top \mathbf{X} [\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I}]^{-1}.$$

Avec une matrice de variables explicatives orthonormées (i.e. $\mathbf{X}^\top \mathbf{X} = \mathbb{I}$), les expressions se simplifient

$$\text{biais}[\widehat{\beta}_\lambda^{\text{ridge}}] = \frac{\lambda}{1+\lambda} \widehat{\beta}^{\text{ols}} \quad \text{et} \quad \text{Var}[\widehat{\beta}_\lambda^{\text{ridge}}] = \frac{\sigma^2}{(1+\lambda)^2} \mathbb{I}.$$

Notons que $\text{Var}[\widehat{\beta}_\lambda^{\text{ridge}}] < \text{Var}[\widehat{\beta}^{\text{ols}}]$. En notant que

$$\text{mse}[\widehat{\beta}_\lambda^{\text{ridge}}] = \frac{k\sigma^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2} \beta^\top \beta,$$

on obtient une valeur optimale pour λ : $\lambda^* = k\sigma^2 / \beta^\top \beta$.

En revanche, si ℓ n'est plus la norme quadratique mais la norme ℓ_1 , le problème (ℓ_1) n'est pas toujours strictement convexe, et en particulier, l'optimum n'est pas toujours unique (par exemple si $\mathbf{X}^\top \mathbf{X}$ est singulière). Mais le fait que ℓ soit strictement convexe $\mathbf{X}\widehat{\beta}$ sera unique. Notons de plus que deux solutions sont forcément cohérentes en terme de signe des coefficients : il n'est pas possible d'avoir $\widehat{\beta}_j < 0$ pour une solution et $\widehat{\beta}_j > 0$ pour une autre. D'un point de vue heuristique, le programme (ℓ_1) est intéressant car il permet d'obtenir dans bon nombre de cas une solution en coin, qui correspond à une résolution de problème de type (ℓ_0) - comme le montre de manière visuelle la Figure 2.

Considérons un modèle très simple: $y_i = x_i \beta + \varepsilon$, avec une pénalité ℓ_1 et une fonction de perte ℓ_2 . Le problème (ℓ_2) s'écrit alors

$$\min \{ \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{x} \beta + \beta \mathbf{x}^\top \mathbf{x} \beta + 2\lambda |\beta| \}$$

La condition du premier ordre est alors

$$-2\mathbf{y}^\top \mathbf{x} + 2\mathbf{x}^\top \mathbf{x} \widehat{\beta} \pm 2\lambda = 0.$$

le signe du dernier terme dépend du signe de $\widehat{\beta}$. Supposons que l'estimateur par moindres carrés (obtenu en posant $\lambda = 0$) soit (strictement positif), autrement dit $\mathbf{y}^\top \mathbf{x} > 0$. Si λ n'est pas trop grand, on peut imaginer que $\widehat{\beta}$ soit du même signe que $\widehat{\beta}^{\text{mco}}$, et donc la condition devient

$$-2\mathbf{y}^\top \mathbf{x} + 2\mathbf{x}^\top \mathbf{x} \widehat{\beta} + 2\lambda = 0.$$

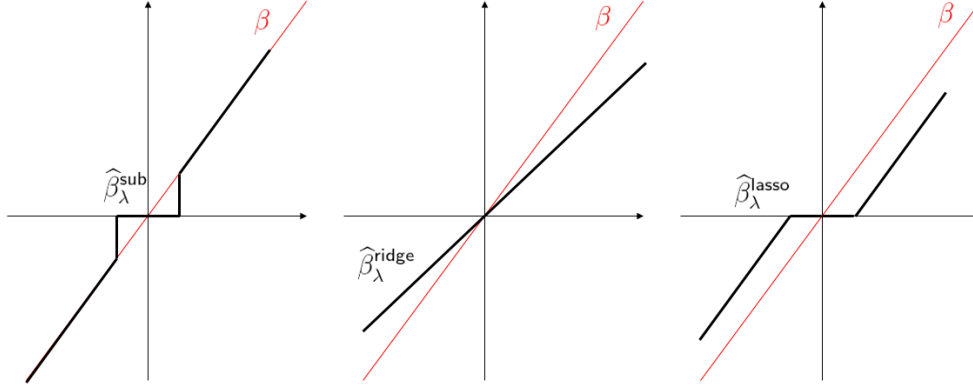


Figure 3: Pénalisation basée sur la norme ℓ_0 , ℓ_1 et ℓ_2 de β , respectivement (inspiré de Hastie *et al.* (2016)).

et la solution est

$$\hat{\beta}_\lambda^{\text{lasso}} = \frac{\mathbf{y}^\top \mathbf{x} - \lambda}{\mathbf{x}^\top \mathbf{x}}.$$

En augmentant λ , on va arriver à un moment où $\hat{\beta}_\lambda = 0$. Si on augmente encore un peu $\hat{\beta}_\lambda$ ne devient pas négatif car dans ce cas le dernier terme de la condition du premier ordre change, et dans ce cas on cherche à résoudre

$$-2\mathbf{y}^\top \mathbf{x} + 2\mathbf{x}^\top \mathbf{x} \hat{\beta} - 2\lambda = 0.$$

dont la solution est alors

$$\hat{\beta}_\lambda^{\text{lasso}} = \frac{\mathbf{y}^\top \mathbf{x} + \lambda}{\mathbf{x}^\top \mathbf{x}}.$$

Mais cette solution est positive (nous avons supposé $\mathbf{y}^\top \mathbf{x} > 0$), et donc il est possible d'avoir en même temps $\hat{\beta}_\lambda < 0$. Aussi, au bout d'un moment, $\hat{\beta}_\lambda = 0$, qui est alors une solution de coin. Les choses sont bien entendu plus compliquées en dimension plus grande (Tibshirani & Wasserman (2016) revient longuement sur la géométrie des solutions) mais comme le note Candès & Plan (2009), sous des hypothèses minimales garantissant que les prédicteurs ne sont pas fortement corrélés, le LASSO obtient une erreur quadratique presque aussi bonne que si l'on dispose d'un oracle fournissant des informations parfaites sur l'ensemble des β_j qui sont non nuls. Moyennant quelques hypothèses techniques supplémentaires, on peut montrer que cet estimateur est « sparsifiant » au sens où le support de $\hat{\beta}_\lambda^{\text{lasso}}$ est celui de β , autrement dit LASSO a permis de faire de la sélection de variables (plus de discussions sur ce point peuvent être obtenues dans Hastie *et al.* (2016)).

De manière plus générale, on peut montrer que $\hat{\beta}_\lambda^{\text{lasso}}$ est 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. Pour comparer les trois techniques, par rapport à l'estimateur par moindres carrés (obtenu quand $\lambda = 0$), si on suppose que les variables explicatives sont orthonormées, alors

$$\hat{\beta}_{\lambda,j}^{\text{sub}} = \hat{\beta}_j^{\text{ols}} \mathbf{1}_{|\hat{\beta}_{\lambda,j}^{\text{sub}}| > b}, \quad \hat{\beta}_{\lambda,j}^{\text{ridge}} = \frac{\hat{\beta}_j^{\text{ols}}}{1 + \lambda} \quad \text{et} \quad \hat{\beta}_{\lambda,j}^{\text{lasso}} = \text{signe}[\hat{\beta}_j^{\text{ols}}] \cdot (|\hat{\beta}_j^{\text{ols}}| - \lambda)_+.$$

3.6 Optimisation et aspects algorithmiques

En économétrie, l'optimisation (numérique) est devenue omniprésente dès que l'on a quitté le modèle Gaussien. Nous l'avons rapidement évoqué dans la section sur la famille exponentielle, et l'utilisation du score de Fisher (descente de gradient) pour résoudre la condition du premier ordre $\mathbf{X}^\top \mathbf{W}(\beta)^{-1}[\mathbf{y} - \hat{\mathbf{y}}] = \mathbf{0}$. En apprentissage, l'optimisation est l'outil central. Et il est nécessaire d'avoir des algorithmes d'optimisation efficaces, pour résoudre des problèmes de la forme :

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\{ \sum_{i=1}^n \ell(y_i, \beta_0 + \mathbf{x}^\top \beta) + \lambda \|\beta\| \right\}.$$

