Complément en ligne – Histoire et fondements des modèles économétriques et d'apprentissage automatique

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 leCs 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, x_i) sont vues comme des réalisations de variables aléatoires (Y_i, X_i). En pratique, la loi jointe du couple (Y, X) nous intéresse toutefois peu : la loi de X est inconnue (toute l'analyse sera faite conditionnellement aux observations x_i) et c'est la loi de Y conditionnelle à X qui nous intéressera. Dans la suite, nous noterons x une observation, $\boldsymbol{x}$ un vecteur d'observations, X une variable aléatoire, et $\boldsymbol{X}$ un vecteur aléatoire et, abusivement, $\boldsymbol{X}$ pourra aussi désigner la matrice des observations individuelles (les x_i), suivant le contexte.

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 paramètres[1]. 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.
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.

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

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

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

On parle de modèle linéaire car $\mathbb{E}[Y|X = x] = \beta_0 + x^T\beta$ est une combinaison linéaire des variables explicatives. C'est un modèle homoscédastique si $\text{Var}[Y|X = x] = \sigma^2$, 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 :

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

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 :

$$(\hat{\beta}_0, \hat{\beta}, \hat{\sigma}^2) = argmax\{\log\mathcal{L}(\beta_0, \beta, \sigma^2|y, x)\}$$

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 $X^T[y - X\hat{\beta}] = 0$, que l'on peut aussi écrire $(X^TX)\hat{\beta} = X^Ty$. Si la matrice X est de plein rang colonne, alors on retrouve l'estimateur classique :

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \beta + (X^TX)^{-1}X^T\varepsilon \quad (2)$$

En utilisant une écriture basée sur les résidus (comme souvent en économétrie), $y = x^T\beta + \varepsilon$. Le théorème de Gauss Markov assure que cet estimateur est l'estimateur linéaire sans biais de variance minimale. On peut alors montrer que $\hat{\beta} \overset{\mathcal{L}}{\sim} \mathcal{N}(\beta, \sigma^2[X^TX]^{-1})$, et en particulier :

$$\mathbb{E}[\hat{\beta}] = \beta \text{ et } \text{Var}[\hat{\beta}] = \sigma^2[X^TX]^{-1}$$

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) sans supposer forcement un modèle Gaussien. Si on suppose que $Y|X = x \overset{\mathcal{L}}{\sim} x^T\beta + \varepsilon$, avec $\mathbb{E}[\varepsilon] = 0$, $\text{Var}[\varepsilon] = \sigma^2$, $\text{Cov}[\varepsilon, X_j] = 0$ pour tout j, alors $\hat{\beta}$ est un estimateur sans biais de $\beta(\mathbb{E}[\hat{\beta}] = \beta)$ et de variance minimale parmi les estimateurs sans biais linéaires, avec $\text{Var}[\hat{\beta}] = \sigma^2[X^TX]^{-1}$. De plus, cet estimateur est asymptotiquement normal $\sqrt{n}(\hat{\beta} - \beta) \overset{\mathcal{L}}{\to} \mathcal{N}(0, \Sigma)$ lorsque $n \to \infty$.

La condition d'avoir une matrice X de plein rang peut être (numériquement) forte en grande dimension. Si elle n'est pas vérifiée, $\hat{\beta} = (X^TX)^{-1}X^Ty$ n'existe pas. Si $\mathbb{I}$ désigne la matrice identité, notons toutefois que $(X^TX + \lambda\mathbb{I})^{-1}X^Ty$ existe toujours, pour $\lambda > 0$. Cet estimateur est appelé l'estimateur ridge de niveau λ (introduit dans les années 1960 par Hoerl (1962), et associé à une régularisation étudiée par Tikhonov, (1963). Cet estimateur apparaît naturellement dans un contexte d'économétrie Bayesienne.

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 :

$$y_i = \beta_0 + x_i{}^T\beta + \varepsilon_i \tag{3}$$

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

$$\hat{\varepsilon}_i = y_i - [\hat{\beta}_0 + x_i{}^T\hat{\beta}]$$

Ces résidus sont l'outil de base pour diagnostiquer la pertinence du modèle. Une extension du modèle décrit par l'équation (1) a été proposée pour tenir compte d'un éventuel caractère hétéroscédastique :

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

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

$$y_i = \beta_0 + x_i{}^T\beta + \sigma^2(x_i) \cdot \varepsilon_i$$

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

$$\varepsilon_i = \frac{y_i - [\beta_0 + x_i{}^T\beta]}{\sigma(x_i)}$$

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 également 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).

Géométrie du modèle linéaire Gaussien

Définissons le produit scalaire dans $\mathbb{R}^n$, $\langle a, b\rangle = a^T b$, et notons $\|\cdot\|$ la norme euclidienne associée, $\|a\| = \sqrt{a^T a}$ (notée $\|\cdot\|_{\ell_2}$ dans la suite). Notons $\mathcal{E}_X$ l'espace engendré par l'ensemble des combinaisons linéaires des composantes x (en rajoutant la constante). Si les variables explicatives sont linéairement indépendantes, X est de plein rang colonne et $\mathcal{E}_X$ est un sous-espace de dimension $p + 1$. Supposons à partir de maintenant que les variables 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 $m(x) = \langle x, \beta \rangle$. L'espace $\mathcal{H}_z = \{x \in \mathbb{R}^k : m(x) = z\}$ est un hyperplan (affine) qui sépare l'espace en deux. Définissons l'opérateur de projection orthogonale sur $\mathcal{H}_0$, $\Pi_X = X[X^T X]^{-1} X^T$. Aussi, la prévision que l'on peut faire pour y est :

$$\hat{y} = X[X^T X]^{-1} X^T y = \Pi_X y \tag{4}$$

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

Le théorème de Pythagore s'écrit ici :

$$\|y\|^2 = \|\Pi_X y\|^2 + \|\Pi_{X^\perp} y\|^2 = \|\Pi_X y\|^2 + \|y - \Pi_X y\|^2 = \|\hat{y}\|^2 + \|\hat{\varepsilon}\|^2$$

qui se traduit classiquement en terme de somme de carrés :

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

Le coefficient de détermination, R^2 s'interprète alors comme le carré du cosinus de l'angle θ entre y et $\Pi_X y$:

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

Une application importante a été obtenue par Frish & Waugh (1933), lorsque l'on partitionne les variables explicatives en deux groupes, $X = [X_1 | X_2]$, de telle sorte que la régression devient :

$$y = \beta_0 + X_1 \beta_1 + X_2 \beta_2 + \varepsilon$$

Frish & Waugh (1933) ont montré qu'on pouvait considérer deux projections successives. En effet, si $y_2^\star = \Pi_{X_1^\perp} y$ et $X_2^\star = \Pi_{X_1^\perp} X_2$, on peut montrer que :

$$\hat{\beta}_2 = [X_2^{\star T} X_2^\star]^{-1} X_2^{\star T} y_2^\star$$

Autrement dit, l'estimation globale est équivalente à l'estimation indépendante des deux modèles si $X_2^\star = X_2$, c'est à dire $X_2 \in \mathcal{E}_{X_1}^\perp$, que l'on peut noter $x_1 \perp x_2$. 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 $y_i = \beta_0 + {x_1}^T \beta_1 + {x_2}^T \beta_2 + \varepsilon_i$, mais le modèle estimé est $y_i = \beta_0 + {x_1}^T b_1 + \eta_i$. L'estimateur du maximum de vraisemblance de b_1 est :

