Sélection de modèles linéaire et non linéaires

Monrocq Christophe

Université René Descartes
Groupe de Recherche en Imagerie Biomédicale
45 rue des Saint-Pères
75006 PARIS
FRANCE
e-mail : monrocq@citi2 fr

Résumé

Nous allons aborder les problèmes d'estimation des erreurs d'apprentissage et de généralisation pour des modèles linéaires ou non. Il est reconnu que cette sélection doit suivre le principe de parcimonie, i.e. que les modèles les plus simples seront choisis prioritairement grâce à une pénalisation des modèles complexes. Mais le problème majeur qui se pose concerne la forme et l'importance du terme de pénalité.
Les nouvelles règles proposées pour la sélection de modèles, parmi lesquelles la règle GAE, reposent sur les estimations des erreurs d'apprentissage et de généralisation. Ces estimations vont aussi permettre de clarifier les liens qui existent entre ces deux erreurs, d'expliciter le terme pénalité précédent et de présenter des critères semblables à ceux d'Akaike pour les modèles linéaires (FPE et AIC) mais aussi valables pour des modèles non linéaires

1 Introduction

Résoudre un problème de classification (supervisée ou non) ou d'approximation de fonctions ne se limite pas à une étape d'optimisation permettant d'estimer les paramètres du modèle, mais nécessite aussi une seconde étape, dite de validation, qui nous indiquera le comportement du modèle en phase d'utilisation. Il s'agit alors d'évaluer l'erreur en généralisation.

L'apprentissage est réalisé en utilisant une base de données constituée de couples (x_i, y_i), où y est la sortie désirée associée à l'entrée x. Comme cette base est de taille finie, un apprentissage sans erreur serait possible en autorisant une complexité suffisante du modèle[1]. Cependant un apprentissage par cœur n'est pas souhaitable car d'une part cette base d'apprentissage ne représente qu'un échantillon possible de la réalité ; d'autre part si les exemples sont bruités, cet apprentissage va avoir l'inconvénient d'apprendre le bruit.
Par conséquent, l'étape de validation devra nous permettre de trouver le meilleur compromis entre l'erreur à l'apprentissage et la complexité du modèle.

Les résultats présentés dans cet article vont permettre d'introduire et d'expliquer ce compromis. Le premier résultat original (théorème 3), d'où découleront les suivants, met à jour une relation entre les erreurs à l'apprentissage et en généralisation, ce qui aura plusieurs conséquences : i) des résultats connus pour le cas linéaire (critères FPE et AIC d'Akaike) seront étendus au cas non linéaire ; ii) une nouvelle approche, "la règle GAE[2]", est proposée pour déterminer la complexité optimale du modèle.

Les problèmes de classification ou d'approximation de fonctions rentrent dans le cadre de la régression. Il s'agit alors de déterminer la valeur d'une variable aléatoire $\mathbf{Y}$ d'après les informations

1. dans le cas de la régression polynomiale, on peut augmenter le degré du polynôme pour passer par tous les points de la base d'apprentissage.

2. GAE = Graph of the Asymptotic Error. La règle GAE consiste à représenter l'Erreur Asympototique des modèles en fonction d'une mesure de la complexité de ceux-ci (par exemple le nombre de paramètres des modèles.)

apportées par d'autres variables constituant le vecteur aléatoire $\mathbf{X}$ des régresseurs. Par exemple, $\mathbf{Y}$ peut être la consommation électrique dans une journée tandis que le vecteur $\mathbf{X}$ regroupe des mesures sur la température, l'ensoleillement, la pluie, le type de jour (férié ou non) Notre problème consiste alors à trouver un modèle $f(\theta, x)$, paramétré par le vecteur $\theta \in \Re^p$ tel qu'une fonction $q(\theta, x, y)$ de l'erreur entre la sortie désirée y et celle délivrée par le modèle $f(\theta, x)$ soit aussi petite que possible. Dans un cadre stochastique où les variables $\mathbf{X}$ et $\mathbf{Y}$ sont aléatoires notre objectif idéal serait de minimiser

$$Q(\theta) = E\left(q(\theta, x, y)\right) = \int q(\theta, x, y) dF(x, y) \tag{1}$$

L'erreur (1) est connue sous le nom d'erreur en généralisation et dépend de la loi de distribution F du couple des variables aléatoires $\mathbf{X}$ et $\mathbf{Y}$. Dans la pratique, cette loi est inconnue et, de ce fait, cette erreur ne peut pas être calculée. Par conséquent, le vecteur θ des paramètres est estimé en appliquant la procédure d'optimisation non pas à l'erreur $Q(\theta)$ mais à l'erreur empirique $Q_N(\theta)$, dite erreur d'apprentissage, évaluée sur la base d'apprentissage qui contient N couples (x_i, y_i) constituant notre seule source d'information:

$$Q_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} q(\theta, x_i, y_i) \tag{2}$$

Comme l'erreur en généralisation nous donne une mesure de la qualité du comportement du modèle, le meilleur modèle sera celui pour lequel cette est minimale; et de ce fait cette erreur constitue la règle idéale pour selectionner le bon modèle. Il s'agit alors d'estimer cette erreur (1). Pour l'estimation de cette erreur, la plupart des procédures sont très coûteuses en calculs (ex. : boostrap, jackknife) et/ou n'utilisent pas toutes les données pour la phase d'apprentissage (ex. : la validation croisée [Stone74, Efron et al 86]).

À l'opposé, les approches proposées pour estimer l'erreur en généralisation, tout comme les critères FPE et AIC, sont basées sur des relations analytiques liant l'erreur à l'apprentissage (2) et l'erreur en généralisation (1), permettant ainsi d'estimer rapidement cette dernière tout en gardant l'ensemble des données à l'apprentissage.

Quant à l'approche basée sur le graphe de l'erreur asympototique (règle GAE), comme les critères FPE et AIC, elle nécessite de construire plusieurs modèles afin de sélectionner le meilleur; de plus, pour chaque modèle on doit réaliser plusieurs apprentissages en faisant varier la taille de la base d'apprentissage Cependant le coût des calculs reste très raisonnable car l'intérêt de cette méthode est d'utiliser des bases d'apprentissage de tailles modestes

La section 2 introduit les notations et deux théorèmes issus de [White81] qui sont à la base des résultats présentés. La section 3 présente les approximations des erreurs moyennes à l'apprentissage et en généralisation. Ces approximations vont ensuite nous servir i) à présenter les critères FPE et AIC dans le cas non linéaire; ii) à proposer de nouvelles approches (section 4) pour choisir la meilleure structure (et donc complexité) de modèle; ces nouvelles approches concernent l'estimation de l'erreur en généralisation et la règle GAE.

Techniquement les calculs effectués n'utilisent guère plus que la formule de Taylor appliquée aux fonctions $Q(\theta)$ et $Q_N(\theta)$. La difficulté provient d'une formulation convenable du cadre théorique et des hypothèses

2 Définition de l'estimateur minimisant le risque empirique

L'étude de la consistance et de la loi limite de l'estimateur $\hat{\theta}_N$ minimisant une fonction d'erreur empirique a été l'objet de nombreuses publications. Dans le cas particulier où la fonction d'erreur est quadratique (estimateur des moindre-carrés), un résumé assez complet est donné par [Antoniadis et al 92] et [Seber et al 89] qui cite des résultats de [Jennrich69, Malinvaud70a, Malinvaud70b, Gallant75, Wu81, White81, Amemiya83]
Les deux théorèmes suivants [White81] généralisent les résultats des auteurs précédents dans un cadre permettant d'utiliser une fonction d'erreur non quadratique et un modèle non correct (voir définition 8). Ces deux théorèmes sont à la base des résultats que nous présentons pour estimer les erreurs moyennes à l'apprentissage et en généralisation

Nous posons les hypothèses suivantes [White81] :

Hypothèse I

La base d'apprentissage est composée de N couples $z_t = (x_t, y_t)$ tels que :

$$y_t = g(x_t) + n_t \qquad t = 1, \ldots, N$$

où la fonction $g(\cdot)$, inconnue dans la pratique, devra être estimée. Les vecteurs aléatoires (x_t, y_t) sont indépendants identiquement distribués de fonction de distribution F sur un espace euclidien Ω. Le bruit est représenté par la variable aléatoire n.

Hypothèse II

Les n_t sont i.i.d. de moyenne nulle et de variance finie λ_0.

Comme la fonction $g(\cdot)$ est inconnue, un modèle $f(\theta, x)$ est utilisé comme estimation de $g(\cdot)$. Le vecteur θ contient les paramètres du modèle.

Hypothèse III

$f(\theta, x)$ représente la fonction réalisée par le modèle. $f(\theta, x)$ est une fonction continue de θ pour x fixé et une fonction mesurable de x pour $\theta \in \Theta$, où Θ est un espace euclidien.