Dans certains cas, au lieu de faire de l'optimisation globale, il suffit de considérer de l'optimisation par coordonnées (largement étudiée dans Daubechies et al. (2004)). Si $f : \mathbb{R}^d \rightarrow \mathbb{R}$ est convexe et différentiable, alors

$$\text{si } \mathbf{x} \text{ vérifie } f(\mathbf{x} + h\mathbf{e}_i) \geq f(\mathbf{x}) \text{ pour tout } h > 0 \text{ et } i \in \{1, \dots, d\}, \text{ alors } f(\mathbf{x}) = \min\{f\},$$

où $\mathbf{e} = (\mathbf{e}_i)$ est la base canonique de \mathbb{R}^d . Cette propriété n'est toutefois pas vraie dans le cas non-différentiable. Mais si on suppose que la partie non-différentiable est séparable (additivement), elle redevient vraie. Plus précisément, si

$$f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^d h_i(x_i) \text{ avec } \begin{cases} g : \mathbb{R}^d \rightarrow \mathbb{R} \text{ convexe-différentiable} \\ h_i : \mathbb{R} \rightarrow \mathbb{R} \text{ convexe.} \end{cases}$$

C'est le cas pour la régression LASSO, $f(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\ell_2} + \lambda\|\boldsymbol{\beta}\|_{\ell_1}$, comme le montre Tsen (2001). On peut alors utiliser un algorithme de descente par coordonnées: à partir d'une valeur initiale $\mathbf{x}^{(0)}$, on considère (en itérant)

$$x_j^{(k)} \in \operatorname{argmin}\{f(x_1^{(k)}, \dots, x_{k-1}^{(k)}, x_k, x_{k+1}^{(k-1)}, \dots, x_n^{(k-1)})\} \text{ pour } j = 1, 2, \dots, n.$$

Ces problèmes algorithmiques peuvent paraître secondaires à des économètres. Ils sont pourtant essentiels en apprentissage machine: une technique est intéressante s'il existe un algorithme stable et rapide, qui permet d'obtenir une solution. Ces techniques d'optimisation sont d'ailleurs transposables : par exemple, on pourra utiliser cette technique de descente par coordonnées dans le cas des méthodes SVM (dit « à support vecteur ») lorsque l'espace n'est pas linéairement séparable, et qu'il convient de pénaliser l'erreur de classification (nous reviendrons sur cette technique dans la prochaine section).

3.7 *In-sample, out-of-sample* et validation croisée

Ces techniques semblent intellectuellement intéressantes, mais nous n'avons pas encore abordé le choix du paramètre de pénalisation λ . Mais ce problème est en fait plus général, car comparer deux paramètres $\widehat{\boldsymbol{\beta}}_{\lambda_1}$ et $\widehat{\boldsymbol{\beta}}_{\lambda_2}$ revient en fait à comparer deux modèles. En particulier, si on utilise une méthode de type LASSO, avec des seuils λ différents, on compare des modèles qui n'ont pas la même dimension. Dans la section 2.9, nous avons abordé le problème de la comparaison de modèles sous l'angle économétrique (en pénalisant les modèles trop complexes). Dans la littérature en apprentissage, juger de la qualité d'un modèle sur les données qui ont servi à le construire ne permet en rien de savoir comment le modèle se comportera sur des nouvelles données. Il s'agit du problème dit de « généralisation ». L'approche classique consiste alors à séparer l'échantillon (de taille n) en deux : une partie qui servira à entraîner le modèle (la base d'apprentissage, *in-sample*, de taille m) et une partie qui servira à tester le modèle (la base de test, *out-of-sample*, de taille $n - m$). Cette dernière permet alors de mesurer un vrai risque prédictif. Supposons que les données soient générées par un modèle linéaire $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i$ où les ε_i sont des réalisations de lois indépendantes et centrées. Le risque quadratique empirique *in-sample* est ici

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}([\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} - \mathbf{x}_i^\top \boldsymbol{\beta}_0]^2) = \mathbb{E}([\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} - \mathbf{x}_i^\top \boldsymbol{\beta}_0]^2),$$

pour n'importe quelle observation i . En supposant les résidus ε Gaussiens, alors on peut montrer que ce risque vaut $\sigma^2 \operatorname{trace}(\mathbf{\Pi}_{\mathcal{X}})/m$ soit $\sigma^2 p/m$. En revanche le risque quadratique empirique *out-of-sample* est ici

$$\mathbb{E}([\mathbf{x}^\top \widehat{\boldsymbol{\beta}} - \mathbf{x}^\top \boldsymbol{\beta}_0]^2)$$

où \mathbf{x} est une nouvelle observation, indépendante des autres. On peut noter que

$$\mathbb{E}([\mathbf{x}^\top \widehat{\boldsymbol{\beta}} - \mathbf{x}^\top \boldsymbol{\beta}_0]^2 | \mathbf{x}) = \sigma^2 \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x},$$

et en intégrant par rapport à \mathbf{x} ,

$$\mathbb{E}([\mathbf{x}^\top \widehat{\boldsymbol{\beta}} - \mathbf{x}^\top \boldsymbol{\beta}_0]^2) = \mathbb{E}(\mathbb{E}([\mathbf{x}^\top \widehat{\boldsymbol{\beta}} - \mathbf{x}^\top \boldsymbol{\beta}_0]^2 | \mathbf{x})) = \sigma^2 \operatorname{trace}(\mathbb{E}[\mathbf{x}\mathbf{x}^\top] \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}]).$$

L'expression est alors différente de celle obtenue *in-sample*, et en utilisation la majoration de Groves & Rothenberg (1969), on peut montrer que

$$\mathbb{E}([\mathbf{x}^\top \widehat{\boldsymbol{\beta}} - \mathbf{x}^\top \boldsymbol{\beta}_0]^2) \geq \sigma^2 \frac{p}{m},$$

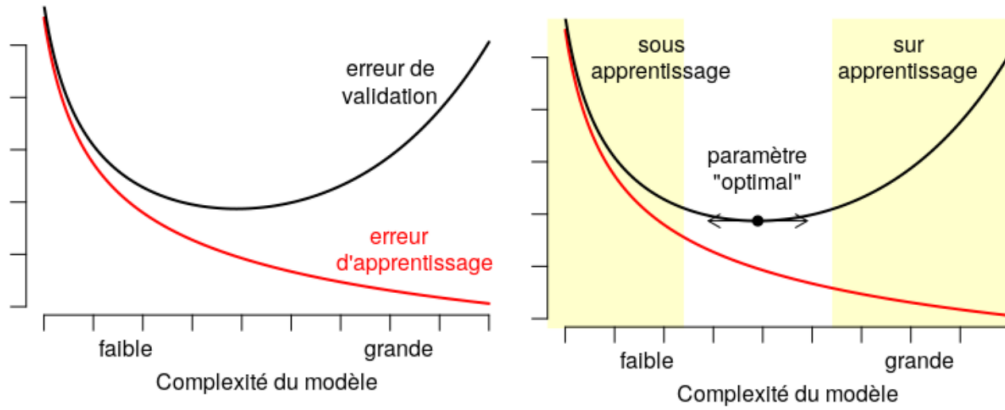


Figure 4: Généralisation, et sur-apprentissage.

ce qui est assez intuitif, finalement. Hormis certains cas simple, il n'y a pas de formule simple. Notons toutefois que si $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I})$, alors $\mathbf{x}^\top \mathbf{x}$ suit une loi de Wishart, et on peut montrer que

$$\mathbb{E}[(\mathbf{x}^\top \hat{\boldsymbol{\beta}} - \mathbf{x}^\top \boldsymbol{\beta}_0)^2] = \sigma^2 \frac{p}{m - p - 1}.$$

Si on regarde maintenant la version empirique: si $\hat{\boldsymbol{\beta}}$ est estimé sur les m premières observations,

$$\hat{\mathcal{R}}^{\text{IS}} = \sum_{i=1}^m [y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}]^2 \quad \text{et} \quad \hat{\mathcal{R}}^{\text{OS}} = \sum_{i=m+1}^n [y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}]^2,$$

et comme l'a noté Leeb (2008), $\hat{\mathcal{R}}^{\text{IS}} - \hat{\mathcal{R}}^{\text{OS}} \approx 2 \cdot \nu$ où ν représente le nombre de degrés de libertés, qui n'est pas sans rappeler la pénalisation utilisée dans le critère d'Akaike.

