Modèles de réseaux de neurones pour l'analyse des séries temporelles ou la régression : Estimation, Identification, Méthode d'élagage SSM

J. Rynkiewicz, M. Cottrell, M. Mangeas, J.F. Yao

SAMOS, Université de Paris 1
90, rue de Tolbiac
75013 Paris
cottrell@univ-paris1.fr

RÉSUMÉ. *Ce papier porte sur la modélisation de séries temporelles ou de régression à l'aide de réseaux de neurones. En nous appuyant sur des résultats récents sur l'estimation des moindres carrés pour les séries temporelles non linéaires, nous proposons une méthodologie complète et explicite pour l'estimation des paramètres (processus d'apprentissage) et pour le choix du modèle (sélection d'architecture). En particulier, nous donnons une solution au problème de l'élagage dans un perceptron multi-couches au moyen d'une méthode pas à pas utilisant un critère de type BIC dont on démontre la consistance.*

ABSTRACT. *This paper deals with neural network modeling for time series analysis or regression. Based on recent results about the least-square estimation for non-linear time series, we propose a complete and feasible methodology for both parameter estimation (learning process) and model selection (architecture selection). In particular, we solve the pruning problem for multilayer perceptron models with a stepwise search method by using a BIC criterion which is proved to be consistent.*

MOTS-CLÉS : *Identification statistique, statistiques asymptotiques*

KEYWORDS: *Statistical Stepwise, Almost sure identification, asymptotic statistics*

1. Introduction

Les perceptrons multi-couches (MLP) ont d'abord été introduits pour résoudre des problèmes complexes de classification. Mais en raison de leur propriété d'approximateur universel (voir Hornik, 1989, [HOR 89] ou Funahashi, 1989, [FUN 89]), ils ont été rapidement utilisés comme modèles de régression non linéaire, et ensuite pour la modélisation des séries temporelles et la prévision. Voir par exemple (Weigend et Gershenfeld, 1994, [WEI 93] ou Cottrell et al., 1995, [COT 95]) pour les références.

Cependant, l'estimation et l'identification de ces modèles utilisent des techniques sophistiquées et il n'est pas facile de déterminer l'architecture adéquate. En effet, ces modèles sont par définition sur-paramétrés, les fonctions d'erreur à minimiser ont de nombreux minima locaux, et la mise en oeuvre s'avère souvent délicate.

De nombreux articles portent sur les techniques d'*élagage* des paramètres inutiles, en particulier dans le cadre des modèles de régression, et les utilisateurs ont étendu les techniques proposées au cas des séries temporelles. Voir par exemple (Le Cun et al., 1990, [CUN 90], Moody, 1992, [MOO 92], Reed, 1993, [REE 93], Murata et al., 1994, [MUR 94], etc.). La plupart de ces papiers fournissent des heuristiques, mais ne se placent pas dans un cadre statistique rigoureux.

Ce papier propose un ensemble de résultats théoriques établis dans le cadre de modèles neuronaux de séries temporelles, qui étendent des résultats connus dans le cadre des modèles statistiques linéaires (Hannan et Deistler [HAN 88]). En fait ces résultats sont également valables dans le cadre des modèles de régression et des modèles mixtes (modèles auto-régressifs comprenant aussi des variables exogènes), mais pour simplifier l'exposé, nous nous plaçons uniquement dans le cadre auto-régressif.

Nous considérons donc une famille de modèles appelés *Neural AutoRegressive model* (*NAR*), définis par :

$$Y_t = f_W(Y_{t-1}, \ldots, Y_{t-p}) + \varepsilon_t \qquad [1]$$

où

– $Y_t \in \mathbb{R}$, mais on peut généraliser au cas multidimensionnel,

– f_W représente une fonction implémentée par un perceptron multi-couches avec une seule unité de sortie,

– $Y_{t-i}, i = 1, \ldots, p$ sont les retards de la série (Y_t),

– ε_t est un bruit i.i.d., d'espérance 0, de variance constante σ^2, par exemple une variable $\mathcal{N}(0, \sigma^2)$, indépendante du passé de la série.