$$\begin{aligned}
\hat{b}_1 &= (X_1^T X_1)^{-1} X_1^T y \\
&= (X_1^T X_1)^{-1} X_1^T [X_{1,i} \beta_1 + X_{2,i} \beta_2 + \varepsilon] \\
&= (X_1^T X_1)^{-1} X_1^T X_1 \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T \varepsilon \\
&= \beta_1 + \underbrace{(X_1' X_1)^{-1} X_1^T X_2 \beta_2}_{\beta_{12}} + \underbrace{(X_1^T X_1)^{-1} X_1^T \varepsilon}_{\nu_i}
\end{aligned}$$

de telle sorte que $\mathbb{E}[\hat{b}_1] = \beta_1 + \beta_{12}$, le biais étant nul uniquement dans le cas où $X_1^T X_2 = 0$ (c'est-à-dire $X_1 \perp X_2$): on retrouve ici une conséquence du théorème de Frisch-Waugh. En revanche, la sur-identification correspond au cas où le vrai modèle serait $y_i = \beta_0 + {x_1}^T \beta_1 + \varepsilon_i$, mais le modèle estimé est $y_i = \beta_0 + {x_1}^T b_1 + {\mathrm{x}_2}^T b_2 + \eta_i$. Dans ce cas, l'estimation est sans biais, au sens où $\mathbb{E}(\hat{b}_1) = \beta_1$ 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 $\hat{b}_2$ 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).

Du paramétrique au non-paramétrique

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

$$\hat{y} = X\,\hat{\beta} = X[X^T X]^{-1} X^T y = \Pi_X y$$

permet de voir la prévision directement comme une transformation linéaire des observations. De manière plus générale, on peut obtenir un prédicteur linéaire en considérant $m(\mathrm{x}) = {s_x}^T y$, où s_x est un vecteur de poids, qui dépend de x, interprété comme un vecteur de lissage. En utilisant les vecteurs $\mathrm{s}_{\mathrm{x}_i}$, calculés à partir des x_i, on obtient une matrice S de taille $n \times n$, et $\hat{y} = Sy$. Dans le cas de la régression linéaire décrite auparavant, $s_x = X[X^T X]^{-1}\mathrm{x}$, $S = X[X^T X]^{-1} X$ et classiquement, $\mathrm{trace}(S)$ est le nombre de colonnes de la matrice X (le nombre de variables explicatives). Dans ce contexte de prédicteurs linéaires, $\mathrm{trace}(S)$ est souvent vu comme un équivalent au nombre de paramètres (ou complexité, ou dimension, du modèle), et $\nu = n - trace(S)$ est alors le nombre de degrés de liberté (Ruppert *et al.*, 2003 ; Simonoff, 1996). Le principe de parcimonie[2] consiste à minimiser cette dimension (la trace de la matrice 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 :

$$\hat{m}_h(x) = {s_x}^T y = \sum_{i=1}^{n} s_{x,i}\, y_i \text{ avec } s_{x,i} = \frac{K_h(x - x_i)}{K_h(x - x_1) + \cdots + K_h(x - x_n)}$$

où $K(\cdot)$ est une fonction noyau, qui attribue une valeur d'autant plus faible que x_i est proche de x, et $h > 0$ est la fenêtre de lissage. L'introduction de ce méta-paramètre h pose un soucis, car il convient de le choisir judicieusement. En faisant des développement limités, on peut montrer que si X a pour densité f :

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

et $\mathrm{Var}[\hat{m}_h(x)] \sim \dfrac{C_3}{nh} \dfrac{\sigma(x)}{f(x)}$

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 C1-I (où le méta-paramètre est ici h^{-1}). L'idée naturelle est alors de chercher à minimiser l'erreur quadratique moyenne, le MSE, $biais[\hat{m}_h(x)]^2 + \mathrm{Var}[\hat{m}_h(x)]$, ce qui donne une valeur optimale pour h de la forme $h^\star = O(n^{-1/5})$, et rappelle la règle de Silverman (1986). En plus grande dimension, pour des variables x continues, on peut utiliser un noyau multivarié, de fenêtre matricielle H :

$$\mathbb{E}[\hat{m}_H(x)] \sim m(x) + \frac{C_1}{2} trace(\mathrm{H}^T m''(\mathrm{x})\mathrm{H}) + C_2 \frac{m'(\mathrm{x})^T H H^T \nabla f(\mathrm{x})}{f(\mathrm{x})}$$

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

Et : $\text{Var}[\hat{m}_H(x)] \sim \frac{C_3}{n \det(H)} \frac{\sigma(x)}{f(x)}$

Si H est une matrice diagonale, avec le même terme h sur la diagonale, alors $h^\star = O(n^{-1/(4+\dim(\mathbf{x}))})$]. 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 :

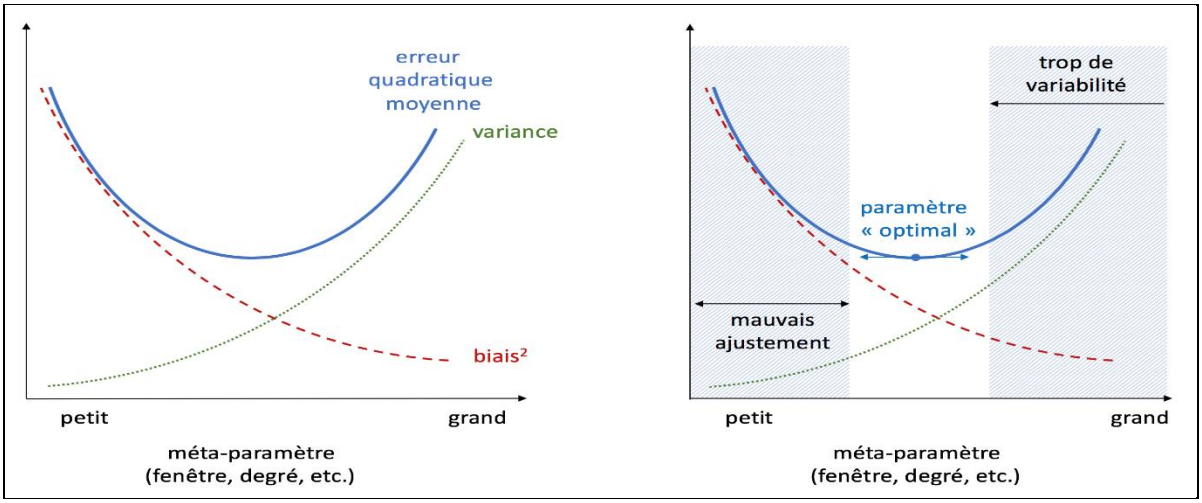
$$MISE[\hat{m}_h] \sim \overbrace{\frac{h^4}{4}\left(\int x^2 k(x)dx\right)^2 \int [m''(x) + 2m'(x)\frac{f\prime(x)}{f(x)}]^2 dx}^{biais^2} + \overbrace{\frac{\sigma^2}{nh}\int k^2(x)dx \cdot \int \frac{dx}{f(x)}}^{variance},$$

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

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

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

Figure C1-I
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)



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 :

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right) \text{ avec } \theta_i = \psi(x_i^T\beta).$$

Les fonctions a, b et c sont spécifiées en fonction du type de loi exponentielle (étudiée abondamment en statistique depuis Darmoix (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 :

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

et la condition du premier ordre s'écrit alors :

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

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

$$\hat{y} = m(\mathrm{x}) = \mathbb{E}[Y|X=x] = g^{-1}(x^T\beta)$$

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 W dépend de β (avec $W = \mathrm{diag}(\nabla g(\hat{y})\mathrm{Var}[y])$ 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 :

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

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

$$\hat{\beta}_{k+1} = [X^T W_k^{-1} X]^{-1} X^T W_k^{-1} z_k$$

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

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

où numériquement $I(\beta) = \phi\cdot[X^T W_{\infty}^{-1} X]$.