Hypothèse IV

Soit $z = (x, y)$ et $q(\theta, z)$ une fonction définie sur $\Theta \times \Omega$ telle que $q(\theta, z)$ soit continue par rapport à θ pour chaque z et mesurable par rapport à z pour chaque θ. Supposons que $|q(\theta, z)| \leq m(z)$, $\forall(\theta, z)$ où $m(\cdot)$ est intégrable par rapport à la loi F des observations, i.e. :

$$\int m(z) dF(z) < \infty$$

Hypothèse V

L'erreur en généralisation

$$Q(\theta) = \int q(\theta, x, y) dF(x, y)$$

a un minimum unique $\theta^* \in \Theta$.

Notations

Nous allons utiliser les notations suivantes :

- $z_t = (x_t, y_t)$ représente les observations de la base d'apprentissage ;
- Z^N : la base d'apprentissage qui contient N observations z_t ;
- le vecteur $\theta \in \Re^p$ regroupe les paramètres du modèle ;

- l'erreur entre les sorties désirées et celles du modèle : $q(\theta, z_t)$ [3] ;
- le critère à minimiser sur la base d'apprentissage :

$$Q_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} q(\theta, z_t)$$

- $\hat{\theta}_N$ est la valeur de θ qui minimise $Q_N(\theta)$: $\hat{\theta}_N = \underset{\theta \in S_j}{\operatorname{argmin}} Q_N(\theta)$
- l'erreur en généralisation pour $\hat{\theta}_N$ fixé :

$$Q(\hat{\theta}_N) = \int q(\hat{\theta}_N, z) dF(z)$$

- l'erreur moyenne en généralisation :

$$J(\hat{\theta}_N) = E_{\hat{\theta}_N}\left(Q(\hat{\theta}_N)\right) \tag{3}$$

où l'espérance est prise sur les $\hat{\theta}_N$ obtenus avec toutes les bases Z^N possibles de taille N;

- θ^* est la valeur de θ qui minimise $Q(\theta)$: $\theta^* = \underset{\theta \in S_j}{\operatorname{argmin}} Q(\theta)$
- si le modèle est correct [4] $\theta_0 = \theta^*$;

Nous rappellons maintenant un théorème précisant les conditions qui assurent que $\hat{\theta}_N$ converge presque surement vers θ^* :

Consistance

Théorème 1 [White81] théorème 2.1

Sous les hypothèses I à V, la suite $(\hat{\theta}_N)_N$ de vecteurs qui minimisent l'erreur empirique

$$Q_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} q(\theta, z)$$

converge vers le minimum θ^* de $Q(\theta)$ pour presque toute base d'apprentissage issue de F.

Ce théorème nécessite les hypothèses I à V mais est valable que le modèle soit linéaire ou non Nous allons maintenant préciser la distribution asymptotique de $\hat{\theta}_N$.

Normalité asymptotique

Nous allons maintenant poser des hypothèses qui permettent d'obtenir la loi limite de l'estimateur minimisant le risque empirique $Q_N(\theta)$.

Nous définissons, quand elles existent

les dérivées suivantes :

$$\dot{q}_i(\theta) = \frac{\partial q(\theta, z)}{\partial \theta_i} \quad , \quad \dot{q}_{ti}(\theta) = \frac{\partial q(\theta, z_t)}{\partial \theta_i}$$

$$\ddot{q}_{ij}(\theta) = \frac{\partial^2 q(\theta, z)}{\partial \theta_i \partial \theta_j} \quad , \quad \ddot{q}_{tij}(\theta) = \frac{\partial^2 q(\theta, z_t)}{\partial \theta_i \partial \theta_j}$$

3 par exemple $q(\theta, z_t) = \|y_t - f(\theta, x_t)\|^2$
4 voir definition 8

les sommes

$$b_{Nij} = \frac{1}{N}\sum_{t=1}^{N} \dot{q}_{ti}(\theta)\dot{q}_{tj}(\theta) \quad , \qquad h_{Nij} = \frac{1}{N}\sum_{t=1}^{N} \ddot{q}_{tij}(\theta)$$

les moyennes

$$b_{ij} = E\left(\dot{q}_i(\theta)\dot{q}_j(\theta)\right) \quad , \qquad h_{ij} = E\left(\ddot{q}_{ij}(\theta)\right)$$

et les matrices

$$\mathbf{H}_N(\theta) = \{h_{Nij}\} \quad , \qquad \mathbf{B}_N(\theta) = \{b_{Nij}\}$$
$$\mathbf{H}(\theta) = \{h_{ij}\} \quad , \qquad \mathbf{B}(\theta) = \{b_{ij}\}$$

et quand les inverses existent, les matrices

$$\mathbf{C}(\theta) = \mathbf{H}(\theta)^{-1}\mathbf{B}(\theta)\mathbf{H}(\theta)^{-1} \quad , \qquad \mathbf{C}_N(\theta) = \mathbf{H}_N^{-1}(\theta)\mathbf{B}_N(\theta)\mathbf{H}_N^{-1}(\theta)$$

Hypothèse VI

$\dot{q}_i(\theta)$, $i = 1, \ldots, p$ sont des fonctions mesurables de z à θ fixé et continûment différentiables de θ pour chaque z.

Hypothèse VII

$|\dot{q}_i(\theta)\dot{q}_j(\theta)|$ et $|\ddot{q}_{ij}(\theta)|$, $i, j = 1, \ldots, p$ sont dominées par des fonctions intégrables par rapport à F pour tout (θ, z).

Hypothèse VIII

θ^* est un point intérieur à Θ et les matrices $\boldsymbol{H}(\theta^*)$ et $\boldsymbol{B}(\theta^*)$ ne sont pas singulières.

d'où le théorème suivant

Théorème 2 [White81] théorème 3 3

Supposons que $q(\theta, z)$ et $Q(\theta)$ vérifient les conditions du théorème 1 et soit $(\hat{\theta}_N)_N$ une suite de vecteurs de Θ qui minimisent l'erreur empirique

$$Q_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} q(\theta, z)$$

où la base d'apprentissage $Z^N = \{z_t\}_{t=1, \ldots, N}$ est issue de la loi F.

Si les hypothèses I à VIII sont vérifiées, alors
a) $\sqrt{N}(\hat{\theta}_N - \theta^*) \approx \mathcal{N}(0, \boldsymbol{C}(\theta^*))$ où $\boldsymbol{C}(\theta^*) = \boldsymbol{H}(\theta^*)^{-1}\boldsymbol{B}(\theta^*)\boldsymbol{H}(\theta^*)^{-1}$
b) $\boldsymbol{C}_N(\theta)$ est un estimateur fortement consistant de $\boldsymbol{C}(\theta)$

3 Approximations des erreurs à l'apprentissage et en généralisation

Les analyses suivantes sont asymptotiques (N grand) et locales ($\hat{\theta}_N$ est dans un voisinage $\mathcal{V}$ de θ^*).

Sous les deux conditions que l'estimateur soit consistant en θ^* et qu'il admette la loi limite précédente, nous allons maintenant présenter des approximations des erreurs à l'apprentissage en généralisation.

Théorème 3

Sous les hypothèses du théorème 2, en effectuant des développements limités d'ordre deux des fonctions d'erreur $Q_N(\theta)$ et $Q(\theta)$ en leur minumum respectif $\hat{\theta}_N$ et θ^*, nous obtenons les estimations suivantes des erreurs associées à l'estimateur $\hat{\theta}_N$ (estimé sur une base d'apprentissage de taille N) :

- *Erreur en généralisation :*

$$Q(\hat{\theta}_N) \approx Q(\theta^*) + \frac{1}{2}(\theta^* - \hat{\theta}_N)^T Q''(\theta^*)(\theta^* - \hat{\theta}_N) \tag{4}$$

- *Erreur à l'apprentissage :*

$$Q_N(\hat{\theta}_N) \approx Q_N(\theta^*) - \frac{1}{2}(\theta^* - \hat{\theta}_N)^T Q''(\theta^*)(\theta^* - \hat{\theta}_N) \tag{5}$$

où $Q''(\theta^*) = \boldsymbol{H}(\theta^*) = \left\{ E\left(\ddot{q}_{ij}(\theta)\right) \right\}_{i,j=1,\ldots,p}$

Preuve

Pour l'erreur en généralisation : on effectue un développement de Taylor à l'ordre deux la fonction $Q(\theta)$ autour de son minimum θ^*.

Pour l'erreur à l'apprentissage : on effectue un développement de Taylor à l'ordre deux la fonction $Q_N(\theta)$ autour de son minimum $\hat{\theta}_N$.