On considère dans la suite un (p, K)-perceptron multi-couches, avec une unité de sortie linéaire, p unités d'entrée linéaires, K unités cachées munies d'une fonction d'activation sigmoïde ϕ de type tangente hyperbolique (fonction impaire).

Alors un modèle (*NAR*) est défini précisément par une équation du type

$$Y_t = f_W(Y_{t-1}, \ldots, Y_{t-p}) + \varepsilon_t = \alpha_0 + \sum_{j=1}^{K} \alpha_j \phi \left(\sum_{i=1}^{p} \beta_{ij} Y_{t-i} + \beta_{0j} \right) + \varepsilon_t \qquad [2]$$

où les hypothèses sont celles de l'équation (1). Les notations sont les notations usuelles :

$$\beta_{ij}, 1 \leq i \leq p, 1 \leq j \leq K$$

est le paramètre correspondant au poids de la connexion entre l'unité d'entrée i et l'unité cachée j,

$$\alpha_j, 1 \leq j \leq K$$

correspond au poids de la connexion entre l'unité cachée j et l'unité de sortie,

$$\beta_{0j}, 1 \leq j \leq K$$

est la constante associée à l'unité cachée j et α_0 est la constante correspondant à l'unité de sortie.

L'équation (2) définit alors un modèle paramétrique ayant une forme fonctionnelle particulière. Nous présentons

– des propriétés asymptotiques des estimateurs des paramètres, pour un modèle donné et un nombre donné T d'observations, ce qui correspond à l'*apprentissage*,

– une méthodologie statistique permettant de choisir le meilleur modèle, (à partir d'un modèle dominant), ce qui correspond au choix de l'*architecture* du réseau.

Nous ne considérons ici que des sorties réelles, mais toutes les propriétés que nous présentons peuvent être étendues aisément au cas multi-dimensionnel.

Soit $W = \{(\alpha_j)_{0 \leq j \leq K}, (\beta_{ij})_{0 \leq i \leq p, 1 \leq j \leq K}\}$ le vecteur paramètre. Notons que sa dimension est $m = (p+2)K + 1$. Etant données $T + p$ observations,

$$(Y_{-p+1}, \ldots, Y_0, Y_1, \ldots, Y_T)$$

de la série, on estime W en minimisant la moyenne des carrés résiduels (*Fonction d'erreur*)

$$S_T(W) = \sum_{t=1}^{t=T} \left(Y_t - f_W(Y_{t-1}, \ldots, Y_{t-p}) \right)^2. \qquad [3]$$

On note

$$\hat{W}_T = \arg min_W S_T(W)$$

l'estimateur des moindres carrés de W.

Si on suppose que $Var(\varepsilon) = \sigma^2 > \mu > 0$, il est équivalent de minimiser

$$LV_T(W) = \ln(\sum_{t=1}^{t=T} (Y_t - f_W(Y_{t-1}, \ldots, Y_{t-p}))^2) = \ln(S_T(W)). \qquad [4]$$

Dans le cas où ε_t est gaussien, $LV_T(W)$ est exactement ce que les statisticiens appellent la log-vraisemblance concentrée.

On notera

$$\tilde{W}_T = \arg min_W \ln LV_T(W).$$

Les deux minimisations sont équivalentes (en fait $\tilde{W} = \hat{W}$, mais nous les distinguons pour rappeler comment ils ont été obtenus), leurs propriétés asymptotiques sont presque les mêmes, mais nous verrons qu'en pratique, il vaut mieux utiliser la seconde.

Le calcul des estimateurs peut être mené en utilisant n'importe quelle méthode de minimisation. Dans ce papier, nous ne traitons pas ce problème et on suppose que $\hat{W}_T$ est le vrai minimum de la fonction Erreur $S_T(W)$. Dans le cas général, non nécessairement gaussien, l'estimateur des moindres carrés est un cas particulier d'*estimateur de minimum de contraste* (voir Guyon, 1995, [GUY 95]).