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.

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 (conditionnelle) suit un modèle linéaire :

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

ou encore :

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

correspondant à la fonction de répartition de la loi logistique. L'estimation de (β_0,β) se fait par maximisation de la vraisemblance :

$$\mathcal{L}=\prod_{i=1}^{n}\left(\frac{e^{\beta_0+x_i^T\beta}}{1+e^{\beta_0+x_i^T\beta}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+x_i^T\beta}}\right)^{1-y_i}$$

On continuera à parler des modèles linéaires car les courbes d'isoprobabilités sont ici les hyperplans parallèles $b_0+x^T\beta$. À ce modèle, popularisé par Berkson (1944), certains préféreront le modèle probit (Berkson, 1951), introduit par Bliss (1934). Dans ce modèle :

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

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

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

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 $y_i=1(y_i^\star>\xi)$ (où ξ sera à fixer). On le voit, ces techniques sont fondamentalement liées à un modèle stochastique sous-jacent. Dans le corps de l'article, nous présentons plusieurs techniques alternatives - tirées de la littérature en apprentissage - pour ce problème de classification (avec deux classes, ici 0 et 1).

Régression en grande dimension

Comme nous l'avions mentionné auparavant, la condition du premier ordre $X^T(X\hat{\beta}-\mathrm{y})=0$ se résout numériquement en effectuant une décomposition QR, pour un coût algorithmique en $\mathrm{O}(np^2)$ opérations (où p est le rang de X^TX). Numériquement, ce calcul peut être long (soit parce que p est grand, soit – dans une moindre mesure – 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 $\{1,\cdots,n\}$. Alors $\hat{\beta}_s=({X_s}^TX_s)^{-1}{X_s}^Ty_s$ est une bonne approximation de $\hat{\beta}$ comme le montre Dhillon *et al.* (2014). Cet algorithme est toutefois dangereux si certains points ont un pouvoir de levier important (i.e. $L_i=x_i(X^TX)^{-1}{x_i}^T$). 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 machine[3] 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 algorithme il y a donc autant d'itération que d'observations. Contrairement à l'algorithme de descente de gradient (ou

[3] Comme, par exemple, celles du langage Python.

méthode de Newton) à chaque itération un seul vecteur de gradient est calculé (et non plus n). Il est néanmoins parfois nécessaire d'exécuter cet 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 $m_\beta(x)$ et y l'algorithme peut se résumer ainsi :
- Étape 0 : Mélange des données
- Étape d'itérations t : Pour $t = 1, \dots, n$, on tire $i \in \{1, \cdots, n\}$ sans remise, on pose :

$$\beta^{t+1} = \beta^t - \gamma_t \frac{\partial l(y_i, m_{\beta^t}(x_i))}{\partial \beta}$$

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).

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 :

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}_{variance\ totale} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{variance\ résiduelle} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{variance\ expliquée} \tag{5}$$

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

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

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 (à une constante près) par :

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

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

$$R^2 = \frac{Déviance(\overline{y}) - Déviance(\hat{y})}{Déviance(\overline{y})} = 1 - \frac{Déviance(\hat{y})}{Déviance(\overline{y})}$$

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

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

où p est le nombre de paramètres du modèle. La mesure de la qualité de l'ajustement va pénaliser les modèles trop complexes. Cette idée va se retrouver dans le critère d'Akaike, où $AIC = Déviance + 2 \cdot p$ ou dans le critère de

Schwarz, $BIC = Déviance + \log(n) \cdot p$. En grande dimension (typiquement $p > \sqrt{n}$), on aura tendance à utiliser un AIC corrigé, défini par :

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

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 :

$$\mathcal{M} = \{m\colon\ m(\mathrm{x}) = s_h(x)^T y \text{ où } S = (s(x_1), \cdots, s(x_n))^T \text{ est la matrice de lissage}\}$$

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

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

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 :

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

(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:

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

On peut montrer que :

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

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

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

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

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'expliciter la relation liant y aux variables 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 le corps de

l'article). 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.

Apprentissage par une machine

Aujourd'hui, on parle d'« apprentissage automatique » pour décrire tout un ensemble de techniques, souvent computationnelles, alternatives à l'approche de 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 automatique 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 cœur, 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 $y \in \{0,1,2,\cdots,8,9\}$. 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, x_i) avec $x_i \in \mathbb{R}^p$. Considérons une distance Δ sur $\mathbb{R}^p$ (la distance Euclidienne ou la distance de Mahalanobis, par exemple). Étant donnée une nouvelle observation $x \in \mathbb{R}^p$, supposons les observations ordonnées en fonction de la distance entre les x_i et x, au sens où $\Delta(x_1, x) \leq \Delta(x_2, x) \leq \cdots \leq \Delta(x_n, x)$ alors on peut considérer comme prédiction pour y la moyenne des k plus proches voisins :

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

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 (page 86) qui surprendra et fera sourire les économètres : « *logistic regression can also be interpreted from a probabilistic perspective* ». 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).

Le tournant des années 1980/1990 et le formalisme probabiliste

On dispose d'un échantillon d'apprentissage, avec des observations (x_i, y_i) où les 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[4]. 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 $\mathcal{R}_{\mathbb{P},f}(m) = \mathbb{P}[m(X) \neq f(X)]$ où $\mathrm{X} \sim \mathbb{P}$, ou écrit de manière équivalente, $\mathcal{R}_{\mathbb{P},f}(m) = \mathbb{P}[\{x \in \mathcal{X} : m(\mathrm{x}) \neq f(\mathrm{x})\}]$. 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 x_i sont obtenus par des tirages indépendants suivant $\mathbb{P}$, et qu'ensuite $y_i = f(\mathrm{x}_i)$.

On peut ici définir le risque empirique d'un modèle m, $\hat{\mathcal{R}}(m) = \frac{1}{n}\sum_{i=1}^{n} 1(m(\mathrm{x}_i) \neq y_i)$.

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 $\mathcal{X}$e contenant deux labels $\{x_1, x_2\}$ et que $\mathbb{P}$ soit telle que $\mathbb{P}(\{x_1\}) = p$ et $\mathbb{P}(\{x_2\}) = 1 - p$. Il est possible 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. Mais on peut aussi noter qu'il est possible de ne jamais observer x_2 (par exemple) parmi n tirages suivant $\mathbb{P}$ (la probabilité est p^n). Il sera alors impossible de trouver $m(x_2)$. 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.