La Figure 4 montre l'évolution respective de $\hat{\mathcal{R}}^{\text{IS}}$ et $\hat{\mathcal{R}}^{\text{OS}}$ en fonction de la complexité du modèle (nombre de degrés dans une régression polynomiale, nombre de noeuds dans des splines, etc). Plus le modèle est complexe, plus $\hat{\mathcal{R}}^{\text{IS}}$ va diminuer (c'est la courbe rouge). Mais ce n'est pas ce qui nous intéresse ici : on veut un modèle qui prédise bien sur de nouvelles données (autrement dit *out-of-sample*). Comme le montre la Figure 4, si le modèle est trop simple, il prédit mal (tout comme sur les données *in-sample*). Mais ce que l'on peut voir, c'est que si le modèle est trop complexe, on est dans une situation de « sur-apprentissage » : le modèle va commencer à modéliser le bruit.

Au lieu de séparer la base en deux, avec une partie des données qui vont servir à calibrer le modèle et une autre à étudier sa performance, il est aussi possible d'utiliser la validation croisée. Pour présenter l'idée générale, on peut revenir au « jackknife », introduit par Quenouille (1949) (et formalisé par Quenouille (1956) et Tukey (1958)) et utilisé en statistique pour réduire le biais. En effet, si on suppose que $\{y_1, \dots, y_n\}$ est un échantillon tiré suivant une loi F_θ , et que l'on dispose d'un estimateur $T_n(\mathbf{y}) = T_n(y_1, \dots, y_n)$, mais que cet estimateur est biaisé, avec $\mathbb{E}[T_n(\mathbf{Y})] = \theta + O(n^{-1})$, il est possible de réduire le biais en considérant :

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

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

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 obtenue avec le modèle estimé, et l'observation manquante :

$$\hat{\mathcal{R}}^{\text{CV}} = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{m}_{(i)}(\mathbf{x}_i))$$

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 ${}_t\hat{y}_{t+1} = \alpha \cdot {}_{t-1}\hat{y}_t + (1 - \alpha) \cdot y_t$, avec $\alpha \in [0, 1]$, et on va considérer :

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

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 $\{1, \dots, n\}$ en k groupes (ou blocs) de même taille, $\mathcal{I}_1, \dots, \mathcal{I}_k$, et notons $\mathcal{I}_{\bar{j}} = \{1, \dots, n\} \setminus \mathcal{I}_j$. En notant $\hat{m}_{(j)}$ construit sur l'échantillon $\mathcal{I}_{\bar{j}}$, on pose alors :

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

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, comme le rappelle James *et al.* (2013).

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

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

où $n_{\bar{b}}$ 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 bostrappé, 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 minimisation du AIC est à rapprocher du critère de validation croisée, et Shao (1997) a montré que la minimisation du BIC correspond à de la validation croisée de type k -fold, avec $k = n / \log n$.

4 Quelques outils d'apprentissage automatique

4.1 Réseaux de Neurones

Les réseaux de neurones sont des modèles semi-paramétriques. Néanmoins, cette famille de modèles peut être appréhendée de la même manière que les modèles non-paramétriques: la structure des réseaux de neurones (présentée par la suite) peut être modifiée afin d'étendre la classe des fonctions utilisées pour approcher une variable d'intérêt. Plus précisément, Cybenko (1989) a démontré que l'ensemble des fonctions neuronales est dense dans l'espace des fonctions continues sur un compact. En d'autres termes, on a un cadre théorique permettant de garantir une forme d'approximation universelle. Il impose en outre une définition d'un neurone et met en avant l'existence d'un nombre de neurones suffisant pour approcher toute fonction continue sur un compact. Ainsi, un phénomène continu peut être approché par une suite de neurones: on appellera cette suite « réseau de neurones à une couche ». Si ce théorème d'approximation universelle est démontré en 1989, le premier neurone artificiel fonctionnel fut introduit par Franck Rosenblatt au milieu du XXième siècle, dans Rosenblatt (1958). Ce neurone, qualifié de nos jours de « neurone élémentaire », porte le nom de « Perceptron ». Il a permis dans ses premières utilisations de 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 formalisées 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 une 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. 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, et de faire le lien avec les notations économétriques usuelles. 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 fini (classiquement $\{0, 1\}$, mais on préférera ici $\{-1, +1\}$);
- un vecteur de paramètres $\mathbf{w} \in \mathbb{R}^p$
- une fonction d'activation $\phi : \mathbb{R} \rightarrow \mathbb{R}$. Cette fonction doit être dans l'idéal monotone, dérivable et bornée (on dira ici « saturante ») afin de s'assurer de certaines propriétés de convergence.

Cette dernière fonction ϕ fait penser aux transformations logistique ou probit, populaires en économétrie (qui sont des fonctions de répartition, à valeur dans $[0, 1]$, idéal quand \mathcal{Y} est l'ensemble $\{0, 1\}$). Pour les réseaux de neurones, on utilisera plutôt la tangente hyperbolique, la fonction arc-tangente ou les fonctions sigmoïdes pour des problèmes de classification sur $\mathcal{Y} = \{-1, +1\}$ (ces dernières évoqueront la transformation logistique de économètres). On appellera neurone toute application $f_{\mathbf{w}}$ de \mathcal{X} dans \mathcal{Y} définie par :

$$y = f_{\mathbf{w}}(\mathbf{x}) = \phi(\mathbf{w}^T \mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}.$$

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

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

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, \mathbf{w} le vecteur de paramètres, y la variable à expliquer et \mathbf{x} le vecteur des variables explicatives de même dimension que \mathbf{w} :

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

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 $Y|\mathbf{X}$ (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 (détaillé dans Cybenko (1989)). Ainsi par exemple, une régression logistique peut être perçue comme un neurone alors que les régressions linéaires généralisées 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. On parlera de structure de réseaux de neurones par couche. Chaque couche de neurones recevant à chaque fois le même vecteur d'observation.

Pour revenir à une analogie plus économétrique, on peut imaginer passer par une étape intermédiaire (on reviendra sur cette construction dans la Figure 5), par exemple en ne faisant pas une régression sur les variables brutes \mathbf{x} mais un ensemble plus faible de variables orthogonales, obtenues par exemple suite à une analyse en composantes principales. Soit \mathbf{A} la matrice associée à cette transformation linéaire, avec \mathbf{A} de taille $k \times p$ si on souhaite utiliser les p premières composantes. Notons \mathbf{z} la transformation de \mathbf{x} , au sens où $\mathbf{z} = \mathbf{A}^T \mathbf{x}$, ou encore $z_j = \mathbf{A}_j^T \mathbf{x}$. Un généralisation du modèle précédant peut être de poser

$$y = f(\mathbf{x}) = \phi(\mathbf{w}^T \mathbf{z}) = \phi(\mathbf{w}^T \mathbf{A}^T \mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X},$$

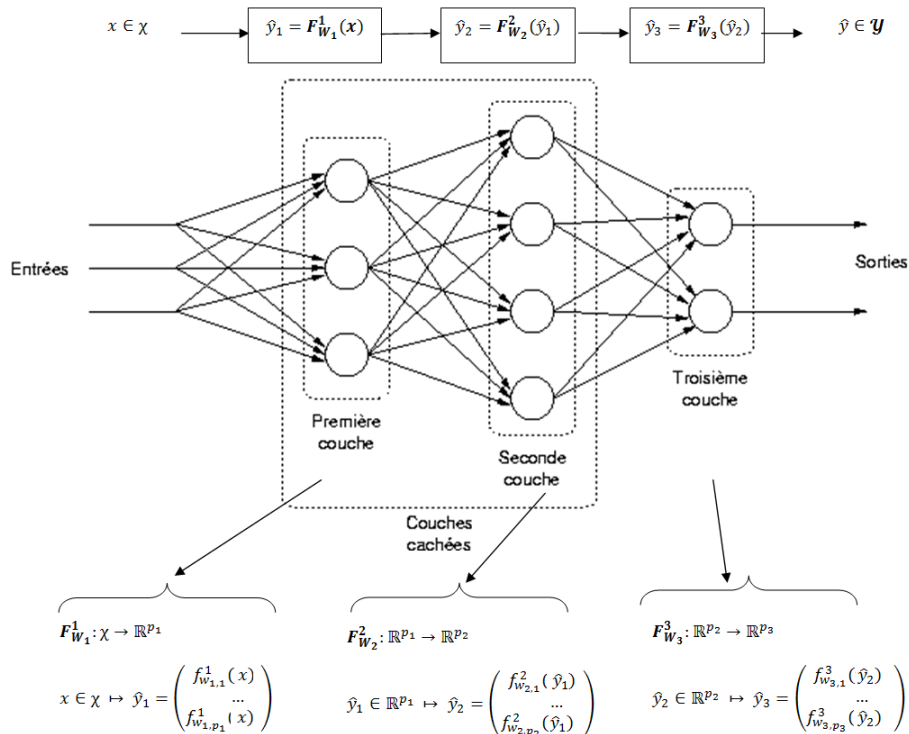
où cette fois $\mathbf{w} \in \mathbb{R}^p$. On a ici une transformation linéaire (en considérant une analyse en composante principale) mais on peut imaginer une généralisation avec des transformée non-linéaire, avec une fonction de la forme

$$y = f(x) = \phi(\mathbf{w}^\top F_A(\mathbf{x})), \quad \forall \mathbf{x} \in \mathcal{X},$$

où F est ici une fonction $\mathbb{R}^k \rightarrow \mathbb{R}^p$. C'est le réseau de neurone à deux couches. Plus généralement, pour formaliser la construction, on introduit les notations suivantes :