Le papier est organisé comme suit : le paragraphe 2 énonce certains résultats mathématiques sur les estimateurs des moindres carrés pour un modèle *NAR* . Le paragraphe 3 fournit des résultats théoriques d'identification presque sûre. Dans le paragraphe 4, on présente une nouvelle méthode d'identification presque sûre du vrai modèle basée sur une méthode de descente pas à pas (*Stepwise Statistical Method, SSM*) qui permet d'élaguer les paramètres inutiles. Enfin, dans le paragraphe 5, on montre sur un exemple simulé comment cette méthode s'utilise. Le dernier paragraphe contient les conclusions et les perspectives.

2. Résultats théoriques

2.1. *L'identifiabilité du MLP à une couche cachée et une sortie*

Pour estimer le vecteur paramètre, une propriété fondamentale permettant d'obtenir des résultats de consistance, est l'identifiabilité du modèle. Cela signifie, que pour une fonction représentable par un MLP donné, il n'y a qu'un seul vecteur paramètre qui représente cette fonction.

Si on considère qu'un perceptron multicouches de dimension (p, K), c'est-à-dire avec p entrées, K unités cachées et une sortie, est une fonction paramétrique sur $\mathbb{R}^m$, avec $m = (p+2) \times K + 1$, le modèle n'est pas identifiable. On peut en effet trouver deux systèmes de paramètres différents qui génèrent les mêmes sorties. Ceci peut être obtenu, par exemple, si deux unités cachées ont des poids en amont strictement

identiques puisque qu'il suffira que la somme de leurs poids en aval soit constante pour représenter la même fonction.

Cependant si on se restreint à un ensemble de paramètres raisonnable les MLP deviennent identifiables, à des opérations triviales près. On parlera alors d'identifiabilité faible.

Nous donnons dans la suite des conditions nécessaires et suffisantes pour que le modèle soit identifiable, au sens faible, dans le cas d'un MLP à une couche cachée, ayant des tangentes hyperboliques pour fonctions d'activation ϕ.

2.1.1. *MLP irréductibles*

Une première précaution à prendre est de ne pas utiliser artificiellement trop d'unités cachées. Par exemple, si on considère un MLP avec une unité cachée dont le poids relié à la sortie est nul, celle-ci est totalement inutile puisque les poids en entrées à cette unité sont quelconques, c'est-à-dire qu'ils n'influencent pas la fonction représentée par ce MLP. Nous allons donc caractériser les MLP n'ayant pas d'unités cachées inutiles de manière manifeste par la notion de MLP irréductible.

Notation 1 *Si $X = (X_1, \cdots, X_p)^T \in \mathbb{R}^p$ est un vecteur d'entrée, on note $\nu_j(X)$ l'impulsion de la j-ème unité cachée :*

$$\nu_j(X) = \beta_{0j} + \sum_{i=1}^{p} \beta_{ij} X_i$$

La dimension p est fixée. Le MLP avec K unités cachées est associé à K applications affines :

$$\begin{array}{c} \mathbb{R}^p \to \mathbb{R} \\ X \longmapsto \nu_j(X) \end{array} .$$

On dira que deux fonctions affines ν_1, ν_2 sont "signe-équivalentes" si $\nu_1 = \nu_2$ ou $\nu_1 = -\nu_2$.

On dira qu'un MLP est réductible si et seulement si il vérifie au moins une des conditions suivantes :

1. Il existe un poids de sortie nul.
2. Il existe au moins deux indices différents $j_1, j_2 \in \{1, \cdots, K\}$ tels que les fonctions ν_{j_1}, ν_{j_2} soient signe-équivalentes
3. Il existe au moins un indice $j \in \{1, \cdots, \}$ tel que la fonction ν_j soit constante.

On dira qu'un MLP est irréductible si ce n'est pas un MLP réductible.