Pour illustrer les liens, supposons, comme dans Wolpert (1996) et Wolpert & Macready (1997) que m appartienne une classe particulière, notée $\mathcal{M}$, contenant 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 $1 - \delta$, $\mathcal{R}_{\mathbb{P},f}(m^\star) \leq \epsilon$ en notant :

$$m^\star \in \underset{m \in \mathcal{M}}{argmin}\{\frac{1}{n}\sum_{i=1}^{n} 1(m(\mathrm{x}_i) \neq y_i)\}$$

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

On 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 $n_{\mathcal{M}} : [0,1]^2 \to \mathbb{N}$ 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 x_i suivant $\mathbb{P}$, étiquetés y_i par f, alors il existe $m \in \mathcal{M}$ tel que, avec

[4] La notation $\{-1, +1\}$ sera ici privilégiée à celle utilisée en économétrie $\{0,1\}$, en lien avec la loi de Bernoulli, et correspondant aux bornes d'une fonction de probabilité.

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 $\{x_1,\cdots x_k\}$, et considérons l'ensemble $\mathcal{E}_k = \{(m(x_1),\cdots,m(x_k)) \text{ pour } m \in \mathcal{M})\}$. 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 $VC(\mathcal{M}) = \sup\{k \text{ tel que } \mathcal{M} \text{ pulvérise } \{x_1,\cdots x_k\}\}$. Par exemple si $\mathcal{X} = \mathbb{R}^k$, et considérons des séparations par des hyperplans passant par l'origine (on dira homogènes), au sens où $m_{\mathrm{w}}(x) = 1_{\pm}(w^T x \geq 0)$. 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 $m_{\mathrm{w},b}(x) = 1_{\pm}(w^T x + b \geq 0)$, 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'apprentisssage : si $\mathcal{M}$ est une classe de dimension $d = VC(\mathcal{M})$, alors il existe des constante 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 Vailiant, 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 rapproché 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.

Le choix de l'objectif et la fonction de perte

On l'a vu, l'apprentissage (automatique, par une machine) se fait en résolvant des problèmes d'optimisation. Et les choix de l'objectif (et de la fonction de perte) sont essentiels, 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 $\{(y_i, \mathrm{x}_i)\}$, 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 :

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

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

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

avec, dans le cas du perceptron :

$$\beta_j^{(k+1)} = \beta_j^{(k)} + \eta(\mathrm{y} - \hat{y}^{(k)})^T x_j$$

Ici $\ell(y, y') = 1(y \neq y')$ est une fonction de perte, qui permettra de donner un prix à une erreur commise, en prédisant $y' = m(\mathrm{x})$ et en observant y. Pour un problème de régression, on peut considérer une erreur quadratique ℓ_2, telle que $\ell(y, m(\mathrm{x})) = (y - m(\mathrm{x}))^2$ ou en valeur absolue ℓ_1, avec $\ell(y, m(\mathrm{x})) = |y - m(x)|$. Ici, pour notre problème de classification, nous utilisions 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 :

$$m^\star(\mathrm{x}) = \underset{m \in \mathcal{M}}{argmin}\left\{\sum_{i=1}^{n} \ell\left(y_i, m(\mathrm{x}_i)\right)\right\} \tag{6}$$

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). On pourra noter $\mathcal{M}_0$ l'espace des hyperplans de $\mathbb{R}^p$ au sens où :

$$m \in \mathcal{M}_0 \text{ signifie } m(\mathrm{x}) = \beta_0 + \beta^T \mathrm{x} \text{ avec } \beta \in \mathbb{R}^p$$

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 $\hat{m}^\star$, dit « oracle », dans une famille d'estimateurs $\mathcal{M}$:

$$\hat{m}^\star = \underset{\hat{m} \in \mathcal{M}}{argmin}\{\mathcal{R}(\hat{m}, m)\}$$

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 $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, symmétrique, qui vérifie l'inégalité triangulaire, et telle que $\ell(\mathrm{x}, \mathrm{y}) = 0$ si et seulement si $\mathrm{x} = \mathrm{y}$. La norme associée est $\|\cdot\|$, telle que $\ell(\mathrm{x}, \mathrm{y}) = \|\mathrm{x} - \mathrm{y}\| = \ell(\mathrm{x} - \mathrm{y}, 0)$ (en utilisant le fait que $\ell(\mathrm{x}, \mathrm{y} + \mathrm{z}) = \ell(\mathrm{x} - \mathrm{y}, \mathrm{z})$ - 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 :

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

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 :

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

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 $x \mapsto \mathbb{E}[Y|X = x]$. Une autre fonction de perte particulièrement intéressante est la perte ℓ_1, $\ell_1(y, m) = |y - m|$. Rappelons que :

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

Le problème d'optimisation

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

est obtenu en économétrie si on suppose que la loi conditionnelle de Y suit une loi de Laplace centrée sur $m(x)$, et en maximisant la (log) vraisemblance (la somme des valeurs absolues des erreurs correspond à la log-vraisenblance 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
$\ell_1(y,m) = |(y-m)(1/2 - 1_{y\leq m})|$, on peut obtenir une généralisation pour $\tau \in (0,1)$:

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

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

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

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

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

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 assymétrique.

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

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

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

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 y, à partir de X, qui donnera une erreur ε_1. À la seconde étape, on estime un modèle m_2 pour ε_1, à partir de X, qui donnera une erreur ε_2, etc. On va alors retenir comme modèle, au bout de k itérations :

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

Ici, l'erreur ε est vue comme la différence entre y et le modèle $m(x)$, 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. Le problème va alors se réécrire comme un problème d'optimisation :

$$m^{(k)} = m^{(k-1)} + \underset{h\in\mathcal{H}}{argmin}\left\{\sum_{i=1}^{n} \ell\left(y_i - m^{(k-1)}(x_i), h(x_i)\right)\right\} \tag{8}$$

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, $\varepsilon_1 = y - m_1(x)$, on posera $\varepsilon_1 = y - \alpha \cdot m_1(x)$ 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 $\hat{\varepsilon} = y - X\hat{\beta}$ sont orthogonaux aux variables explicatives, 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 :

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

puis on considère l'apprentissage suivant :

$$m^{(k)} = m^{(k-1)} + \underset{h\in\mathcal{H}}{argmin}\sum_{i=1}^{n} \ell\left(y_i, m^{(k-1)}(x_i) + h(x_i)\right) \tag{9}$$

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

$$m^{(k)} = m^{(k-1)} - \gamma_k \sum_{i=1}^{n} \nabla_{m^{(k-1)}} \ell\left(y_i, m^{(k-1)}(x_i)\right), \tag{10}$$

où :

$$\gamma_k = \underset{\gamma}{argmin}\sum_{i=1}^{n} \ell\left(y_i, m^{(k-1)}(x_i) - \gamma\nabla_{m^{(k-1)}}\ell(y_i, m^{(k-1)}(x_i))\right)$$

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

$$r_{i,k} = -\frac{\partial\ell(y_i, m(x_i))}{\partial m(x_i)}\Big|_{m(\mathbf{x})=m^{(k-1)}(x)} \text{ pour } i = 1, \cdots, n.$$

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

$$\gamma_k = \underset{\gamma\in\mathbb{R}}{argmin}\left\{\sum_{i=1}^{n} \ell\left(y_i, m^{(k-1)}(x_i) + \gamma h^\star(x_i)\right)\right\}$$

puis on met à jour le modèle en posant $m_k(x) = m_{k-1}(x) + \gamma_k h^\star(x)$. 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 un 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 $m_k(x) = m_{k-1}(x) + \nu\gamma_k h^\star(x)$, 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.