- $K \in \mathbb{N}^*$: nombre de couches;
- $\forall k \in \{1, \dots, K\}$, p_k représente le nombre de neurones dans la couche k ;
- $\forall k \in \{1, \dots, K\}$, 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 $l \in \{1, \dots, p_k\}$, $w_{k,l} \in \mathbb{R}^{p_{k-1}}$ désigne le vecteur de poids associé au neurone élémentaire l de la couche k ;
- on appellera $W = \{W_1, \dots, W_K\}$, l'ensemble des paramètres associés au réseau de neurones.
- $F_{W_k}^k : \mathbb{R}^{p_{k-1}} \rightarrow \mathbb{R}^{p_k}$ 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 $k \in \{1, \dots, K\}$;
- on appellera $F = F_W = F^1 \circ \dots \circ F^K$ la fonction de transfert associée au réseau global. A ce titre, si $\mathbf{x} \in \mathcal{X}$, on pourra noter $\hat{\mathbf{y}} = F_W(\mathbf{x})$.

Figure 5: Exemple de notations associées aux réseaux de neurones multicouche.



La Figure 5 permet d'illustrer les notations présentées ici⁸. 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 $\mathbf{x} \in \mathcal{X}$, de couche de sortie pour la couche fournissant en « output » la prédiction $\hat{\mathbf{y}} \in \mathcal{Y}$. Les autres couches sont couramment appelées couches cachées.

⁸Source: <http://intelligenceartificielle.org>.

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 $\{1, \dots, K\}$. Chaque fonction d'activation associée à chaque neurone (chaque cercle de la Figure 5) 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 $k \in \{1, \dots, K\}$), 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 $(y_i, \mathbf{x}_i) \in \mathcal{Y} \times \mathcal{X}$ est de minimiser le risque empirique :

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

Afin d'illustrer les propos, intéressons nous à l'exemple suivant qui illustrera également la démarche opérée. Supposons que nous observons un phénomène y aux travers de n observations $y_i \in [-1, 1]$. On souhaiterait expliquer ce phénomène à partir des variables explicatives \mathbf{x} que l'on suppose à valeurs réelles. La « théorie de l'approximation universelle » nous indique qu'un réseau à une couche de neurones devrait permettre de modéliser le phénomène (sous hypothèse qu'il soit continue). On note toutefois que ce théorème ne donne pas de vitesse de convergence. Il est alors laissé à l'utilisateur le choix de la structure. Ainsi par exemple une première structure pourrait être un simple neurone dont la fonction d'activation serait la fonction tangente hyperbolique.

On aurait ainsi comme premier modèle:

$$y_1 = \tanh(w_0 + w_1 x)$$

où les paramètres w_0, w_1 sont les paramètres à optimiser de sorte que sur les données d'apprentissage, le risque empirique soit minimal.

Si l'on suit toujours la philosophie du théorème d'approximation universelle, en ajoutant plusieurs neurones, l'erreur est censée diminuer. Cependant, ne connaissant pas la fonction à estimer, on ne peut l'observer qu'aux travers de l'échantillon. Ainsi, mécaniquement, on s'attend à ce que plus on ajoute de paramètres, plus l'erreur sur la base d'apprentissage diminue. L'analyse de l'erreur sur la base de test permet alors d'évaluer notre capacité à généraliser (cf partie précédente).

On peut ainsi s'intéresser à un second modèle qui cette fois utilise plusieurs neurones. Par exemple, considérons le modèle

$$y_2 = w_a \tanh(w_0 + w_1 x) + w_b \tanh(w_2 + w_3 x) + w_c \tanh(w_4 + w_5 x)$$

où les paramètres w_0, \dots, w_5 ainsi que w_a, w_b, w_c sont les paramètres à optimiser. Calibrer un réseaux de neurones revient alors à réitérer ces étapes de modification de la structure jusqu'à minimisation du risque sur la base de test.

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 $W^* = (W_1, \dots, W_K)$ de sorte que :

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

De cette formule apparaît l'importance du choix de la fonction ℓ . Cette fonction de perte 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 préfère 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. Ceci explique l'analogie avec la philosophie des modèles non-paramétriques faite auparavant.

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).

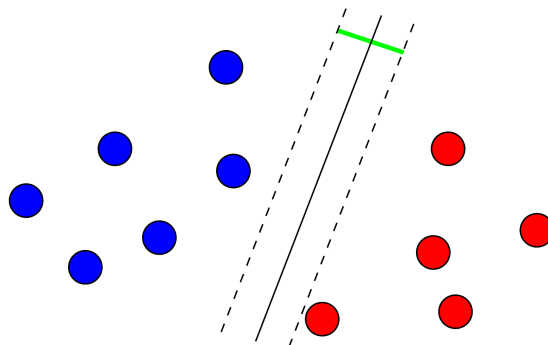
L’extraction d’informations peut, par exemple, se faire grâce à la convolution. Procédé non supervisé, il a permis notamment d’obtenir d’excellentes performances dans l’analyse d’image. Techniquement, cela peut s’apparenter à une transformation à noyaux (comme utilisé dans les techniques SVM). Si une 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 précédent. 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 profond, en particulier les réseaux 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 type « *deep learning* » sont plus difficiles à appréhender car ils nécessitent beaucoup de jugement empirique. En effet, si aujourd’hui les bibliothèques 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

Comme nous l’avions noté auparavant, dans les problèmes de classification en apprentissage machine (comme en traitement du signal) on préférera avoir des observations dans l’ensemble $\{-1, +1\}$ (plutôt que $\{0, 1\}$, comme en économétrie). Avec cette notation, 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 Vector 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.

Figure 6: Schéma d’illustration d’un SVM à marge, Vert (2017).



Si un plan peut être caractérisé entièrement par un vecteur directeur \mathbf{w} orthogonal à ce dernier et une constante b , appliquer un algorithme SVM à un ensemble de $n \in \mathbb{N}^*$ points \mathbf{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 LASSO (distance quadratique sous contrainte linéaire). Plus particulièrement, on sera amené à résoudre :

$$(\mathbf{w}^*, b^*) = \underset{\mathbf{w}, b}{\operatorname{argmin}} \{ \|\mathbf{w}\|^2 \} = \underset{\mathbf{w}, b}{\operatorname{argmin}} \{ \mathbf{w}^\top \mathbf{w} \},$$

$$\text{sous contrainte } \forall i \in \{1, \dots, n\}, \begin{cases} \boldsymbol{\omega}^\top \mathbf{x}_i + b \geq +1 & \text{lorsque } y_i = +1 \\ \boldsymbol{\omega}^\top \mathbf{x}_i + b \leq -1 & \text{lorsque } y_i = -1 \end{cases}$$

La contrainte peut être relâchée en autorisant que dans un sous-ensemble, un point puisse ne pas être du 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* »). De manière heuristique, comme en pratique, bien souvent, on ne peut pas avoir $y_i(\boldsymbol{\omega}^\top \mathbf{x}_i + b) - 1 \geq 0$ pour tout $i \in \{1, \dots, n\}$, on relâche en introduisant des variables positives ξ telle que

$$\begin{cases} \boldsymbol{\omega}^\top \mathbf{x}_i + b \geq +1 - \xi_i & \text{lorsque } y_i = +1 \\ \boldsymbol{\omega}^\top \mathbf{x}_i + b \leq -1 + \xi_i & \text{lorsque } y_i = -1 \end{cases} \quad (12)$$

avec $\xi_i \geq 0$. On a une erreur de classification si $\xi_i > 1$, et on va alors introduire une pénalité, un coût à payer pour chaque erreur commise. On cherche alors à résoudre un problème quadratique

$$\min \left\{ \frac{1}{2} \boldsymbol{\omega}^\top \boldsymbol{\omega} + C \mathbf{1}^\top \mathbf{1}_{\xi > 1} \right\}$$

sous la contrainte (12), qui pourra être résolu de manière numérique très efficacement par descente de coordonnées (décrit auparavant).