Les résultats découlent du fait que le reste intégral de chacun des développements de Taylor à l'ordre deux s'annule sous les hypothèses posées. ■

3.1 Moyennes des erreurs issues des apprentissages sur toutes les bases d'apprentissage possibles de taille N

Nous allons maintenant calculer les moyennes des erreurs (4) et (5) sur toutes les bases d'apprentissage de taille N. Pour un apprentissage ayant lieu sur une base d'apprentissage Z^1 de taille N, nous obtenons un estimateur $\hat{\theta}_N{}^1$, l'erreur $Q_N(\hat{\theta}_N{}^1)$ sur la base d'apprentissage et l'erreur (théorique) en généralisation $Q(\hat{\theta}_N{}^1)$. En réalisant des apprentissages sur des bases d'apprentissage différentes, nous obtenons différents estimateurs :

base d'apprentissage	estimateur	erreur à l'apprentissage	erreur en généralisation
Z^1	$\hat{\theta}_N{}^1$	$Q_N(\hat{\theta}_N{}^1)$	$Q(\hat{\theta}_N{}^1)$
Z^2	$\hat{\theta}_N{}^2$	$Q_N(\hat{\theta}_N{}^2)$	$Q(\hat{\theta}_N{}^2)$
Z^3	$\hat{\theta}_N{}^3$	$Q_N(\hat{\theta}_N{}^3)$	$Q(\hat{\theta}_N{}^3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
on calcule la moyenne des erreurs :		$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} Q_N(\hat{\theta}_N{}^k)$	$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} Q(\hat{\theta}_N{}^k)$
		$= E\left(Q_N(\hat{\theta}_N)\right)$	$= E\left(Q(\hat{\theta}_N)\right) = J(\hat{\theta}_N)$

En calculant les moyennes des erreurs à l'apprentissage (5) et en généralisation (4), nous obtenons alors le théorème suivant :

Théorème 4

En effectuant des développements limités d'ordre deux des fonctions d'erreur $Q_N(\theta)$ et $Q(\theta)$, sous les hypothèses du théorème 2, nous obtenons les estimations suivantes des erreurs moyennes associées à l'estimateur $\hat{\theta}_N$ (estimé sur une base d'apprentissage de taille N) :

- *Erreur moyenne en généralisation :*

$$J(\hat{\theta}_N)=E\left(Q(\hat{\theta}_N)\right) \approx E\left(Q_N(\hat{\theta}_N)\right)+tr\left(Q''(\theta^*)E\left((\theta^*-\hat{\theta}_N)(\theta^*-\hat{\theta}_N)^T\right)\right) \quad (6)$$

$$\approx Q(\theta^*)+\frac{1}{2}tr\left(Q''(\theta^*)E\left((\theta^*-\hat{\theta}_N)(\theta^*-\hat{\theta}_N)^T\right)\right) \quad (7)$$

- *Erreur moyenne à l'apprentissage :*

$$E\left(Q_N(\hat{\theta}_N)\right) \approx Q(\theta^*)-\frac{1}{2}tr\left(Q''(\theta^*)E\left((\theta^*-\hat{\theta}_N)(\theta^*-\hat{\theta}_N)^T\right)\right) \quad (8)$$

- *La somme des erreurs moyennes à l'apprentissage et en généralisation est indépendante de N*

$$E\left(Q_N(\hat{\theta}_N)\right)+J(\hat{\theta}_N) \approx 2Q(\theta^*) \quad (9)$$

Interprétation: *les erreurs moyennes en généralisation et à l'apprentissage sont approximativement égales à la somme de deux termes : un premier terme dépendant du modèle et du problème (loi théorique des observations) et d'un second terme dépendant du hessien de la fonction d'erreur et d'une statistique des bases d'apprentissage de taille N représentée par $E\left((\theta^*-\hat{\theta}_N)(\theta^*-\hat{\theta}_N)^T\right)$.*

Preuve

Nous partons des résultats du théorème 3 (équations (5) et (4)) donnant les estimations des erreurs en généralisation et à l'apprentissage, et nous prenons la moyenne sur toutes les bases d'apprentissage de taille N. ■

En utilisant le théorème 2 donnant la loi limite de l'estimateur $\hat{\theta}_N$, nous pouvons expliciter le terme $E\left((\theta^*-\hat{\theta}_N)(\theta^*-\hat{\theta}_N)^T\right)$, d'où le corollaire suivant du théorème 4 :

Corollaire 5

Sous les hypothèses du théorème 2, les erreurs moyennes à l'apprentissage et en généralisation associées à un estimateur (consistant) minimisant le risque empirique sont :

Erreur moyenne en généralisation

$$J(\hat{\theta}_N) \approx E\left(Q_N(\hat{\theta}_N)\right)+\frac{1}{N}tr\left(\boldsymbol{B}(\theta^*)\boldsymbol{H}^{-1}(\theta^*)\right) \quad (10)$$

$$\approx Q(\theta^*)+\frac{1}{2}\frac{1}{N}tr\left(\boldsymbol{B}(\theta^*)\boldsymbol{H}^{-1}(\theta^*)\right) \quad (11)$$

Erreur moyenne à l'apprentissage

$$E\left(Q_N(\hat{\theta}_N)\right) \approx Q(\theta^*)-\frac{1}{2}\frac{1}{N}tr\left(\boldsymbol{B}(\theta^*)\boldsymbol{H}^{-1}(\theta^*)\right) \quad (12)$$

Remarque 6

- L'équation (11) montre que l'erreur moyenne en généralisation $J(\hat{\theta}_N)$ atteint sa valeur minimale $Q(\theta^*)$ quand $N \to \infty$; si nous ne disposons que d'un nombre fini d'exemples alors l'erreur moyenne en généralisation sera toujours supérieure à $Q(\theta^*)$;
- Inversement, nous constatons que l'erreur moyenne à l'apprentissage (12) croît quand $N \to \infty$;
- $\lim_{N\to\infty} E\left(Q_N(\hat{\theta}_N)\right) = \lim_{N\to\infty} J(\hat{\theta}_N) = Q(\theta^*)$

Remarque 7

Les équations (11) et (12) montrent que les erreurs moyennes en généralisation et à l'apprentissage sont distribuées sur une droite en fonction de $1/N$. Les deux droites sont de pentes opposées ($\pm\frac{1}{2}\text{tr}\left(\mathbf{B}(\theta^*)\mathbf{H}^{-1}(\theta^*)\right)$) et se coupent sur l'axe des ordonnées en $Q(\theta^*)$ (i.e. quand la base d'apprentissage est de taille infinie).
Nous exploiterons cette remarque dans la section 4 pour proposer de nouvelles méthodes d'estimation de $J(\hat{\theta}_N)$.

3.2 Application au coût quadratique: critère FPE[5]

$$\begin{aligned} q(\theta, z_t) &= (y_t - f(\theta, x_t))^2 = e^2(z_t, \theta) \\ Q_N(\theta) &= \frac{1}{N}\sum_{t=1}^{N} e^2(z_t, \theta) \\ Q(\theta) &= \int e^2(z, \theta) dF(z) \end{aligned}$$

Nous appliquons le corollaire 5 avec :

$$\mathbf{H}(\theta) = Q''(\theta) = 2E\left(\frac{\partial f(\theta, z)}{\partial \theta}\frac{\partial f(\theta, z)}{\partial \theta}^T + e(z, \theta)\frac{\partial^2 f(\theta, z)}{\partial \theta^2}\right) \quad (13)$$

$$\mathbf{B}(\theta) = 4E\left(\frac{\partial f(\theta, z)}{\partial \theta}e(z, \theta)\left[\frac{\partial f(\theta, z)}{\partial \theta}e(z, \theta)\right]^T\right) \quad (14)$$

$$\mathbf{C}(\theta^*) = \mathbf{H}^{-1}(\theta^*)\mathbf{B}(\theta^*)\mathbf{H}^{-1}(\theta^*) \quad (15)$$

Modèle correct

Pour obtenir une estimation de $J(\hat{\theta}_N)$ nous devons estimer $\text{tr}\left(\mathbf{B}(\theta^*)\mathbf{H}^{-1}(\theta^*)\right)$, ce qui nous amène à l'hypothèse de modèle correct.

5. FPE = Final Prediction Error

Définition 8 Modèle correct

La base d'apprentissage Z^N est constituée de couples $z_t = (x_t, y_t), t = 1, \ldots, N$, de fonction de distribution $F(z)$ et vérifiant la relation :

$$y_t = g(x_t) + n_t \tag{16}$$

Le modèle $f(\theta, x_t)$ paramétré par θ est dit correct s'il existe θ_0 tel que $g(x_t) = f(\theta_0, x_t)$.
La relation liant les entrées x et les sorties désirées y est alors :

$$y_t = f(\theta_0, x_t) + n_t \tag{17}$$

où n_t modélise un bruit indépendant, identiquement distribué (iid) et additif sur les mesures, de moyenne nulle et de variance λ_0 finie : $\lambda_0 = En_t^2$.

Lorsque la taille de la base d'apprentissage est suffisante, l'erreur minimale en généralisation est atteinte si le modèle est correct. L'objectif idéal peut être alors de déterminer la structure du modèle correct ; cependant si l'échantillon d'apprentissage est trop petit, un modèle moins complexe que le modèle correct peut donner de meilleurs résultats en généralisation.