Pénalisation et choix de variables

Dans la section sur l'économétrie, nous avions évoqué le principe de parcimonie. 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 :

$$\left(\hat{\beta}_{0,\lambda}, \hat{\beta}_\lambda\right) = argmin\left\{\sum_{i=1}^{n} \ell\left(y_i, \beta_0 + x^T\beta\right) + \lambda\, pénalisation(\beta)\right\} \tag{11}$$

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 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* »)[5]. 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 :

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

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. On parlera de « sparsité ». Soit s le nombre de covariables (réellement) pertinentes, $s = \#\mathcal{S}$ avec $\mathcal{S} = \{j = 1, \cdots, p; \beta_j \neq 0\}$. Si on note X_S la matrice constituée des variables pertinentes (en colonnes), alors on suppose que le vrai modèle est de la forme $y = {x_S}^T\beta_S + \varepsilon$. Intuitivement, un estimateur intéressant serait alors $\hat{\beta}_S = [{x_S}^T X_S]^{-1} X_S y$, 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 :

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

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

[5] Le terme « lasso » peut être vu comme une référence à Breiman (1995), qui avait proposé une méthode de « garrottage » (en anglais *garrote*) pour faire de la sélection de variables.

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

Tableau C 1-1
Optimisation contrainte et régularisation

	Optimisation contrainte	Pénalisation	
Meilleur groupe	$\underset{\beta;\|\beta\|_{\ell_0}\leq s}{argmin}\left\{\sum_{i=1}^{n} \ell\,(y_i, \beta_0 + x^T\beta)\right\}$	$\underset{\beta,\lambda}{argmin}\left\{\sum_{i=1}^{n} \ell\,(y_i, \beta_0 + x^T\beta) + \lambda\|\beta\|_{\ell_0}\right\}$	$(\ell 0)$
Lasso	$\underset{\beta;\|\beta\|_{\ell_1}\leq s}{argmin}\left\{\sum_{i=1}^{n} \ell\,(y_i, \beta_0 + x^T\beta)\right\}$	$\underset{\beta,\lambda}{argmin}\left\{\sum_{i=1}^{n} \ell\,(y_i, \beta_0 + x^T\beta) + \lambda\|\beta\|_{\ell_1}\right\}$	$(\ell 1)$
Ridge	$\underset{\beta;\|\beta\|_{\ell_2}\leq s}{argmin}\left\{\sum_{i=1}^{n} \ell\,(y_i, \beta_0 + x^T\beta)\right\}$	$\underset{\beta,\lambda}{argmin}\left\{\sum_{i=1}^{n} \ell\,(y_i, \beta_0 + x^T\beta) + \lambda\|\beta\|_{\ell_2}\right\}$	$(\ell 2)$

Considérons les problèmes d'optimisation du Tableau C1-1. Si on considère le problème classique où ℓ est la norme quadratique, les deux problèmes de l'équation $(\ell 1)$ du Tableau C1-1 sont équivalents, au sens où, pour toute solution $(\beta^\star, s)$ au problème de gauche, il existe $\lambda^\star$ tel que $(\beta^\star, \lambda^\star)$ soit solution du problème de droite; et inversement. Le résultat est également vrai pour les problèmes $(\ell 2)$[6]. Il s'agit en effet de problèmes convexes. En revanche, les deux problèmes $(\ell 0)$ ne sont pas équivalents : si pour $(\beta^\star, \lambda^\star)$ solution du problème de droite, il existe $s^\star$ tel que $\beta^\star$ 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
$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2) \leq \underbrace{\mathbb{E}({x_S}^T\hat{\beta}_S - x^T\beta_0]^2)}_{=\sigma^2 \#S} \cdot (4\log p + 2 + o(1))$.
Notons que dans ce cas :
$$\hat{\beta}_{\lambda,j}^{sub} = \begin{cases} 0 \text{ si } j \notin \mathcal{S}_\lambda \\ \hat{\beta}_j^{ols} \text{ si } j \in \mathcal{S}_\lambda, \end{cases}$$
où $\mathcal{S}_\lambda$ désigne l'ensemble des coordonnées non nulle lors de la résolution de $(\ell 0)$.
Le problème $(\ell 2)$ 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 :

$$\hat{\beta}_\lambda^{ridge} = (X^TX + \lambda\mathbb{I})^{-1}X^Ty = (X^TX + \lambda\mathbb{I})^{-1}(X^TX)\,\hat{\beta}^{ols}$$

Aussi, on peut en déduire que :
$biais[\hat{\beta}_\lambda^{ridge}] = -\lambda[\mathrm{X^TX} + \lambda\mathbb{I}]^{-1}\ \hat{\beta}^{ols}$ et $Var[\hat{\beta}_\lambda^{ridge}] = \sigma^2[X^TX + \lambda\mathbb{I}]^{-1}X^TX[X^TX + \lambda\mathbb{I}]^{-1}$.

Avec une matrice de variables explicatives orthonormées (i.e. $X^TX = \mathbb{I}$), les expressions se simplifient :
$biais[\hat{\beta}_\lambda^{ridge}] = \dfrac{\lambda}{1+\lambda}\,\hat{\beta}^{ols}$ et $Var[\hat{\beta}_\lambda^{ridge}] = \dfrac{\sigma^2}{(1+\lambda)^2}\,\mathbb{I}$.

Notons que $Var[\hat{\beta}_\lambda^{ridge}] < Var[\hat{\beta}^{ols}]$. Et en notant que :

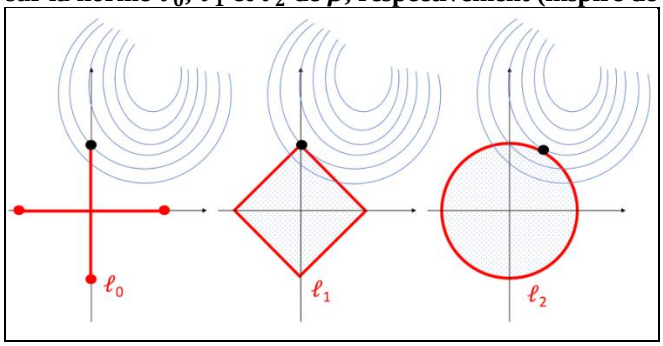
[6] Pour $(\ell 1)$, s'il y a équivalence au niveau théorique, il peut exister des soucis numériques car il n'y a pas forcément unicité de la solution.

$$mse[\hat{\beta}_\lambda^{ridge}] = \frac{k\sigma^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\beta^T\beta$$

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

En revanche, si ℓ n'est plus la norme quadratique mais la norme ℓ_1, le problème ($\ell 1$) n'est pas toujours strictement convexe, et en particulier, l'optimum n'est pas toujours unique (par exemple si X^TX est singulière). Mais si ℓ est strictement convexe, alors $X\hat{\beta}$ sera unique. Notons de plus que deux solutions sont forcément cohérentes en terme de signe des coefficients : il n'est pas possible d'avoir $\hat{\beta}_j < 0$ pour une solution et $\hat{\beta}_j > 0$ pour une autre. D'un point de vue heuristique, le programme ($\ell 1$) 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 ($\ell 0$) - comme le montre de manière visuelle la figure C1-II.