S'il n'est pas possible de séparer les points, une autre astuce possible 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 définissant les transformations $T(\cdot)$ et les produits scalaires via un noyau $K(\mathbf{x}_1, \mathbf{x}_2) = \langle T(\mathbf{x}_1), T(\mathbf{x}_2) \rangle$. L'un des choix les plus courants pour une fonction de noyau est la fonction de base radiale (noyau gaussien) $K(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\|\mathbf{x}_1 - \mathbf{x}_2\|^2)$. Il n'existe néanmoins pas de règles à ce jour permettant de choisir le « meilleur » noyau. Comme mentionné au début de la section précédente, cette technique est basé sur de la minimisation de distance, et il n'a aucune prévision de la probabilité d'être positif ou négatif (mais une interprétation probabiliste est néanmoins possible, comme le montre Grandvalet *et al.* (2005), par exemple).

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 $H(\mathbf{x})$ est associée à la quantité de désordre dans les données \mathbf{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 que toutes les feuilles soient pures, c'est à dire composées 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: on laisse l'arbre grossir, puis on coupe certaines branches *a posteriori* (ce qui est différent de l'introduction d'un critère d'arrêt *a priori* au processus de croissance de l'arbre - par exemple en imposant une taille minimale aux feuilles, ou d'autres critères discutés dans Breiman (2001a)).

à un nœud donné, constitué de n_0 observations (\mathbf{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}_g et \mathcal{I}_d . Soit I le critère d'intérêt, comme l'entropie du nœud (ou plutôt du nœud vu en tant que feuille):

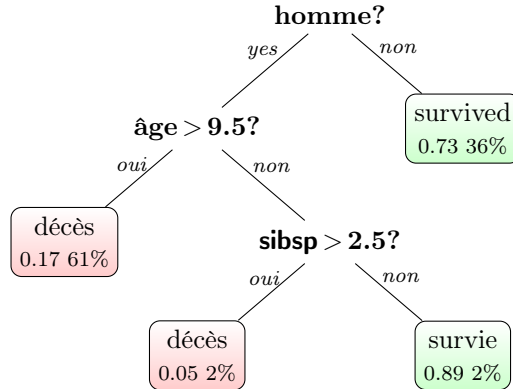
$$I(\mathbf{y}_0) = -n_0 p_0 \log p_0 \quad \text{où} \quad p_0 = \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} y_i,$$

ou la variance du nœud:

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

ce dernier étant également l'indice d'impureté de Gini.

Figure 7: 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 $I(\mathbf{y}_0) - [I(\mathbf{y}_g) + I(\mathbf{y}_d)]$ 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 $\mathcal{I}_g = \{i \in \mathcal{I}_0 : x_{k,i} < s\}$ et $\mathcal{I}_d = \{i \in \mathcal{I}_0 : x_{k,i} > s\}$, 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 7 sur les données du Titanic⁹), 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 $\{(\mathbf{x}_i, y_i)\}$. Ces ensembles d'arbres - naturellement appelés « forêts » - une fois agrégés donnent souvent de bien meilleurs résultats que les arbres isolés, mais ils sont difficiles à interpréter. Ces techniques ressemblent toutefois beaucoup à ce qui est fait lorsque l'on utilise les techniques de bootstrap 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 d'arbres. C'est l'agrégation 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 prévision.

Les forêts aléatoires, ou « *random forests* » reposent sur le même principe que le « *bagging* », mais en plus, lors de la construction d'un arbre de classification, à chaque branche, un sous-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'amplifier la variabilité entre les différents arbres et d'obtenir, au final, une forêt composée d'arbres moins corrélés les uns aux autres.

4.4 Sélection de modèle de classification

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

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

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

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

⁹Ce 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>.

| | $y = 0$ | $y = 1$ | |
|-----------------|----------------------------------|----------------------------------|----------------------------------|
| $\hat{y}_s = 0$ | TN _s | FN _s | TN _s +FN _s |
| $\hat{y}_s = 1$ | FP _s | TP _s | FP _s +TP _s |
| | TN _s +FP _s | FN _s +TP _s | n |

Table 2: Matrice de confusion, ou tableau de contingence pour un seuil s donné.

pour $(u, v) \in \{0, 1\}$. La Table 2 présente un tel tableau, avec le nom de chacun des éléments : TP (*true positive*) sont les vrais positifs, correspondant aux 1 prédits en 1, TN (*true negative*) sont les vrais négatifs, correspondant aux 0 prédits en 0, FP (*false positive*) sont les faux positifs, correspondant aux 0 prédits en 1, et enfin FN (*false negative*) sont les faux négatifs, correspondant aux 1 prédits 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. 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* ») :

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

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

Les deux grandeurs intensivement utilisées en apprentissage automatique 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 \mathbf{N}^\perp le tableau de contingence correspond aux cas indépendants (défini à partir de \mathbf{N} dans le test d'indépendance du chi-deux). On pose alors

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

alors que

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

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 symétriques (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 autant 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 d'apprentissage automatique se sont orientées autour des problèmes de classification (avec éventuellement plus de 2 modalités¹⁰), 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 (5), avec un algorithme proche de celui de la méthode CART

¹⁰Par exemple dans le cas de reconnaissance de lettres ou de chiffres.

décrite auparavant. Dans le contexte de la classification, on calculait, à chaque nœud (dans le cas de l'indice d'impureté de Gini) en sommant sur la feuille de gauche $\{x_{k,i} < s\}$ et celle de droite $\{x_{k,i} > s\}$

$$I = \sum_{i:x_{k,i} < s} \bar{y}_g(1 - \bar{y}_g) + \sum_{i:x_{k,i} > s} \bar{y}_d(1 - \bar{y}_d)$$

où \bar{y}_g et \bar{y}_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

$$I = \sum_{i:x_{k,i} < s} (y_i - \bar{y}_g)^2 + \sum_{i:x_{k,i} > s} (y_i - \bar{y}_d)^2$$

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 la régression, on utilise la moyenne des prédictions, sur tous les arbres.

Dans la partie précédente, nous avons présenté la dimension « apprentissage » de l'apprentissage automatique 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 (8) sous la forme :

$$m^{(k)}(\mathbf{x}) = m^{(k-1)}(\mathbf{x}) + \alpha_k \operatorname{argmin}_{h \in \mathcal{H}} \left\{ \sum_{i=1}^n (y_i, m^{(k-1)}(\mathbf{x}) + h(\mathbf{x}))^2 \right\}$$

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

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

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

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

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

Tout comme les modèles additifs, les fonctions g_1, \dots, g_k sont à estimer, tout comme les directions $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_k$. Cette écriture est relativement générale, et permet de tenir compte d'interactions 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

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

autrement dit $g_1(x) = x^2/4$, $g_2(x) = -x^2/4$, $\boldsymbol{\omega}_1 = (1, 1)^\top$ et $\boldsymbol{\omega}_2 = (1, -1)^\top$. Dans la version simple, avec $k = 1$, avec une fonction de perte quadratique, on peut utiliser un développement de Taylor pour approcher $[y_i - g(\boldsymbol{\omega}^\top \mathbf{x}_i)]^2$, et construire classiquement un algorithme itératif. Si on dispose d'une valeur initiale $\boldsymbol{\omega}_0$, notons que

$$\sum_{i=1}^n [y_i - g(\boldsymbol{\omega}^\top \mathbf{x}_i)]^2 \approx \sum_{i=1}^n g'(\boldsymbol{\omega}_0^\top \mathbf{x}_i)^2 \left[\boldsymbol{\omega}^\top \mathbf{x}_i + \frac{y_i - g(\boldsymbol{\omega}_0^\top \mathbf{x}_i)}{g'(\boldsymbol{\omega}_0^\top \mathbf{x}_i)} - \boldsymbol{\omega}_0^\top \mathbf{x}_i \right]^2$$

qui correspondrait à l'approximation dans les modèles linéaires généralisés sur la fonction $g(\cdot)$ qui était la fonction de lien (supposée connue). On reconnaît 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.¹¹ 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 de l'apprentissage automatique 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 de l'apprentissage automatique 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 jeu 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).¹² Une variable dépendante binaire est artificiellement créée, 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 :

$$y^* = \gamma + \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim G(0, 1), \quad (13)$$

où \mathbf{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 observée, mais seulement y , avec :

$$y = \begin{cases} 1 & \text{si } y^* > \xi, \\ 0 & \text{si } y^* \leq \xi. \end{cases} \quad (14)$$

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

$$\mathbb{P}(Y = 1) = G(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}) \quad (15)$$

où $\beta_0 = \gamma - \xi$.¹³ 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 $\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$ 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 :

$$\mathbb{P}[Y = 1 | \mathbf{X} = \mathbf{x}] = G(h(\mathbf{x})) \quad (16)$$

¹¹Entre autres, voir Hastie *et al.* (2009) et James *et al.* (2013).