Si nous nous plaçons sous les hypothèses de la définition 8 (i.e. de modèle correct : $\exists \theta^* = \theta_0$ et $e(z, \theta_0) = n$), nous obtenons :

$$\begin{cases} Q(\theta_0) = \lambda_0 \quad \text{la variance du bruit } n \\ E\left(e(z,\theta_0)\,\dfrac{\partial^2 f(\theta_0,z)}{\partial\theta^2}\right) = E\left(n\,\dfrac{\partial^2 f(\theta_0,z)}{\partial\theta^2}\right) = 0 \Longrightarrow \mathbf{H}(\theta_0) = 2E\left(\dfrac{\partial f(\theta_0,z)}{\partial\theta}\,\dfrac{\partial f(\theta_0,z)}{\partial\theta}^T\right) \quad (18) \\ \mathbf{B}(\theta_0) = 4\lambda_0 E\left(\dfrac{\partial f(\theta_0,z)}{\partial\theta}\,\dfrac{\partial f(\theta_0,z)}{\partial\theta}^T\right) = 2\lambda_0\mathbf{H}(\theta_0) \quad (19) \end{cases}$$

En appliquant le corollaire 5 à un modèle correct appris en minimisant un coût quadratique, nous obtenons la proposition suivante :

Proposition 9

Sous les hypothèses du théorème 2 et de la définition 8 (modèle correct), les estimations des erreurs moyennes associées à l'estimateur $\hat{\theta}_N$ minimisant l'erreur quadratique

$$Q_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \|y_i - f(\theta, x_i)\|^2$$

sont :

- *Erreur moyenne en généralisation*

$$J(\hat{\theta}_N) \approx Q(\theta_0) + \frac{\lambda_0 p}{N} = \lambda_0(1 + \frac{p}{N}) \tag{20}$$

- *Erreur moyenne à l'apprentissage*

$$E\left(Q_N(\hat{\theta}_N)\right) \approx Q(\theta_0) - \frac{\lambda_0 p}{N} = \lambda_0(1 - \frac{p}{N}) \tag{21}$$

Preuve

En utilisant (18) et (19) nous avons :

$$\frac{1}{N}\mathrm{tr}\left(\mathbf{B}(\theta_0)\mathbf{H}^{-1}(\theta_0)\right) = 2\frac{\lambda_0 p}{N} \tag{22}$$

où λ_0 est la variance du bruit iid et additif sur les sorties désirées et p le nombre de paramètres du modèle. En injectant la relation (22) dans les équations (11) et (12) du corollaire 5, nous obtenons les résultats désirés. ■

Remarque 10

- Les équations précédentes (20) et (21) montrent que l'erreur moyenne en généralisation $J(\hat{\theta}_N)$ atteint sa valeur minimale $Q(\theta_0)$ quand $N \to \infty$; si nous ne disposons que d'un nombre fini d'exemples, alors l'erreur moyenne en généralisation sera toujours supérieure à $Q(\theta_0)$. Nous constatons aussi que $J(\hat{\theta}_N)$ croît quand le nombre de paramètres p augmente.
- Nous constatons que l'erreur moyenne à l'apprentissage décroît quand p augmente et croît quand $N \to \infty$.

En partant de la relation (20), nous obtenons la proposition suivante :

Proposition 11

Lorsque la variance λ_0 du bruit iid et additif sur les réponses est inconnue, si le modèle est correct, nous pouvons estimer l'erreur moyenne en généralisation $J(\hat{\theta}_N)$ par :

$$J(\hat{\theta}_N) \approx FPE = \frac{1+p/N}{1-p/N} Q_N(\hat{\theta}_N) = \frac{1+p/N}{1-p/N}\frac{1}{N}\sum_{t=1}^{N} n^2(t,\hat{\theta}_N) \tag{23}$$

Nous retrouvons le critère FPE présenté par Akaike et utilisé pour la sélection de la complexité d'un modèle linéaire ; mais la présentation faite ici est valable que le modèle soit linéaire ou non, avec un bruit indépendant et additif sur les réponses.

Preuve

En supposant que $Q_N(\hat{\theta}_N) \approx E\left(Q_N(\hat{\theta}_N)\right)$, d'après la relation (21) un estimateur non biaisé de la variance λ_0 du bruit sur les sorties y est :

$$\hat{\lambda}_N = \frac{Q_N(\hat{\theta}_N)}{1-(p/N)} \tag{24}$$

d'où la proposition en remplaçant λ_0 par $\hat{\lambda}_N$ dans la relation (20) ■

3.3 Application au coût de vraisemblance : critère AIC[6]

$$\begin{aligned} Q_N(\theta) &= -\text{ (log-vraisemblance de la base d'apprentissage)} \\ &= -L_N(\theta) \end{aligned}$$

Sous des conditions générales (voir [White82], [White92] chap. 14) et notamment que le modèle soit correct, l'estimateur du maximum de vraisemblance est asymptotiquement normalement distribué :

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \approx \mathcal{N}(0, \mathcal{F}^{-1})$$

où $\mathcal{F} = -Q''(\theta_0)$ est la matrice d'information de Fisher

6. AIC = Akaike's Information Criterion

De ce fait

$$E_{\hat{\theta}_N}(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T = \frac{1}{N}\left(Q''(\theta_0)\right)^{-1}$$

d'où

$$\mathrm{tr}\left(Q''(\theta_0)E_{\hat{\theta}_N}(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T\right) = \frac{1}{N}\mathrm{tr}\left(Q''(\theta_0)\left(Q''(\theta_0)\right)^{-1}\right) = \frac{1}{N}\dim\,\theta_0 = \frac{p}{N}$$

en appliquant la relation 6 de la proposition 4 nous retrouvons le critère AIC d'Akaike [Akaike74] :

$$\boxed{J(\hat{\theta}_N) = -L_N(\theta) + \frac{p}{N}} \tag{25}$$

Lorsque les erreurs sur les réponses y suivent une distribution normale, le critère AIC consiste à chercher le minimum de :

$$-L_N(\theta) + \frac{p}{N} \quad = \quad -\ln\left[\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\exp\left(-\frac{\sum_{i=1}^{N}\|y_i - f(\hat{\theta}_N, x_i)\|^2}{2\hat{\sigma}^2}\right)\right] + \frac{p}{N}$$

Si la variance du bruit sur les réponses est estimée par

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\|y_i - f(\hat{\theta}_N, x_i)\|^2$$

nous obtenons que la minimisation du critère (25) est équivalente à celle de :

$$\boxed{AIC = \frac{1}{2}\ln\hat{\sigma}^2 + \frac{p}{N}} \tag{26}$$

Les critères (25) et (26) sont constitués de deux termes :

1. le premier terme caractérise la qualité de l'apprentissage puisqu'il s'agit de l'erreur sur la base d'apprentissage ;
2. le second terme pénalise la complexité du modèle (en fonction de la taille de la base d'apprentissage).

Remarque 12

Dans la littérature, les critères FPE et AIC sont utilisés pour estimer l'erreur en généralisation et, de ce fait, pour en déduire le modèle de complexité optimale, i.e. donnant l'erreur minimale en généralisation. Cependant, dans cette section ces critères ont été obtenus sous une hypothèse de modèle correct. La question qui se pose est alors : est-il valide d'utiliser ces critères pour estimer l'erreur en généralisation lorsque le modèle est insuffisamment ou trop complexe ?

Si nous utilisons un modèle trop complexe mais qui contient le modèle correct en tant que sous modèle, alors l'utilisation des critères FPE et AIC est certainement valide. Ce qui sera d'ailleurs vérifié lors des simulations effectuées dans la section 4 dans le cadre du coût quadratique.

Par contre, si le modèle n'est pas suffisament complexe, la validité de l'utilisation de ces critères semble moins évidente puisque le modèle est alors biaisé [Geman et al.92].

Des travaux en cours portent sur l'évaluation de l'erreur en généralisation des modèles de complexité insuffisante. À notre connaissance, seul [Kannurpatti et al.91] a obtenu un critère semblable au FPE pour la sélection de modèles, i.e. valable pour des modèles pas assez ou trop complexes. Cependant ses résultats ont été établis dans le cas linéaire et sous des hypothèses contraignantes

Résultats des simulations Les simulations effectuées ont prouvé que les critères FPE et AIC peuvent servir à déterminer le modèle ayant l'erreur minimale en généralisation Ceci pour des modèles linéaires et non linéaires : régression polynomiale [7] et réseaux neuronaux du type Perceptron Multi-Couches.

Cependant les limitations théoriques présentées dans la remarque 12 précédente nous amènent à présenter et à détailler d'autres approches pour sélectionner le bon modèle, dont notamment la "règle GAE"

4 Nouvelle estimation de l'erreur moyenne en généralisation et règle GAE de sélection

Jusqu'alors nous avons déterminé la complexité optimale en estimant l'erreur moyenne en généralisation à partir de l'erreur sur une **seule** base d'apprentissage de taille N (critères FPE et AIC).

Les nouvelles approches que nous proposons maintenant concernent un problème très voisin : en supposant que nous connaissons les erreurs moyennes à l'apprentissage et/ou en généralisation associées à des modèles appris sur m bases d'apprentissage $\mathcal{A}_i$ de tailles croissantes N_i, $i = 1, \ldots, m$, est-il possible d'en déduire les erreurs moyennes à l'apprentissage et en généralisation si la base d'apprentissage avait contenu N observations avec $N \neq N_i$?