Notation 2 *On note $\mathcal{N}_{p,K}$ l'ensemble des MLP avec p entrées, K unités cachées, qui sont irréductibles.*

Remarque 1 *Si $K = 0$, $\mathcal{N}_{p,0}$ sont les fonctions affines de $\mathbb{R}^p \longrightarrow \mathbb{R}$.*

2.1.2. *Identifiabilité des MLP irréductibles*

Dorénavant, on ne considère plus que des MLP irréductibles, c'est-à-dire ne comprenant pas d'unités cachées manifestement inutiles ou faisant exactement double emploi. Remarquons que cette restriction n'empêche pas les MLP d'être sur-paramétrés. Il reste des transformations triviales qui ne changent pas la fonction représentée par le MLP. Par exemple, si on choisit l'unité cachée j, que l'on change le signe de tous les poids β_{ij} pour $1 \leq i \leq p$ et que l'on change aussi le signe du poids de sortie α_j associé à cette unité cachée, comme la fonction tangente hyperbolique est impaire, cela ne changera pas la fonction F_W. Notons $\zeta_j(MLP)$ le MLP résultant de cette transformation.

Une autre possibilité est d'interchanger deux unités cachées j_1 et j_2, ainsi que les poids correspondants, on note $\eta_{j_1,j_2}(MLP)$ le MLP correspondant. On dira que deux MLP M_1 et M_2 sont équivalents ($M_1 \mathcal{R} M_2$), si et seulement si il existe une transformation ψ, qui soit la composée d'applications ζ_j et η_{j_1,j_2} telle que $M_1 = \psi(M_2)$. Pour que les fonctions MLP soient identifiables, on considère donc les classes de MLP équivalents. Il s'agit ici d'une identifiabilité "faible" c'est-à-dire à une transformation ψ près. Sussmann (cf [SUS 92]) prouve alors que :

Théorème 1 *Les classes de MLP irréductibles équivalents sont identifiables i.e.*

$$f_W = f_{W'} \Leftrightarrow W \mathcal{R} W'.$$

En conclusion, on peut dire que si on se place dans l'espace des MLP irréductibles, à une fonction ne correspond qu'une classe de MLP, c'est-à-dire qu'il n'y a qu'un nombre fini de MLP qui représentent cette fonction et que l'on sait tous les écrire à partir de l'un d'entre eux. L'intérêt de considérer des MLP irréductibles est que cela évite que la matrice hessienne de la fonction d'erreur soit non inversible sur un ensemble de paramètres d'intérieur non vide. On peut ne considérer que des MLP irréductibles en imposant des contraintes comme "la valeur absolue d'un poids reliant une unité cachée à la sortie reste toujours supérieure à un millième", et de même pour les autres conditions.

Mais en fait dans la pratique, puisque on choisit les poids initiaux au hasard, la probabilité que se vérifie l'une des trois conditions de *réductibilité* est extrêmement faible. Les remarques précédentes sont importantes pour les résultats théoriques qui suivent.

On obtient aussi des résultats semblables si le nombre de sorties est quelconque (cf la thèse de J. Rynkiewicz [RYN 00]).

2.2. *Propriétés statistiques*

Considérons le modèle *NAR* défini par l'équation (2). Nous noterons $(w_i)_{1 \leq i \leq m}$ où $m = (p+2)K + 1$ les composantes du vecteur paramètre W et W_0 sa vraie valeur (inconnue).

Récemment, (Mangeas et Yao, 1997, [MAN 97d]) ont étudié les propriétés asymptotiques de l'estimateur des moindres carrés pour un processus *NAR* général. Leur approche est principalement basée sur la théorie de la stabilité pour les chaînes de Markov (Duflo, 1990, [DUF 90]; Meyn and Tweedie, 1993, [MEY 94]).

Soit $Y^{(p)} = (Y_t^{(p)})_{t>0}$ le *processus vectoriel*, défini par $Y_t^{(p)} = (Y_t, \ldots, Y_{t-p+1})$ pour $t > 0$. La suite $(Y_t^{(p)})$ est une chaîne de Markov homogène sur l'espace $\mathbb{R}^p$. Le vecteur $(y_1, \ldots, y_p) \in \mathbb{R}^p$ est noté $\tilde{y}$. Les hypothèses ci-dessous entraînent la stabilité de la chaîne $Y^{(p)}$ et en particulier, cette chaîne a une distribution invariante unique μ_{W_0}.