¹²C'est le jeu de données *Carseats* de la bibliothèque ISLR.

¹³ $\mathbb{P}[Y = 1] = \mathbb{P}[Y^* > \xi] = \mathbb{P}[\gamma + \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon > \xi] = \mathbb{P}[\varepsilon > \xi - \gamma - \mathbf{x}^\top \boldsymbol{\beta}] = \mathbb{P}[\varepsilon < \gamma - \xi + \mathbf{x}^\top \boldsymbol{\beta}]$. En posant $\gamma - \xi = \beta_0$, on obtient $\mathbb{P}[Y = 1] = G(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta})$. En général, on suppose que le terme d'erreur est de variance σ^2 , auquel cas les paramètres du modèle (15) deviennent β_0/σ et $\boldsymbol{\beta}/\sigma$, ce qui veut dire que les paramètres du modèle latent (13) ne sont pas identifiables, ils sont estimés à un paramètre d'échelle près.

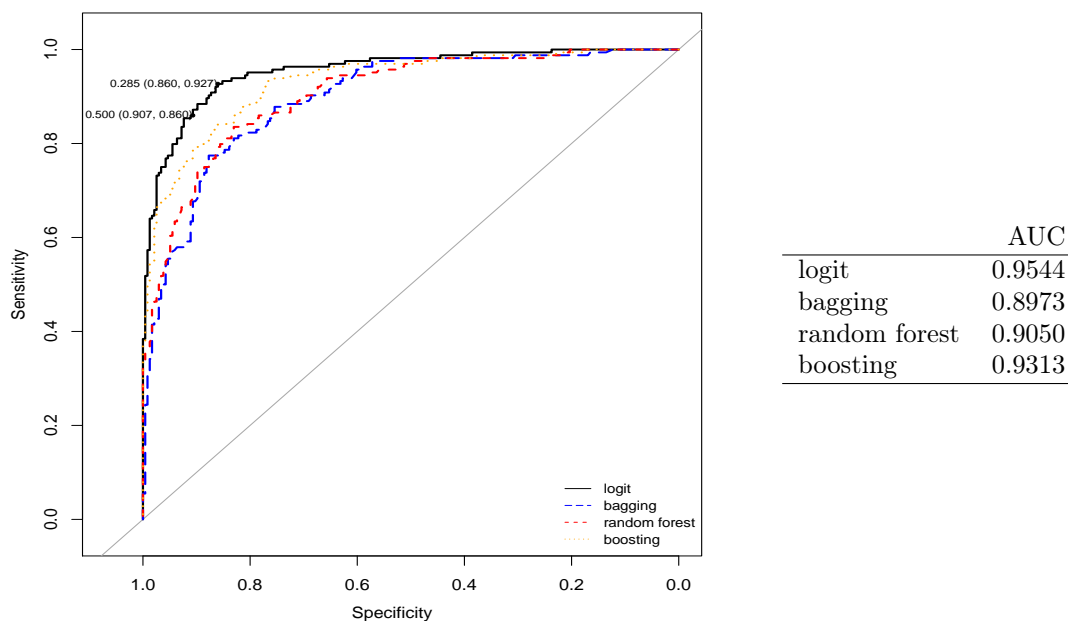


Figure 8: Ventes de sièges auto: courbes ROC et aires sous la courbe (AUC).

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 *a 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 (15) avec la loi logistique pour G , et le modèle (16) 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 8 présente la courbe ROC, ainsi que l'aire sous la courbe (AUC), pour les estimations logit, « *bagging* », forêt aléatoire 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 (on parle de « *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 8, 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 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 jeu de données contient 85 variables sur les caractéristiques démographiques de 5822 individus.¹⁴ La variable dépendante (Purchase) indique si l'individu a acheté une assurance caravane, c'est une variable binaire, égale à « oui » ou « non »

¹⁴C'est le jeu de données Caravan de la bibliothèque ISLR sous R.

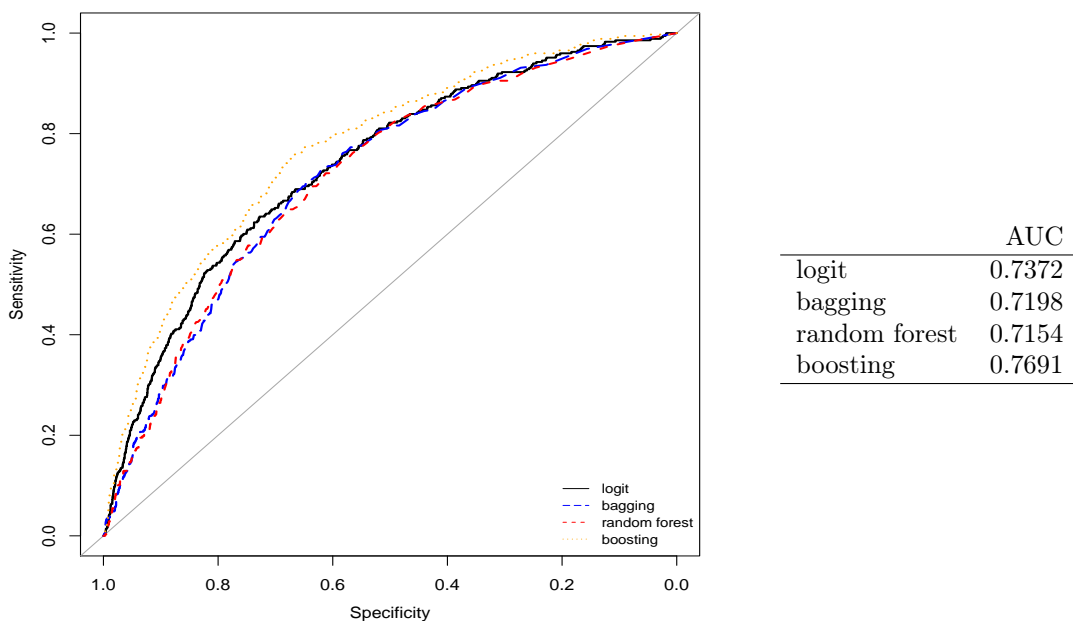


Figure 9: Achat d’assurance: courbes ROC et aires sous la courbe (AUC).

| | 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 |
| <i>random forest</i> | 0.9892 | 0.0316 | 0.6345 | 0.6954 |
| boosting | 0.9987 | 0.0000 | 0.6860 | 0.7385 |

Table 3: Achat d’assurance: sensibilité au choix du seuil de classification.

». Dans le jeu de données, seulement 6% des individus ont pris une telle assurance. Les classes sont donc fortement déséquilibrées.

Nous estimons le modèle (15) avec la loi logistique et le modèle (16) 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 9 présente la courbe ROC, ainsi que l’aire sous la courbe (AUC), pour les estimations logit, « *bagging* », forêt aléatoire 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 » si la probabilité estimée qu’il réponde « non » est supérieure à 0.5, sinon c’est « oui »). La partie gauche du tableau 3 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 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 déséquilibrées

(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 3 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 4. Une approche mentionnée avant qui peut être utile est le LASSO, en pénalisant convenablement la norme ℓ_1 du vecteur de paramètres β . 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 premières variables considérées comme non nulles (pour un λ assez grand) sont les deux premières 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 nœuds 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 4 sont présentées les variables par ordre d'importance décroissante, lorsque l'indice utilisé est l'indice d'impureté de Gini.