Plus précisément le problème posé est :

- **Information disponible**: nous disposons d'estimations des erreurs moyennes à l'apprentissage et/ou en généralisation de modèles appris sur m bases d'apprentissage $(\mathcal{A}_i)_{i=1}^m$ de tailles N_i avec N_i "petit"
- **Questions posées** : i) Est-ce-que l'erreur moyenne en généralisation $J(\hat{\theta}_N)$ décroît significativement si la taille de la base d'apprentissage augmente?
  ii) Quel est le meilleur modèle (dans une famille finie de modèles)?

Ce problème peut se poser dans le milieu industriel où les problèmes de collecte de données et notamment les coûts des campagnes d'acquisition de données nécessitent d'avoir des outils permettant d'estimer l'utilité de réaliser des collectes supplémentaires de données.

Les deux approches (estimation de la généralisation et règle GAE), que nous proposons maintenant, reposent sur une propriété obtenue en partant des équations (11) et (12) du corollaire 5 que nous rappelons ici :

Erreur moyenne en généralisation

$$J(\hat{\theta}_N) \approx Q(\theta^*) + \tfrac{1}{2}\tfrac{1}{N}\mathrm{tr}\left(\mathbf{B}(\theta^*)\mathbf{H}^{-1}(\theta^*)\right) \quad (11)$$

Erreur moyenne à l'apprentissage

$$E\left(Q_N(\hat{\theta}_N)\right) \approx Q(\theta^*) - \tfrac{1}{2}\tfrac{1}{N}\mathrm{tr}\left(\mathbf{B}(\theta^*)\mathbf{H}^{-1}(\theta^*)\right) \quad (12)$$

7 mais linéaire par rapport aux paramètres

d'où la propriété :

Propriété 13

Les équations (11) et (12) du corollaire 5 peuvent se mettre sous la forme suivante :

$$f = b + a * \frac{1}{N} \quad avec \quad a = \pm\frac{1}{2} tr\left(\boldsymbol{B}(\theta^*)\boldsymbol{H}^{-1}(\theta^*)\right) \quad , \quad b = Q(\theta^*)$$

Ce qui signifie que les erreurs moyennes à l'apprentissage sont distribuées selon une droite. Une propriété identique s'applique aux erreurs moyennes en généralisation mais avec une droite de pente opposée. Cette pente vaut $a = \pm\frac{1}{2} tr\left(\boldsymbol{B}(\theta^*)\boldsymbol{H}^{-1}(\theta^*)\right)$.

Ces deux droites se coupent au point $b = Q(\theta^*)$ *qui constitue l'erreur asymptotique du modèle (i.e. quand la base d'apprentissage est de taille infinie). Voir les figures 2 et 6a.*

Algorithme 1: Estimation de l'erreur moyenne en généralisation

La nouvelle approche que nous proposons pour estimer l'erreur moyenne en généralisation $J(\hat{\theta}_N)$ est contituée de trois étapes :

1. estimer les erreurs moyennes à l'apprentissage sur des bases d'apprentissage de tailles $(N_i)_{i=1}^m$;
2. effectuer une régression linéaire des erreurs moyennes à l'apprentissage, i.e. des points $\left\{\frac{1}{N_i}, E\left(Q_{N_i}(\hat{\theta}_{N_i})\right)\right\}_{i=1}^m$

   $$\text{éq (12)} \Longrightarrow E\left(Q_{N_i}(\hat{\theta}_{N_i})\right) \approx Q(\theta^*) - \frac{1}{2}\frac{1}{N_i}\text{tr}\left(\mathbf{B}(\theta^*)\mathbf{H}^{-1}(\theta^*)\right) = b_{app} + \frac{1}{N_i}a_{app}$$

   $\Longrightarrow$ détermination de a_{app} et b_{app}
3. estimer l'erreur moyenne en généralisation : d'après la propriété 13, les droites pour les erreurs d'apprentissage et de généralisation ont des pentes opposées ($a_{app} = -a_{gene}$) et mêmes ordonnées à l'origine ($b_{app} = b_{gene}$), d'où :

   $$\text{éq. (11)} \Longrightarrow \quad \boxed{J(\hat{\theta}_N) \approx b_{app} - \frac{1}{N}a_{app}}$$

Avec cette approche il est nécessaire de constituer des bases d'apprentissage $(A_i)_{i=1}^m$ pour effectuer la régression linéaire des erreurs moyennes à l'apprentissage[8]. Mais en contre partie il **n'est pas** nécessaire de garder des données pour évaluer les erreurs en généralisation des modèles appris sur ces bases d'apprentissage, puisque l'erreur en généralisation est donnée par la relation précédente.

Remarque 14

La propriété 13 est assez similaire mais aussi différente de la "predictive method" présentée par[Cortes et al 94] avec une approche mécanique statistique. Notre approche à base de développements de Taylor semble plus simple. Des travaux ultérieurs devraient permettre de comparer les domaines de validité des deux approches ainsi que les hypothèses posées. Par exemple, l'approche de [Cortes et al 94] n'est théoriquement applicable qu'à des modèles comprenant beaucoup de paramètres ; ce qui n'est pas le cas des travaux que nous présentons.

8. en fait pour chaque N_i, il est nécessaire de constituer plusieurs bases d'apprentissage de taille N_i pour obtenir une estimation correcte de l'erreur moyenne d'apprentissage pour chaque N_i ; par exemple en utilisant des méthodes de réchantillonnage (boostrap, jackknife).

Simulations numériques

Les simulations suivantes vont porter sur deux points :

1. valider la propriété 13 ;
2. valider l'algorithme 1 pour l'estimation de l'erreur moyenne en généralisation $J(\hat{\theta}_N)$;
3. de présenter et de valider la nouvelle règle graphique GAE de sélection de modèle

Régression linéaire polynomiale

Nous prenons l'exemple de régression polynomiale[9] présenté par la figure 1. Il s'agit d'approcher la fonction polynomiale $1-2*x^2-x^3$ sur l'intervalle [-2.5, 1.5]. Les bases de données sont composés de couples $z_i = (x_i, y_i)$ tels que :

$$y_i = E_{\mathbf{Y}}\left(\mathbf{X} = x_i\right) + n_i = (1 - 2 * x_i^2 - x_i^3) + n_i \quad \text{avec} \quad n \sim \mathcal{N}(0,1)$$

Nous allons considérer des modèles de complexité croissante dont la complexité est définie par le degré p du polynôme utilisé pour réaliser l'approximation. L'approximation optimale est donné par le polynôme de degré 3, i.e. constitué de 4 paramètres.

Nous allons utiliser des bases d'apprentissage $(A_i)_{i=1}^{10}$ de taille N_i = 20, 25, 30, 35, 40, 50, 100, 200, 400 ou 600 observations

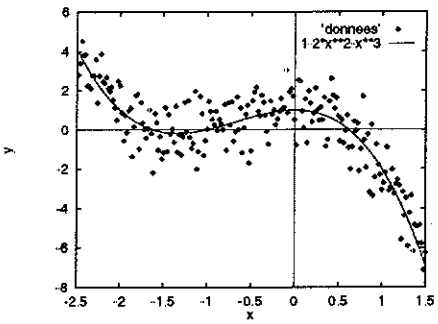


FIG 1 - *Base d'apprentissage bruitée (200 observations) et fonction originale* $g(x) = 1-2*x^2-x^3$

1 - Validation de la propriété 13.

Les réponses y_i ont été bruitées avec un bruit de variance $\lambda_0 = 1.0$. De ce fait, d'après la propriété 13, lorsque le modèle est correct, nous devrions trouver que les erreurs moyennes à l'apprentissage et en généralisation sont distribuées (en fonction de $1/N$) selon des droites :

- d'ordonnée à l'origine : $b = Q(\theta^*) = Q(\theta_0) = \lambda_0 = 1.0$
- de pente : $a = \pm\frac{1}{2}\mathrm{tr}\left(\mathbf{B}(\theta_0)\mathbf{H}^{-1}(\theta_0)\right) = \pm\lambda_0 p$, où p est le nombre de paramètres.

Dans le cadre des simulations, si le modèle est correct, nous constatons :

- Fig. 2a : que les erreurs moyennes à l'apprentissage et en généralisation sont distribuées selon des droites (en fonction de $1/N$) ;
- tableau 1, ligne grisée (= modèle correct) :
  - que l'erreur asymptotique[10] est pratiquement égale à la variance $\lambda_0 = 1.0$ du bruit ;
  - que la pente de la droite vaut $a = -4.18$ ou $4.24 \approx \pm\lambda_0 p = \pm 4$;

ces résultats valident la propriété 13.