Figure C1-II
Pénalisation basée sur la norme ℓ_0, ℓ_1 et ℓ_2 de β, respectivement (inspiré de Hastie *et al.*, 2016)



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'ensembles des β_j non nuls. Moyennant quelques hypothèses techniques supplémentaires, on peut montrer que cet estimateur est « sparsistant » au sens où le support de $\hat{\beta}_\lambda^{lasso}$ 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 $\hat{\beta}_\lambda^{lasso}$ 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.

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 $\hat{\beta}_{\lambda_1}$ et $\hat{\beta}_{\lambda_2}$ 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. Précédemment, 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 $y_i = x_i{}^T\beta_0 + \varepsilon_i$ où les ε_i sont des réalisations de lois indépendantes et centrées. Le risque quadratique empirique *in-sample* est ici :

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left([x_i{}^T\hat{\beta} - x_i{}^T\beta_0]^2\right) = \mathbb{E}([x_i{}^T\hat{\beta} - x_i{}^T\beta_0]^2),$$

pour n'importe quelle observation i. En supposant les résidus ε Gaussiens, alors on peut montrer que ce risque vaut $\sigma^2\text{trace}(\Pi_X)/m$ soit $\sigma^2 p/m$. En revanche le risque quadratique empirique *out-of-sample* est ici :
$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2)$
où x est une nouvelle observation, indépendante des autres. On peut noter que :
$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2|x) = \sigma^2 x^T(X^TX)^{-1}x$
et en intégrant par rapport à x,
$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2) = \mathbb{E}(\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2|x)) = \sigma^2 trace(\mathbb{E}[xx^T]\mathbb{E}[(X^TX)^{-1}])$.

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

$$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2) \geq \sigma^2\frac{p}{m}$$

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

$$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2) = \sigma^2\frac{p}{m-p-1}.$$

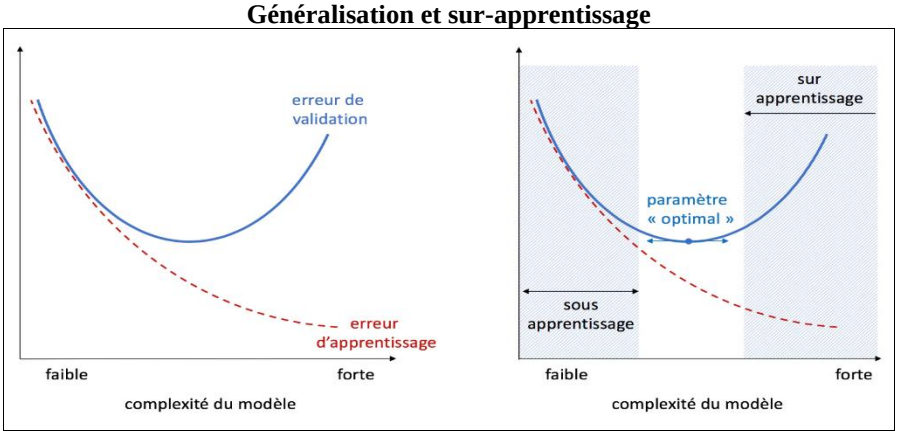
Si on regarde maintenant la version empirique : si $\hat{\beta}$ est estimé sur les m premières observations,

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

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 C1-IV 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, en bas). 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 C1-III, 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. Cette figure n'est pas sans rappeler la Figure C1-I.

Figure C1-III
Généralisation et sur-apprentissage



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) relativement utilisé en statistique pour réduire le biais. En effet, si on suppose que $\{y_1, \cdots, y_n\}$ est un échantillon tiré suivant une loi F_θ, et que l'on dispose d'un estimateur $T_n(y) = T_n(y_1, \cdots, y_n)$, mais que cet estimateur est biaisé, avec $\mathbb{E}[T_n(Y)] = \theta + O(n^{-1})$, il est possible de réduire le biais en considérant :

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

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

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 :

$$\hat{\mathcal{R}}^{\,\mathrm{CV}} = \frac{1}{n}\sum_{i=1}^{n} \ell\,(y_i, \hat{m}_{(i)}\,(x_i))$$

On parlera ici de méthode « *leave-one-out* » (loocv).

Cette technique n'est pas sans faire penser à celle souvent retenue 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 :

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

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

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

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 boostrappés. Soit $\mathcal{I}_b$ un échantillon de taille n obtenu en tirant avec remise dans $\{1, \cdots, n\}$ pour savoir quelles observations (y_i, x_i) seront gardées dans la population d'apprentissage (à chaque tirage). Notons $\mathcal{I}_{\overline{b}} = \{1, \cdots, n\} \setminus \mathcal{I}_b$. En notant $\hat{m}_{(b)}$ construit sur l'échantillon $\mathcal{I}_b$, on pose alors :

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

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

Bibliographie

Ahamada, I. & Flachaire, E. (2011). *Non-Parametric Econometrics*. Oxford: 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(1), 21–37.
https://doi.org/10.1016/0304-4076(77)90052-5
Aldrich, J. (2010). The Econometricians' Statisticians, 1895-1945. *History of Political Economy*, 42(1), 111–154.
https://doi.org/10.1215/00182702-2009-064
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(3), 505–529.
https://doi.org/10.1016/0378-4266(94)90007-8
Angrist, J. D. & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*, 114(2), 533–575.
https://doi.org/10.1162/003355399556061
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(2), 3–30.
https://doi.org/10.1257/jep.24.2.3
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? *The Quarterly Journal of Economics*, 106(4), 979–1014.
https://doi.org/10.2307/2937954
Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, 177–187.
https://doi.org/10.1007/978-3-7908-2604-3_16
Bajari, P., Nekipelov, D., Ryan, S. P. & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, 105(5), 481–485.
https://doi.org/10.1257/aer.p20151021
Bazen, S. & Charni, K. (2015). Do earnings really decline for older workers? AMSE, *Working Paper* 2015-11.
https://halshs.archives-ouvertes.fr/halshs-01119425
Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.
Belloni, A., Chernozhukov, V. & Hansen, C. (2010). Inference for High-Dimensional Sparse Econometric Models. *Advances in Economics and Econometrics*, 245–295
https://doi.org/10.1017/CBO9781139060035.008
Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica*, 80(6), 2369–2429.
https://doi.org/10.3982/ECTA9626
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(1), 289–300.
https://www.jstor.org/stable/2346101
Berger, J. O. (1985). *Statistical decision theory and Bayesian Analysis* (2nd ed.). New York, Berlin: Springer-Verlag.
Berk, R. A. (2008). *Statistical Learning from a Regression Perspective*. New York: Springer Verlag.
Berkson, J. (1944). Applications of the Logistic Function to Bioassay. *Journal of the American Statistical Association*, 39(227), 357–365.
https://doi.org/10.1080/01621459.1944.10500699
Berkson, J. (1951). Why I Prefer Logits to Probits. *Biometrics*, 7(4), 327–339.
https://doi.org/10.2307/3001655
Bernardo, J. M. & Smith, A. F. M. (2000). *Bayesian Theory*. New York: John Wiley.
Berndt, E. R. (1990). *The Practice of Econometrics: Classic and Contemporary*. Reading, Mass: 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.
http://www3.stat.sinica.edu.tw/statistica/oldpdf/a7n11.pdf
Bishop, C. (2006). *Pattern Recognition and Machine Learning*. New York: 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(1), 356–364.
https://doi.org/10.1016/j.eswa.2012.07.051
Bliss, C. I. (1934). The method of probits. *Science*, 79(2037), 38–39.
https://doi.org/10.1126/science.79.2037.38
Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M.K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4), 929–965.
https://doi.org/10.1145/76359.76371