Avec l'approche *stepwise* et l'approche LASSO, on reste sur des modèles logistiques linéaires. Dans le cas des forêts aléatoires (et des arbres), des interactions 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 prédictives (troisième variable la plus importante).

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):

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{ed} + \beta_2 \text{exp} + \beta_3 \text{exp}^2 + \beta_4 \text{fe} + \varepsilon \quad (17)$$

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 (17) 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

| 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 4: Crédit: choix de variables, tri séquentiel, par approche *stepwise*, par fonction d'importance dans une forêt aléatoire et par LASSO.

| $\widehat{\mathcal{R}}^{10-CV}$ | Modèle (17) | | Modèle (18) | | |
|---------------------------------|-------------|---------|-------------|----------|----------|
| | OLS | Splines | Bagging | R.Forest | Boosting |
| out-of-sample | 0.2006 | 0.2004 | 0.2762 | 0.2160 | 0.2173 |

Table 5: Salaire: analyse de validation croisée par blocs ($K = 10$) : performances de l'estimation des modèles linéaire (17) et entièrement non-paramétrique (18).

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 (17) impose également que la différence salariale entre les hommes et les femmes est indépendante 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'interactions*).

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

$$\log(\text{wage}) = m(\text{ed}, \text{exp}, \text{fe}) + \varepsilon \quad (18)$$

où $m(\cdot)$ est une fonction quelconque. Il a l'avantage de pouvoir tenir compte de relations non-linéaires quelconques et d'interactions 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 sur R.¹⁵ Nous estimons les deux modèles et utilisons une analyse de validation croisée par 10 blocs pour sélectionner la meilleure approche. Le modèle paramétrique (17) est estimé par Moindres Carrés Ordinaires (OLS). Le modèle entièrement non-paramétrique (18) 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 5 présente les résultats de la validation croisée en 10 blocs (*10-fold cross-validation*). Le meilleur modèle est celui qui minimise le critère $\widehat{\mathcal{R}}^{10-CV}$. Les résultats montrent que le modèle (17) est au

¹⁵C'est le jeu de données CPS1985 de la bibliothèque AER.

| $\widehat{\mathcal{R}}^{10-CV}$ | Modèle (19) | Modèle (20) | | |
|---------------------------------|-------------|-------------|----------|----------|
| | 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 6: Prix des logements à Boston: analyse de validation croisée par blocs ($K = 10$): performances de l'estimation des modèles linéaire (19) et entièrement non-paramétrique (20).

moins aussi performant que le modèle (18), ce qui suggère que le modèle paramétrique (17) 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 jeu 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'âge moyen des maisons (age) et le pourcentage de ménages dont la catégorie socio-professionnelle est peu élevée (lstat).¹⁶

Considérons le modèle de régression linéaire suivant :

$$\text{medv} = \alpha + \mathbf{x}^T \boldsymbol{\beta} + \varepsilon \quad (19)$$

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

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

$$\text{medv} = m(\mathbf{x}) + \varepsilon. \quad (20)$$

L'estimation de ce modèle avec les méthodes du noyau ou les splines peut être problématique, car le nombre de variables est relativement élevé (il y a ici 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 (19) est estimé par Moindres Carrés Ordinaires (OLS) et le modèle entièrement non-paramétrique (20) 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 6 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 servi à estimer le modèle, pour calculer l'erreur quadratique. à partir des résultats in-sample, les méthodes de « *bagging* » et de forêt aléatoire paraissent incroyablement plus performantes que l'estimation OLS du modèle linéaire (19), le critère $\widehat{\mathcal{R}}^{10-CV}$ 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 $\widehat{\mathcal{R}}^{10-CV}$ 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 (19) et du modèle entièrement non-paramétrique (20) 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'interactions sont présentes dans la relation entre le prix des logements, medv , et les variables explicatives \mathbf{x} . Ce résultat nous conduit à chercher une meilleure spécification paramétrique.

à partir du modèle paramétrique (19), et afin de prendre en compte d'éventuelles non-linéarités, le modèle additif généralisé (GAM) suivant peut être considéré :

$$\text{medv} = m_1(x_1) + m_2(x_2) + \dots + m_{13}(x_{13}) + \varepsilon, \quad (21)$$

¹⁶C'est le jeu 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 7: Prix des logements à Boston: mesures de l’importance de chacune des variables dans l’estimation de forêt aléatoire du modèle (20), en considérant tout l’échantillon.

où m_1, m_2, \dots, 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’interactions.

L’estimation du modèle additif généralisé (21) par la méthode des *splines*, dans le cadre d’une analyse de validation croisée par 10-blocs, donne une valeur $\hat{\mathcal{R}}^{10-CV} = 13.643$. Par rapport au modèle paramétrique (19), il y a un gain important (13.643 vs. 24.082). Mais la différence avec le modèle entièrement non-paramétrique (20) 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’interactions entre les variables sont présents. Nous pourrions inclure dans le modèle les variables d’interactions 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’interactions 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 de forêt aléatoire s’accompagne de mesures de l’importance de chacune des variables dans l’estimation du modèle (décrit dans la section précédente). Le tableau 7 présente ces mesures dans le cadre du modèle (20), 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 d’interactions liées à ces deux variables seulement, qui sont les plus importantes.

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

$$\text{medv} = m_1(x_1) + \dots + m_{13}(x_{13}) + (\text{rm}:x) \gamma + (\text{lstat}:x) \delta + \varepsilon, \quad (22)$$

où $(\text{rm}:x)$ représente les variables d’interactions de **rm** avec toutes les autres variables de x et $(\text{lstat}:x)$ représente les variables d’interactions de **lstat** avec toutes les autres variables de x .¹⁷ 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 10. 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 noeud.

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

$$\text{medv} = \alpha + \mathbf{x}^T \boldsymbol{\beta} + (\text{dis} - 2)_+ \theta + (\text{rm}:x) \gamma + (\text{lstat}:x) \delta + \varepsilon \quad (23)$$

¹⁷On a $(\text{rm}:x) = [\text{rm} \times \text{chas}, \text{rm} \times \text{nox}, \text{rm} \times \text{age}, \text{rm} \times \text{tax}, \text{rm} \times \text{indus}, \text{rm} \times \text{rad}, \text{rm} \times \text{dis}, \text{rm} \times \text{lstat}, \text{rm} \times \text{crim}, \text{rm} \times \text{black}, \text{rm} \times \text{zn}, \text{rm} \times \text{ptratio}]$ et $(\text{lstat}:x) = [\text{lstat} \times \text{chas}, \text{lstat} \times \text{nox}, \text{lstat} \times \text{age}, \text{lstat} \times \text{tax}, \text{lstat} \times \text{indus}, \text{lstat} \times \text{rad}, \text{lstat} \times \text{dis}, \text{lstat} \times \text{crim}, \text{lstat} \times \text{black}, \text{lstat} \times \text{zn}, \text{lstat} \times \text{ptratio}]$.

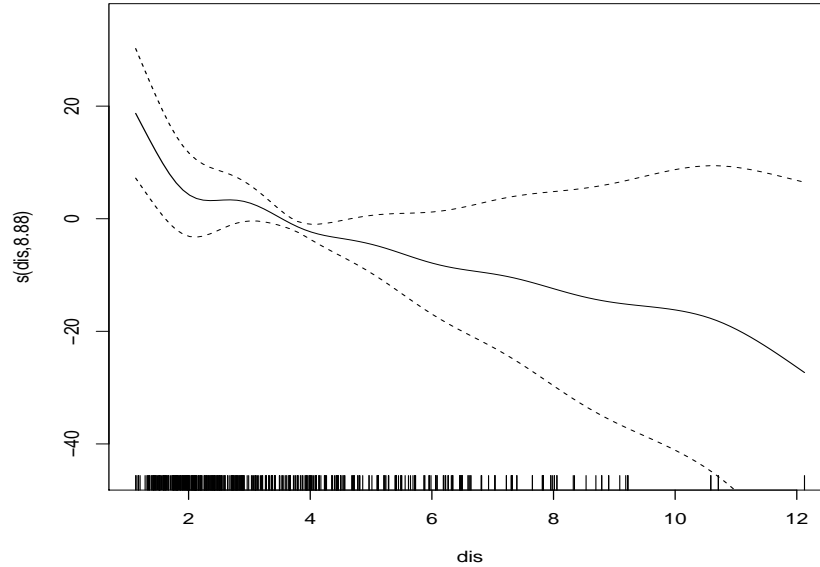


Figure 10: Estimation de la relation $m_7(x_7)$ dans le modèle additif généralisé (22), où $x_7 = \text{dis}$.