9. linéaire par rapport aux paramètres et polynomiale par rapport aux entrées x
10. = ordonnées à l'origine des droites = colonnes b_{app} et b_{gene} du tableau 1

2 - Sélection de la complexité optimale: la règle GAE

Nous allons maintenant présenter une nouvelle règle graphique de sélection de modèle: **la règle GAE**. Cette nouvelle règle se déduit de l'observation des colonnes b_{gene} et b_{app} du tableau 1 représentant les erreurs asymptotiques des modèles

Le tableau 1 contient les ordonnées à l'origine et les pentes des droites de régression des erreurs moyennes à l'apprentissage et en généralisation; ceci pour des modèles de complexité croissante (ici des polynômes de degré 1 à 6).

degré du polynôme	nb. de paramètres	ordonnée à l'origine		pente de la droite		qualité de la régression
		b_{app}	b_{gene}	a_{app}	a_{gene}	q
1	2	2.84	2.86	7.67	2.50	1.00
2	3	2.48	2.50	-3.80	4.85	1.00
3	4	1.00	1.01	-4.18	4.24	1.00
4	5	1.00	1.01	-5.08	5.29	1.00
5	6	1.00	1.01	-6.25	6.59	1.00
6	7	1.00	1.01	-7.19	7.96	1.00

TAB 1 - *Valeurs des pentes et des ordonnées à l'origine des droites de régression des erreurs moyennes à l'apprentissage et en généralisation. La ligne grisée détermine le modèle correct, i.e de complexité optimale. En premier lieu, si le modèle est de complexité supérieure ou égale à la complexité optimale, nous constatons que les erreurs asymptotiques (= colonnes b_{app} et b_{gene}) sont égales à la variance $\lambda_0 = 1.0$ du bruit: le biais du modèle est pratiquement nul. Cependant plus le modèle est complexe, plus l'erreur en généralisation augmente car la pente (a_{gene}) devient plus forte.*

L'étude des colonnes b_{app} et b_{gene} du tableau 1 révèle que les erreurs asymptotiques restent égales à $\lambda_0 = 1.0$ dès que la complexité est supérieure ou égale à celle du modèle optimal: ce point est intéressant car il prouve que l'utilisation de modèles complexes permet d'atteindre l'erreur asymptotique minimale due au bruit. Par contre, si le modèle n'est pas suffisamment complexe, il est alors biaisé [Geman et al 92], ce qui se traduit par b_{app} et $b_{gene} >> \lambda_0$.

La sélection du modèle optimal est réalisée en traçant b_{app} en fonction du nombre p de paramètres du modèle; et en décidant à partir de quelle valeur de p les segments de droites joignant les points deviennent pratiquement horizontaux. Cette valeur de p est alors prise comme étant la complexité optimale. Dans le cadre de cette simulation, le modèle correct est un polynôme de degré 3: figure 3.

Algorithme 2: le Graphe de l'Erreur Asymptotique pour la sélection de modèle

Cette nouvelle règle graphique GAE de sélection de la complexité optimale (sous hypothèse d'une base d'apprentissage de taille infinie) est constituée de trois étapes.
Pour chaque modèle:

1. estimer les erreurs moyennes à l'apprentissage sur des bases d'apprentissage de tailles $(N_i)_{i=1}^m$;
2. effectuer une régression linéaire des erreurs moyennes à l'apprentissage, i.e. des points $\left\{\frac{1}{N_i}, E\left(Q_{N_i}(\hat{\theta}_{N_i})\right)\right\}_{i=1}^m$

   éq. (12) $\Longrightarrow E\left(Q_{N_i}(\hat{\theta}_{N_i})\right) \approx Q(\theta^*) - \frac{1}{2}\frac{1}{N_i}\mathrm{tr}\left(\mathbf{B}(\theta^*)\mathbf{H}^{-1}(\theta^*)\right) = b_{app} + \frac{1}{N_i}a_{app}$

   $\Longrightarrow$ on obtient l'ordonnée à l'origine b_{app} des droites (une droite est associée à un modèle de complexité fixée).
3. représenter l'ordonnée à l'origine b_{app} en tant que fonction de la complexité des modèles.

Cette représentation graphique est d'utilisation similaire au "test du coude[11]" introduit par [Cattel66] pour la sélection du nombre de composantes à garder lors d'une Analyse en Composantes Principales. Le "test du coude" consiste à tracer les valeurs propres l_k en fonction du rang k des composantes. Une autre méthode, populaire en météorologie, consiste à tracer $\log(l_k)$, et non l_k, en fonction de k ; cette méthode est connue sous le nom de "log-eigenvalue (ou LEV) diagram" [Craddock et al 69, Farmer71]. Pour une présentation de ces deux méthodes, se reporter à [Jolliffe82].

Notons aussi que le modèle de complexité optimale est celui dont les termes a_{gene} et b_{gene} vérifient : la pente a_{gene} (= -a_{app}) est minimale parmi les modèles d'erreur asymptotique b_{gene} (= b_{app}) minimale. Sur figure 2b sont tracées les droites définies par les termes a_{gene} et b_{gene} pour les modèles de complexité supérieure ou égales à celle du modèle optimal (et correct). Nous constatons que les droites ont la même ordonnée à l'origine[12] $b_{gene} = \lambda_0$: le modèle correct est celui de pente a_{gene} minimale, i.e. celui dont l'erreur en généralisation est minimale à N fixé et qui augmente le moins vite (en fonction de $1/N$).

Un autre résultat intéressant apparaît en examinant la colonne a_{gene} (ou a_{app}) du tableau 1 : dès que la complexité des modèles est supérieure ou égale à celle du modèle correct, la pente de la droite en généralisation[13] est approximativement proportionnelle au nombre p de paramètres du modèle : $a_{gene} \approx \lambda_0 p \approx -a_{app}$. Ce qui n'est visiblement pas le cas lorsque le modèle est biaisé (= pas assez complexe) ; de plus dans ce cas $a_{app} \neq -a_{gene}$.

Nous obtenons donc une réponse partielle à la question que nous nous posions à la remarque 12 : l'utilisation du critère FPE pour la sélection de modèles est justifiée pour des modèles de complexité supérieure ou égale à la complexité du modèle optimal et englobant ce dernier. Par contre l'utilisation du critère FPE pour estimer l'erreur en généralisation de modèles pas assez complexes et plus discutable car le modèle est biaisé et, de ce fait, l'estimation de la variance λ_0 par la relation (21) n'est plus valide.

3 - Nouvelles estimations de l'erreur en généralisation : algorithme 1

Dans le paragraphe précédent, nous avons étudié l'erreur asymptotique du modèle, i.e. quand la base d'apprentissage est de taille **infinie**. Notons cependant que la règle GAE est construite pour permettre la sélection du meilleur modèle indépendamment de la taille de la base d'apprentissage.

Maintenant, nous désirons sélectionner le meilleur modèle mais aussi connaître son comportement en phase d'utilsation, i.e. estimer l'erreur en généralisation du modèle lorsque la taille de la base d'apprentissage est **finie** (ce qui est toujours le cas en pratique). L'approche adoptée est celle de l'algorithme 1.

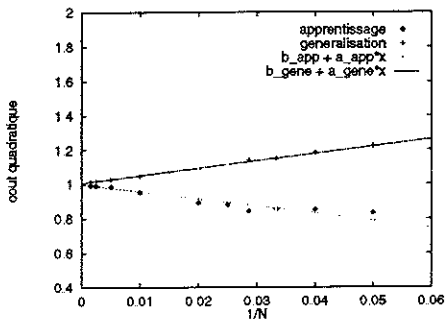
En effet, l'algorithme 1 va nous servir à estimer l'erreur moyenne en généralisation d'un modèle qui aurait été appris sur une base d'apprentissage de taille finie, par exemple $N = 50$ observations, à partir des erreurs moyennes à l'apprentissage obtenues lors de l'apprentissage du modèle sur des bases de "petites" tailles $(N_i)_{i=1}^m$ (Fig. 4 a).

Résultat : La figure 4a représente les erreurs obtenues sur des bases d'apprentissage de tailles $(N_i)_{i=1}^m = 20, 25, 30, 35, 50$ observations. Nous constatons qu'elles sont assez différentes. **Seules** les erreurs d'apprentissage pour $N_i = 20, 25, 30, 35$ servent à l'estimation de l'erreur moyenne en généralisation $J(\hat{\theta}_N)$ pour $N = 50$. On utilise l'algorithme 1 pour chaque modèle :

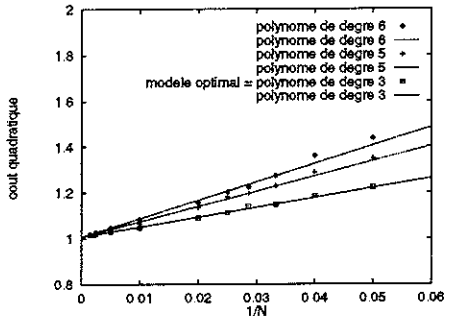
Pour chaque modèle :

1. estimer les erreurs moyennes à l'apprentissage pour $N_i = 20, 25, 30, 35$;
2. calculer la régression linéaire des erreurs moyennes à l'apprentissage : on obtient a_{app} et b_{app} ; en déduire la droite de régression des erreurs moyennes en généralisation ;
3. d'après cette dernière droite, en déduire une estimation de l'erreur en généralisation pour $N = 50$.