Breiman, L. Fiedman, J., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Londres: Chapman & Hall/CRC.
Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4), 373–384. https://doi.org/10.1080/00401706.1995.10484371
Breiman, L. (2001a). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3), 199–231. https://doi.org/10.1214/ss/1009213726
Breiman, L. (2001b). Random forests. *Machine learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
Brown, L. D. (1986). *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Hayworth, CA, USA: Institute of Mathematical Statistics.
Bühlmann, P. & van de Geer, S. (2011). *Statistics for High Dimensional Data: Methods, Theory and Applications*. Heidelberg, New York: Springer Verlag.
Candès, E. & Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5), 2145–2177. https://doi.org/10.1214/08-AOS653
Clarke, B. S., Fokoué, E. & Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning*. New York: Springer Verlag.
Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1023/A:1022627411411
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. https://doi.org/10.1109/PGEC.1965.264137
Cover, T. M. & Hart, P. (1965). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964
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(4), 303–314. https://doi.org/10.1007/BF02551274
Darmois, G. (1935). Sur les lois de probabilités à 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 https://doi.org/10.1002/cpa.20042
Davison, A. C. (1997). *Bootstrap*. Cambridge: Cambridge University Press.
Davidson, R. & MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
Davidson, R. & MacKinnon, J. G. (2003). *Econometric Theory and Methods*. Oxford: Oxford University Press.
Duo, Q. (1993). *The Formation of Econometrics*. Oxford: Oxford University Press
Debreu, G. (1986). Theoretic Models: Mathematical Form and Economic Content. *Econometrica*, 54(6), 1259–1270. https://doi.org/10.2307/1914299
Dhillon, P., Lu, Y. Foster, D. P. & Ungar, L. H. (2014). New Subsampling Algorithms for Fast Least Squares Regression. In: Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q. (Eds.), *Advances in Neural Information Processing Systems 26*. New York: Curran Associates. http://papers.nips.cc/paper/5105-new-subsampling-algorithms-for-fast-least-squares-regression.pdf
Efron, B. & Tibshirani, R. (1993). *Bootstrap*. Londre : 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(358), 314–329. https://doi.org/10.2307/2231790
Flach, P. (2012). *Machine Learning*. Cambridge: 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. https://doi.org/10.1214/aos/1176325766
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*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.374.7489&rep=rep1&type=pdf
Frisch, R. & Waugh, F.V. (1933). Partial Time Regressions as Compared with Individual Trends. *Econometrica*. 1(4), 387–401. https://doi.org/10.2307/1907330
Galton, Edgeworth, Frish, and prospects for quantile regression in Econometrics (1998). Conference on Principles of Econometrics, Madison.
Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494), 746–762. https://doi.org/10.1198/jasa.2011.r10138
Givord, P. (2010). Méthodes économétriques pour l'évaluation de politiques publiques. *Economie & Prévision*, 204-205, 1–28.

https://www.cairn.info/revue-economie-et-prevision-2014-1-page-1.htm
Grandvalet, Y., Mariéthoz, J., & Bengio, S. (2005). Interpretation of SVMs with an application to unbalanced classification. *Advances in Neural Information Processing Systems*, 18.
https://papers.nips.cc/paper/2763-a-probabilistic-interpretation-of-svms-with-an-application-to-unbalanced-classification
Groves, T. & Rothenberg, T. (1969). A note on the expected value of an inverse matrix. *Biometrika*, 56(3), 690–691.
https://doi.org/10.1093/biomet/56.3.690
Haavelmo, T. (1944). The probability approach in econometrics, *Econometrica*, 12, iii–vi, 1–115.
https://doi.org/10.2307/1906935
Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. Londres: Chapman & Hall/CRC.
Hastie, T., Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning. New York: Springer Verlag.
Hastie, T., Tibshirani, W. & Wainwright, M. (2015). *Statistical Learning with Sparsity*. Londres: Chapman CRC.
Hastie, T., Tibshiriani, R. & Tibshiriani, R. J. (2016). Extended Comparisons of Best Subset Selection, Forward Stepwise Selection and the Lasso.
https://arxiv.org/abs/1707.08692
Haultefœuille, X. (d') & Givord, P. (2014). La régression quantile en pratique. *Économie & Statistiques*, 471, 85–111.
https://www.persee.fr/doc/estat_0336-1454_2014_num_471_1_10484
Hebb, D. O. (1949). The organization of behavior. New York: Wiley.
Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1), 153–161.
https://doi.org/10.2307/1912352
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(3), 748–755.
https://doi.org/10.1162/003465303322369867
Hendry, D F. & Krolzig, H.-M. (2001). Automatic Econometric Model Selection. London: Timberlake Press.
Herbrich, R., Keilbach, M., Graepel, T. Bollmann-Sdorra, P. & Obermayer, K. (1999). Neural Networks in Economics. In: Brenner, T. (Ed.), *Computational Techniques for Modelling Learning in Economics*, pp. 169–196. Boston, MA: Springer Verlag.
https://doi.org/10.1007/978-1-4615-5029-7_7
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.
Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.
https://doi.org/10.1080/01621459.1986.10478354
Hyndman, R., Koehler, A. B., Ord, J. K. & Snyder, R. D. (2009). Forecasting with Exponential Smoothing. Springer Verlag.
James, G., D. Witten, T. Hastie, & R. Tibshirani (2013). *An introduction to Statistical Learning*. Springer Series in Statistics.
Khashman, A. (2011). Credit risk evaluation using neural networks: Emotional versus conven- tional models. *Applied Soft Computing*, 11(8), 5477–5484.
https://doi.org/10.1016/j.asoc.2011.05.011
Kean, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, 156(1), 3–20.
https://doi.org/10.1016/j.jeconom.2009.09.003
Kleiner, A., Talwalkar, A., Sarkar, P. & Jordan, M. (2012). *The Big Data Bootstrap*.
https://arxiv.org/abs/1206.6415
Koch, I. (2013). *Analysis of Multivariate and High-Dimensional Data*. Cambridge: Cambridge University Press.
Koenker, R. (2003). *Quantile Regression*. Cambridge: 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(448), 1296–1309.
https://doi.org/10.1080/01621459.1999.10473882
Kolda, T. G. & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3), 455–500.
https://doi.org/10.1137/07070111X
Koopmans, T. C. (1957). *Three Essays on the State of Economic Science*. New York: 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(1), 159–174.
https://doi.org/10.2307/2529310
LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep Learning. *Nature*, 521, 436–444.
https://doi.org/10.1038/nature14539
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.
https://doi.org/10.3150/08-BEJ127
Lemieux, T. (2006). The « Mincer Equation » Thirty Years After Schooling, Experience, and Earnings. In: Grossbard, S. (Ed.), *Jacob Mincer A Pioneer of Modern Labor Economics*, pp. 127–145; Springer Verlag.
https://doi.org/10.1007/0-387-29175-X_11
Li, J. & J. S. Racine (2006). *Nonparametric Econometrics*. Princeton: 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.
https://doi.org/10.5705/ss.202017.0034
Lin, H. W., Tegmark, M. & Rolnick, D. (2016). Why does deep and cheap learning work so well?
https://arxiv.org/abs/1608.08225
Lucas, R. E. (1976). Econometric Policy Evaluation: A Critique. *Carnegie-Rochester Conference Series on Public Policy*, 19–46.
https://doi.org/10.1016/S0167-2231(76)80003-6
Mallows, C.L. (1973). Some Comments on C_p. *Technometrics*, 15(4), 661–675.
https://doi.org/10.2307/1267380
McCullogh, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4), 115–133.
https://doi.org/10.1007/BF02478259
Mincer, J. (1974). *Schooling, experience and earnings*. New York: Columbia University Press.
Mitchell, T. (1997). *Machine Learning*. New York: 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(302), 415–434.
https://doi.org/10.1080/01621459.1963.10500855
Morgan, M. S. (1990). *The history of econometric ideas*. Cambridge: Cambridge University Press.
Mohri, M., Rostamizadeh, A. & Talwalker, A. (2012). Foundations of Machine Learning. Cambridge, Mass: MIT Press.
Mullainathan, S. & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.
https://doi.org/10.1257/jep.31.2.87
Müller, M. (2011). Generalized Linear Models In*:* Gentle, J. E., Härdle, W. K. & Mori, Y. (Eds.), *Handbook of Computational* Statistics. Springer Verlag.
Murphy, K. R. (2012). *Machine Learning: a Probabilistic Perspective*. Cambridge, Mass: MIT Press.
Murphy, K. M. & Welch, F. (1990). Empirical Age-Earnings Profiles. *Journal of Labor Economics*, 8(2), 202–229.
https://doi.org/10.1086/298220
Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and its Applications*, 9(1), 141–2.
https://doi.org/10.1137/1109020
Natarajan, B. K. (1995). Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing* (SICOMP), 24(2), 227–234.
https://doi.org/10.1137/S0097539792240406
Nevo, A. & Whinston, M. D. (2010). Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference. *Journal of Economic Perspective*, 24(2), 69–82.
https://doi.org/10.1257/jep.24.2.69
Neyman, J. (1923). Sur les applications de la théorie des probabilités aux expériences agricoles : Essai des principes. Mémoire de master, republié dans *Statistical Science*, 5(4), 463–472.
https://doi.org/10.1214/ss/1177012031
Nisbet, R., Elder, J. & Miner, G. (2011). *Handbook of Statistical Analysis and Data Mining Applications*. New York: Academic Press.
Okun, A. (1962). Potential GNP: Its measurement and significance. *Proceedings of the Business and Economics Section of the American Statistical Association*, 98–103.
https://milescorak.files.wordpress.com/2016/01/okun-potential-gnp-its-measurement-and-significance-p0190.pdf
Orcutt, G. H. (1952). Toward a partial redirection of econometrics. *Review of Economics and Statistics*, 34(3), 195–213.
https://doi.org/10.2307/1925626
Pagan, A. & Ullah, A. (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(11), 559—572.
https://doi.org/10.1080/14786440109462720
Platt, J. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, 10, 61–74.
http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639
Portnoy, S. (1988). Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity. *Annals of Statistics*, 16(1), 356–366.
https://doi.org/10.1214/aos/1176350710
Quenouille, M. H. (1949). Problems in Plane Sampling. *The Annals of Mathematical Statistics*, 20(3), 355–375.
https://doi.org/10.2307/2332914
Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika*, 43(3-4), 353–360.
https://doi.org/10.2307/2332914
Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.
https://doi.org/10.1007/BF00116251