| | Modèle (19) | Modèle (21) | Modèle (23) |
|--|-------------|-------------|-------------|
| $\widehat{\mathcal{R}}^{10\text{-CV}}$ | OLS | Splines | OLS |
| out-of-sample | 24.082 | 13.643 | 11.759 |

Table 8: Prix des logements à Boston: analyse de validation croisée par blocs ($K = 10$) : performances de l'estimation du modèle linéaire (19) et du modèle linéaire (23) incluant les effets d'interactions et une non-linéarité par morceaux.

où $(\text{dis} - 2)_+$ est égal à la valeur de son argument si ce dernier est positif, et à 0 sinon. Par rapport au modèle linéaire initial, ce modèle inclut une relation linéaire par morceaux avec la variable dis , ainsi que des effets d'interactions entre rm , lstat et chacune des autres variables de \mathbf{x} .

Le tableau 8 présente les résultats de la validation croisée en 10 blocs (*10-fold cross-validation*) de l'estimation des modèles paramétriques (19) et (23), estimés par Moindres Carrés Ordinaires (OLS), et du modèle additif généralisé (21) estimé par les splines. Il montre que l'ajout des variables d'interactions et de la relation linéaire par morceaux dans le modèle (23) donne des résultats beaucoup plus performants que le modèle initial (19): le critère $\widehat{\mathcal{R}}^{10\text{-CV}}$ est divisé par plus de deux, il passe de 24.082 à 11.759. En comparant ces résultats avec ceux du tableau 6, on constate également que le modèle paramétrique (23), estimé par OLS, est aussi performant que le modèle général (20) estimé par *boosting* ($\widehat{\mathcal{R}}^{10\text{-CV}} = 11.789$). La différence avec les méthodes *bagging* et forêt aléatoire n'est quant à elle pas très importante ($\widehat{\mathcal{R}}^{10\text{-CV}} = 9.59, 9.407$)

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'interactions appropriées.

6 Conclusion

Si les « deux cultures » (ou les deux communautés) de l'économétrie et de l'apprentissage automatique se sont développées en parallèle, le nombre de passerelles entre les deux ne cesse d'augmenter. Alors que Varian (2014) présentait les apports importants de l'économétrie à la communauté de l'apprentissage automatique, 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 d'apprentissage automatique 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 d'apprentissage automatique 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 de l'apprentissage automatique 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.
- Bottou, L. (2010) Large-Scale Machine Learning with Stochastic Gradient Descent *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT’2010)*, 177–187.
- 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.
- Bickel, P.J., Gotze, F. & van Zwet, W. (1997). Resampling fewer than n observations: gains, losses and remedies for losses. *Statistica Sinica*, **7**, 1-31.
- 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.
- Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M.K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, **36**:4, 929–965.
- Breiman, L. (2001a). Statistical Modeling: The Two Cultures. *Statistical Science*, **16**:3, 199–231.
- Bühlmann, P. & van de Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. Springer Verlag.
- 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.
- Candès, E. & Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, **37**:5, 2145–2177.
- 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.
- Cover, T.M. (1965). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on Electronic Computers*, **14**:3, 326–334.
- Cover, T.M. & Hart, P. (1965). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**:1, 21 – 27.
- Cover, T.M. & Thomas, J. (1991). Elements of Information Theory. Wiley.
- 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 probabilités a estimation exhaustive. *Comptes Rendus de l'Académie des Sciences, Paris*, **200** 1265–1266.
- Daubechies, I. Defrise, M. & De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with sparsity constraint. *Communications on Pure and Applied Mathematics*, **57**:11, 1413–1457
- 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.
- Dhillon, P., Lu, Y. Foster, D.P. & Ungar, L.H. (2014). New Subsampling Algorithms for Fast Least Squares Regression. *in Advances in Neural Information Processing Systems 26*, Burges, Bottou, Welling, Ghahramani & Weinberger Eds., Curran Associates.
- 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.
- Foster, D.P. & George, E.I. (1994). The Risk Inflation Criterion for Multiple Regression. *The Annals of Statistics*, **22**:4, 1947–1975.
- Friedman, J.H. (1997). Data Mining and Statistics: What's the Connection. *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*.
- 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**.
- Groves, T. & Rothenberg, T. (1969). A note on the expected value of an inverse matrix. *Biometrika*, **56**:3, 690–691.

- 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.
- Hastie, T., Tibshirani, R. & Tibshirani, R.J. (2016). Extended comparisons of best subset selection, forward stepwise selection and the Lasso. *ArXiv*, <https://arxiv.org/abs/1707.08692>.
- 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.
- Hendry, D F. & Krolzig, H.-M. (2001). Automatic Econometric Model Selection. Timberlake Press.
- 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.
- Hoerl, A.E. & Kennard, R.W. (1981). Ridge regression: biased estimation for nonorthogonal problems *This Week’s Citation Classic*, ISI, <http://bit.ly/2H9LGiD>
- 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 conventional models. *Applied Soft Computing*, **11**, 5477–5484.
- Kean, M.P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, **156**, 3–20.
- leiner, A., Talwalkar, A., Sarkar , P. & Jordan, M. (2012). The Big Data Bootstrap. arXiv:1206.6415 .
- 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.
- Koenker, R. & Machado, J. (1999). Goodness of fit and related inference processes for quantile regression *Journal of the American Statistical Association*, **94**, 1296-1309.
- 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.
- Leeb, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli* **14**:3, 661–690.
- 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.

- Li, C., Li, Q., Racine, J. & Zhang, D. (2017). Optimal Model Averaging Of Varying Coefficient Models. *Department of Economics Working Papers 2017-01*, McMaster University.
- 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.
- McCulloch, W.S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:4, 115–133.
- 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 Statistical 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.
- Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing (SICOMP)*, 24 227–234.
- 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, republicé 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 Matematik, 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.
- Shai, S.S. & Shai, B.D. (2014). Understanding Machine Learning From Theory to Algorithms. Cambridge University Press.
- Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association* **88**:(422), 486–494.
- Shalev-Shwartz, S. & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.
- Shao, J. (1997). An Asymptotic Theory for Linear Model Selection. *Statistica Sinica*, **7**, 221–264.
- Shapire, R.E. & Freund, Y. (2012). Boosting. MIT Press.
- Silverman, B.W. (1986) Density Estimation. Chapman & Hall.
- 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.
- Tibshirani, R. & Wasserman, L. (2016). A Closer Look at Sparse Regression. <http://bit.ly/2FrGQ32>
- 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.
- Tropp, (2011). Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3:1, 115–126.
- Tsen, P. (2001). Convergence of a block coordinate descent for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:3, 475–494.
- 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.
- Vailiant, L.G. (1984). A theory of the learnable. *Communications of the ACM* 27:11, 1134–1142.
- 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.
- Vert, J.P. (2017). Machine learning in computational biology. ENSAE.

- 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.
- Watt, J., Borhani, R. & Katsaggelos, A. (2016). Machine Learning Refined : Foundations, Algorithms, and Applications. Cambridge University Press.
- White, H. (1989). Learning in Artificial Neural Networks: A Statistical Perspective. *Neural Computation*, **1**:4, 425–464.
- Widrow, B. & Hoff, M.E. Jr. (1960). Adaptive Switching Circuits. IRE WESCON Convention Record, **4**:96–104.
- Wolpert, D.H., Macready, W.G. (1997), No Free Lunch Theorems for Optimization, *IEEE Transactions on Evolutionary Computation* **1**, 67.
- Wolpert, David (1996), The Lack of A Priori Distinctions between Learning Algorithms, *Neural Computation*, 1341-1390.
- Working, E. J. (1927). What do statistical ‘demand curves’ show? *Quarterly Journal of Economics*, **41**:212–35.
- Yu, K. & Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, **54**, 437–447.
- Zinkevich M.A., Weimer, M., Smola, A. & Li, L. (2010). Parallelized Stochastic Gradient. *Advances in neural information processing systems*, 2595–2603.