11. "Scree test" en anglais
12. notons qu'alors $b_{gene} = Q(\theta_0) = \lambda_0 = 1.0$
13. ou à l'apprentissage avec un signe opposé.



(a) Modèle correct: régression linéaire des erreurs moyennes à l'apprentissage et en généralisation.



(b) Erreurs en généralisation du modèle correct et de modèles trop complexes. Le modèle correct est représenté par la droite de pente optimale.

FIG 2 - *Les bases d'apprentissage $(A_i)_{i=1}^m$ utilisées pour la régression linéaire contiennent N_i = 20,25,30,35,40,50,100,200,400 ou 600 observations. Moyennes sur 50 tirages. Nous constatons que les erreurs à l'apprentissage et à la généralisation sont correctement ajustées par des droites. De plus, nous voyons que l'erreur en généralisation croît avec la complexité du modèle (Fig. (b)). Cette augmentation devient d'autant plus sensible que la base d'apprentissage est de taille réduite. L'erreur asymptotique ($N \to \infty$) vaut 1.0, ce qui correspond à la variance λ_0 du bruit sur les réponses y.*

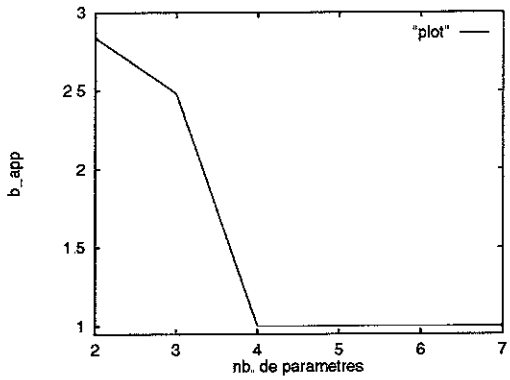
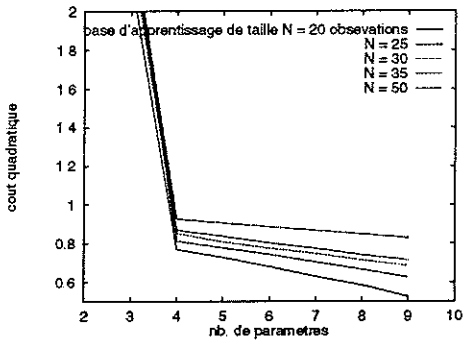


FIG. 3 - *La règle GAE: b_{app} en fonction du nombre de paramètres. Le modèle optimal contient 4 paramètres.*

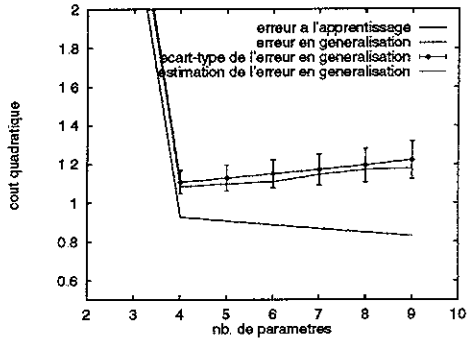
Trois courbes sont représentées sur la figure 4b :

- l'erreur (décroissante) à l'apprentissage ;
- l'erreur en généralisation (en fait estimée sur une grosse base de données) ;
- l'estimation de l'erreur en généralisation par l'algorithme 1.

Nous constatons que l'algorithme 1 fournit une estimation de très bonne qualité puisque la différence reste inférieure à l'écart-type [14]



(a) Erreurs moyennes à l'apprentissage pour des bases d'apprentissage de tailles variables. L'erreur à l'apprentissage augmente avec N et décroît avec la complexité du modèle.



(b) $J(\hat{\theta}_N) \approx b_{app} - a_{app}/N$

FIG. 4 - *À partir des erreurs moyennes à l'apprentissage du modèle appris sur des bases d'apprentissage $(A_i)_{i=1}^m$ de taille $N_i = 20, 25, 30, 35$ observations (Fig. (a)), nous pouvons en déduire des estimations de l'erreur moyenne en généralisation $J(\hat{\theta}_N)$ que nous aurions obtenue si ce modèle avait été appris sur une base d'apprentissage de taille $N = 50$ observations. Moyennes sur 100 tirages.*

Régression neuronale

Nous prenons un problème de régression plus complexe que celui étudié précédemment. Comme exemple de modèle non-linéaire, nous utilisons un réseau neuronal de type Perceptron Multi-Couches comportant une seule couche cachée. Les cellules cachées et la cellule de sortie utilisent la fonction sigmoïde à valeurs dans [-1,1] comme fonction de passage.

Nous désirons déterminer le nombre minimal de cellules cachées pour réaliser une approximation de la fonction de régression $g(x)$ sur l'intervalle [-2.5,1.5]

Les bases d'apprentissage sont constituées de N couples $(x_i, y_i)_{i=1}^N$ dont les réponses y_i sont obtenues en bruitant la fonction $g(x)$:

$$y_i = g(x_i) + n \quad \text{avec} \quad n \sim \mathcal{N}(0, 0.04)$$

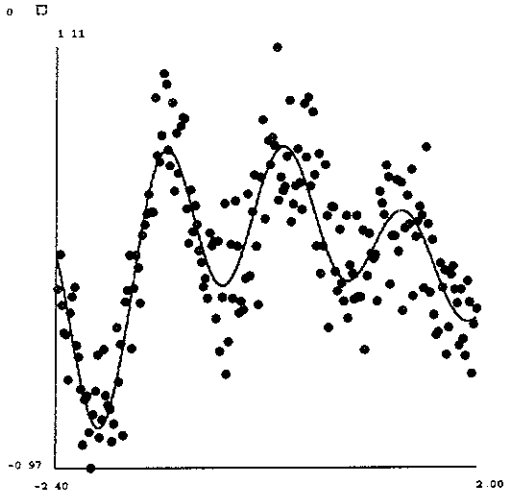
Des réseaux neuronaux PMC comportant un nombre croissant de cellules cachées sont appris sur des bases de taille $N_i = 20, 25, 30, 40$ ou 50 observations. Nous calculons alors les erreurs à l'apprentissage et en généralisation [15].

1 - Validation la propriété 13

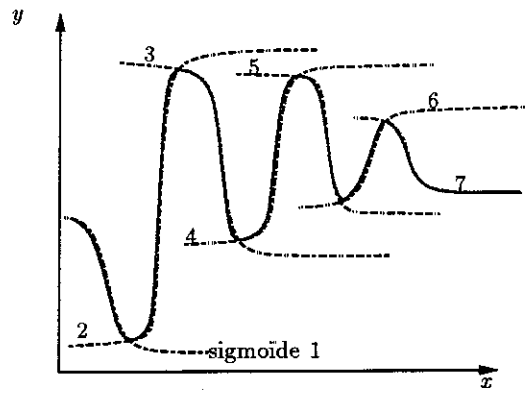
D'après le tableau 2, des modèles non-linéaires tels que les réseaux neuronaux PMC vérifient la propriété 13 : les erreurs à l'apprentissage et en généralisation sont bien ajustées par des droites (en fonction de $1/N$).

14. pour un modèle et une taille de base d'apprentissage fixés, la moyenne et l'écart-type de l'erreur en généralisation sont estimés en apprenant le modèle sur 100 bases d'apprentissage (de mêmes tailles) et en calculant les erreurs en généralisation sur une base de test de taille suffisante pour permettre des estimations correctes de cette erreur. Se reporter à la section 3.1.

15. les erreurs en généralisation sont évaluées en testant les modèles sur une base de généralisation comportant 1000 observations.



(a) Base d'apprentissage bruitée (200 observations) et fonction originale
$g(x) = 30 * cos(2x) * cos(3x)/(x+7)^2$



(b) Approximation grossière de la courbe avec 7 sigmoïdes: une bosse est approximativement approchée par deux sigmoïdes. Cette approximation "avec les mains" est seulement utilisable pour une fonction d'une seule variable.