Reiersøl, O. (1945). Confluence analysis of means of instrumental sets of variables. *Arkiv. for Mathematik, Astronomi Och Fysik*, 32.
Rosenbaum, P. & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70, 41–55.
https://doi.org/10.21236/ada114514
Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
https://doi.org/10.1037/h0042519
Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5), 688–701.
https://doi.org/10.1037/h0037350
Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.
Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 44(1).
https://doi.org/10.1147/rd.33.0210
Schultz, H. (1930). *The Meaning of Statistical Demand Curves*. Chicago: University of Chicago.
Shai, S. S. & Shai, B. D. (2014). *Understanding Machine Learning From Theory to Algorithms*. Cambridge: Cambridge University Press.
Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, 88(422), 486–494.
https://doi.org/10.2307/2290328
Shalev-Shwartz, S. & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge: 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*. Cambridge, Mass: MIT Press.
Silverman, B.W. (1986). *Density Estimation*. London: 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.
https://doi.org/10.1111/j.2517-6161.1977.tb01603.x
Tam, K. Y. & Kiang, M. Y. (1992). Managerial Applications of Neural Networks: The Case of Bank Failure Predictions. *Management Science*, 38(7), 926–947.
https://doi.org/10.1287/mnsc.38.7.926
Tan, H. (1995). *Neural-Network model for stock forecasting*. MSc Thesis, Texas Tech. University. https://bit.ly/2UplmYu
Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B.*, 58(1), 267–288.
https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
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(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(1), 24–36.
https://doi.org/10.2307/1907382
Tropp, (2011). Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3(1), 115–126.
https://doi.org/10.1142/S1793536911000787
Tseng, P. (2001). Convergence of a block coordinate descent for nondifferentiable minization. *Journal of Optimization Theory and Applications*, 109(3), 475–494.
https://doi.org/10.1023/A:1017501703105
Tufféry, S. (2001). *Data Mining and Statistics for Decision Making*. New York: Wiley Interscience.
Tukey, J. W. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, 29(2), 614–623.
https://doi.org/10.1214/aoms/1177706647
Vailiant, L.G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11), 1134–1142.
https://doi.org/10.1145/1968.1972
Vapnik, V. (1998). *Statistical Learning Theory*. New York: 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(2), 264–280.
https://doi.org/10.1137/1116025
Varian, H.R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.
https://doi.org/10.1257/jep.28.2.3

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.
https://doi.org/10.1177/1471082X14561155
Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, 26(4), 359–372.
https://www.jstor.org/stable/25049340
Watt, J., Borhani, R. & Katsaggelos, A. (2016). *Machine Learning Refined: Foundations, Algorithms, and Applications*. Cambridge: Cambridge University Press.
White, H. (1989). Learning in Artificial Neural Networks: A Statistical Perspective. *Neural Computation*, 1(4), 425–464.
https://doi.org/10.1162/neco.1989.1.4.425
Widrow, B. & Hoff, M. E. Jr. (1960). Adaptive Switching Circuits. *IRE WESCON Convention Record*, 4, 96–104.
https://apps.dtic.mil/dtic/tr/fulltext/u2/241531.pdf
Wolpert, D. H. & Macready, W. G. (1997). No Free Lunch Theorems for Optimization, *IEEE Transactions on Evolutionary Computation*, 1(1), 67.
https://doi.org/10.1109/4235.585893
Wolpert, David (1996). The Lack of A Priori Distinctions between Learning Algorithms, *Neural Computation*, 1341-1390.
https://doi.org/10.1162/neco.1996.8.7.1341
Working, E. J. (1927). What Do Statistical "Demand Curves" Show? *The Quarterly Journal of Economics*, 41(2), 212–35.
https://doi.org/10.2307/1883501
Yu, K. & Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4), 437–447.
https://doi.org/10.1016/S0167-7152(01)00124-9
Zinkevich M. A., Weimer, M., Smola, A. & Li, L. (2010). Parallelized Stochastic Gradient. *Advances in neural information processing systems*, 2595–2603.
https://papers.nips.cc/paper/4006-parallelized-stochastic-gradient-descent.pdf