FIG. 5 - *Exemple de régression neuronale*

nombre de cellules cachées	nombre de paramètres	ordonnée à l'origine		pente de la droite		qualité de la régression	
		b_{app}	b_{gene}	a_{app}	a_{gene}	app.	gene.
2	7	0.11	0.09	-1.14	0.13	0.91	0.47
3	10	0.08	0.08	-0.76	0.25	0.93	0.98
4	13	0.07	0.07	-0.72	0.35	0.91	1.00
5	16	0.07	0.07	-0.95	0.32	0.98	1.00
6	19	0.06	0.06	-1.00	0.10	0.90	0.95
7	22	0.04	0.04	-0.78	0.65	0.98	1.00
8	25	0.04	0.04	-0.88	0.66	0.97	1.00
9	28	0.04	0.04	-0.79	0.76	0.98	1.00
10	31	0.04	0.04	-0.80	0.79	0.98	1.00
11	34	0.04	0.04	-0.82	0.79	0.98	1.00
12	37	0.04	0.04	-0.79	0.84	0.91	1.00
13	40	0.04	0.04	-0.80	0.86	0.94	1.00

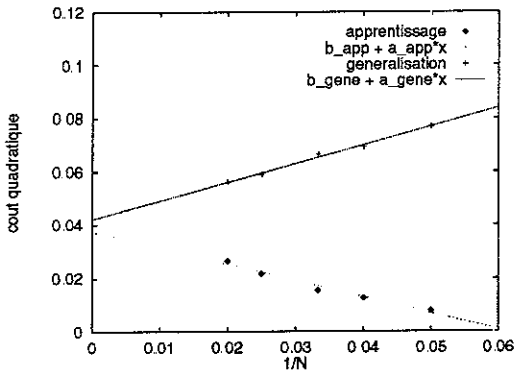
TAB. 2 - *Valeurs des pentes et des ordonnées à l'origine des droites de régression des erreurs moyennes à l'apprentissage et en généralisation. La ligne grisée détermine le modèle optimal, i.e. de complexité optimale. En premier lieu, si le modèle est de complexité supérieure ou égale à la complexité optimale, nous constatons que les erreurs asymptotiques (= colonnes b_{app} et b_{gene}) sont égales à la variance $\lambda_0 = 0.04$ du bruit: le biais du modèle est pratiquement nul. Cependant plus le modèle est complexe, plus l'erreur en généralisation augmente (colonne a_{gene}) (voir le texte pour plus de détails). La qualité de la régression est donnée par q (voir [Press et al.90] p. 525).*

Par exemple, si nous traçons ces droites pour le modèle optimal (Fig. 6a), nous constatons que l'erreur quadratique asymptotique est de l'ordre de 0.04, ce qui correspond à la variance du bruit sur les réponses y_i.

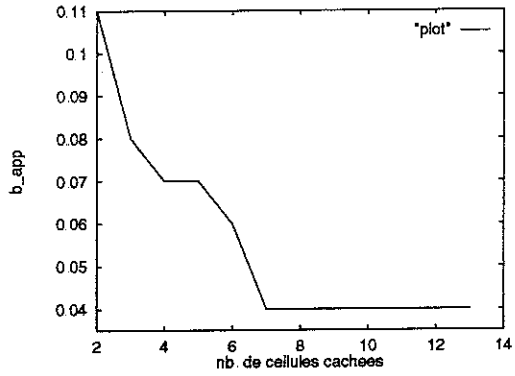
2 - Sélection de la complexité optimale: la "règle GAE"

Nous utilisons la représentation graphique de "l'Erreur Asymptotique" pour sélectionner le modèle optimal: figure 6b

Nous recherchons le point au delà duquel le graphe de l'erreur asymptotique devient plus ou moins horizontal; ce point donnant le modèle optimal i.e. d'erreur minimale en généralisation Pour cet exemple, le modèle optimal comporte 7 cellules cachées.



(a) Régression linéaire des erreurs moyennes à l'apprentissage et en généralisation pour le modèle optimal (7 cellules cachées). Les bases utilisées pour la régression linéaire contiennent 20, 25, 30, 40 ou 50 observations. Moyennes sur 100 tirages



(b) La règle GAE : b_{app} en fonction du nombre de cellules cachées. Le modèle optimal contient 7 cellules cachées.

FIG. 6 - *Sélection du bon nombre de cellules cachées d'un modèle neuronal du type Perceptron Multi-Couches.*

5 Conclusion

L'ensemble des résultats présentés reposent sur des estimations des erreurs moyennes à l'apprentissage et en généralisation. Nous avons alors constaté, propriété 13, que ces deux erreurs sont liées: elles sont distribuées selon des droites de pentes opposées et d'ordonnées à l'origine identiques, en fonction de $1/N$, où N est la taille de la base d'apprentissage.

Ces deux estimations nous ont aussi permis de comprendre le principe de parcimonie pour la sélection d'un modèle: plus celui-ci est complexe, plus l'erreur en généralisation augmente; mais aussi d'expliciter la forme du terme de pénalité (corollaire 5) qui permet de construire des règles de sélection de modèles (critères FPE, AIC).

En ce qui concerne la sélection de modèle, ces deux estimations sont à l'origine de plusieurs résultats originaux:

- appliquées à la fonction d'erreur quadratique (ou log-vraisemblance), elles permettent de dériver le critère FPE (ou AIC) pour des modèles linéaires ou non linéaires;
- de nouvelles approches utilisant la propriété 13 ont été proposées pour sélectionner la bonne structure de modèle:
  - soit l'estimation de l'erreur en généralisation de modèles de complexités croissantes,
  - soit la "règle GAE" qui représente le Log. de l'Erreur Asymptotique des modèles en fonction de leur complexité (figures 3 et 6b).

Dans la pratique, pour sélectionner le meilleur modèle, deux approches sont possibles :

1 algorithme 2 : on veut juste sélectionner le meilleur modèle : on utilise alors la règle de sélection GAE ;

2 algorithme 1 : on veut sélectionner le meilleur modèle et aussi connaître l'erreur qu'il obtiendra sur des données en phase d'utilisation, i e estimer l'erreur en généralisation du modèle lorsque celui-ci est appris sur une base d'apprentissage de taille N

Références

[Akaike74] Akaike (H) – A new look at the statistical model identification *IEEE Trans. Automatic. Control*, vol. 19, 1974, pp 716–723

[Amemiya83] Amemiya (T.) – *Handbook of econometrics*, chap Non-linear regression models, pp. 333–389 – Amsterdam, Z Griliches and M. D Intriligator (Eds); North-Holland, 1983volume I

[Antoniadis et al 92] Antoniadis (A.), Berruyer (J) et Carmona (R) – *Régression non linéaire et applications*. – Economica, 1992, *Collection Économie et Statistiques avancées Série : École Nationale de la Statistique et de l'Administration Économique et du Centre d'Étude des Programmes Économiques*.

[Cattel66] Cattel (R. B). – The scree test for the number of factors *J. Multi. Behav Res.* [Jolliffe82], pp. 245–276.

[Cortes et al.94] Cortes (C), Jacquel (L D), Solla (S A) et Vapnik (Vladimir). – Learning curves: Asymptotic values and rate of convergence. *NIPS*, éd. par Moody (John E.), Hanson (Steven J) et Lippmann (Richard P.), pp 327–334 – 1994. generalization error, model selection

[Craddock et al.69] Craddock (J M.) et Flood (C R). – Eigenvectors for representing the 500 mb geopotential surface over the northern hemisphere. *Q. J R Met. Soc.* [Jolliffe82], pp 576–593.

[Efron et al.86] Efron (B.) et Tibshirani (R J.) – Boostrap methods for standard errors, confidence intervals and other measures of statistical accuracy *Statistical Science*, vol 1, 1986, pp 54–77

[Farmer71] Farmer (S A.) – An investigation into the results of principal component analysis of data derived from random numbers *Statistician* [Jolliffe82], pp. 63–72.

[Gallant75] Gallant (A.R) – Nonlinear regression *Am. Stat.*, vol 29, n° 2, 1975, pp 73–81.

[Geman et al 92] Geman (S), Bienenstock (E) et Doursat (R) – Neural networks and the bias/variance dilemma *Neural Computation*, vol 4, n° 1, 1992, pp 1–58

[Jennrich69] Jennrich (R I.) – Asymtotic properties of nonlinear least squares estimators *Ann Math Stat.*, vol. 40, 1969, pp 633–643

[Jolliffe82] Jolliffe (I T.) – *Principal Component Analysis* – Springer-Verlag, 1982, *Series in Statistics*.

[Kannurpatti et al 91] Kannurpatti (R) et Hart (G.W.) – Memoryless nonlinear system identification with unknown model order *IEEE Trans. Information Theory*, vol 37, n° 5, Sept 1991, pp 1440–1450

[Malinvaud70a] Malinvaud (E.) – The consistency of non linear regression. *Ann. Math. Stat.*, vol. 41, 1970, pp. 956–969.

[Malinvaud70b] Malinvaud (E.) – *Statistical methods of econometrics* – Amsterdam, North Holland, 1970.

[Press et al. 90] Press (W. H.), Flannery (B. P.), Teukolsky (S. A.) et Vetterling (W. T.) – *Numerical Recipes in C. The art of scientific computing.* – Cambridge University Press, 1990.

[Seber et al. 89] Seber (G. A. F.) et Wild (C. J.) – *Nonlinear regression* – Wiley series in probability and mathematical statistics, 1989.

[Stone74] Stone (M.) – Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, vol. Ser. B, n° 36, 1974, pp. 111–147.

[White81] White (H.) – Consequences and detection of misspecified nonlinear regression models. *Journal of the American statistical association*, vol. 76, 1981, pp. 419–433.

[White82] White (H.) – Maximum likelihood estimation of misspecified models. *Econometrica*, vol. 50, 1982, pp. 1–25.

[White92] White (H.) – *Artificial Neural Networks, Approximation and Learning Theory* – Blackwell, 1992.

[Wu81] Wu (C. F.) – Asymptotic theory of nonlinear least squares estimation. *Ann. Stat.*, vol. 9, 1981, pp. 501–513.