Dans l'article (Mangeas et Yao, 1997, [MAN 97d]), les auteurs montrent le théorème suivant :

Théorème 2 Consistance forte et normalité asymptotique *Pour le modèle (2), avec $\phi(x) = \tanh(x)$, supposons que:*

1. $(\varepsilon_t)_{t>0}$ est une suite i.i.d., indépendante des états initiaux $(Y_{-p+1}, \ldots, Y_0)$, telle que $E\ \varepsilon_t^6 < \infty$,

2. W appartient à un sous-ensemble compact $\mathcal{W}$ de l'espace euclidien $\mathbb{R}^m$ de dimension m, tel que $W_0 \in \overset{\circ}{\mathcal{W}}$ (intérieur de $\mathcal{W}$).

3. (Condition d'identifiabilité) Pour tout W différent de W_0, $f_W \neq f_{W_0}$ dans le sens qu'il existe un $\tilde{y} \in \mathbb{R}^p$ tel que $f_W(\tilde{y}) \neq f_{W_0}(\tilde{y})$.

4. La matrice de dimension $m \times m$

$$\Sigma_0 = \int_{\mathbb{R}^p} \left[\frac{\partial}{\partial w_i} f_W(\tilde{y}) \frac{\partial}{\partial w_j} f_W(\tilde{y})\right]_{1 \leq i,j \leq m} \mu_{W_0}(d\tilde{y}), \qquad [5]$$

est définie positive.

Alors

(a) L'estimateur $\hat{W}_T$ est fortement consistant, c'est-à-dire qu'il converge presque sûrement vers W_0 quand T tend vers $+\infty$.

(b) Indépendamment de la distribution initiale de la chaîne de Markov $Y^{(p)}$, le terme $\sqrt{T}\left[\hat{W}_T - W_0\right]$ converge en loi vers la distribution gaussienne multidimensionnelle $\mathcal{N}(0, \sigma^2 \Sigma_0^{-1})$.

En ne considérant que des MLP irréductibles, on assure que les hypothèses 3 et 4 sont vérifiées.

Remarquons que ces résultats permettent de construire des intervalles de confiance et des tests d'hypothèse sur la nullité des paramètres (voir le paragraphe 4). La variance résiduelle σ^2 est estimée en pratique par $\hat{\sigma}^2 = \frac{1}{T} S_T(\hat{W}_T)$, et la matrice Σ_0 par $\hat{\Sigma}_0 = \frac{1}{2T} \nabla^2 S_T(\hat{W}_T)$ qui peut aussi être approchée par

$$\frac{1}{T}\sum_{t}[\nabla f_{\hat{W}_T}(Y_t^{(p)})][\nabla f_{\hat{W}_T}(Y_t^{(p)})]'.$$

Il faut remarquer que les hypothèses sont très faibles par rapport aux hypothèses habituelles de normalité.

On peut montrer un théorème analogue pour l'estimateur $\tilde{W}_T$, mais alors il faut supposer que le bruit a un moment fini d'ordre 12, c'est-à-dire que $E(\varepsilon_t^{12}) < \infty$ (voir Rynkiewicz [RYN 00]).

3. Identification presque sûre

3.1. *La capacité de généralisation*

Une des principales difficultés intervenant dans l'utilisation de fonctions de plus en plus complexes pour l'estimation statistique des processus est le phénomène de sur-apprentissage. Si on utilise un modèle trop complexe sur trop peu de données, on aboutit à la modélisation du bruit qui a engendré les données sur lesquelles on estime le modèle. On introduit ainsi un biais dans le modèle qui compromet fortement la validité de ses résultats sur de nouvelles données du même processus. On dit alors que le modèle "généralise" mal. Il est donc apparu fondamental de contrôler la complexité du modèle pour assurer que l'erreur de celui-ci reste faible, non seulement sur les données que l'on observe, mais aussi sur des données futures, non encore observées, provenant du même phénomène.

Vapnik [VAP 95] propose par exemple d'utiliser le principe de minimisation du risque structurel. Ce principe définit un compromis entre la qualité d'approximation et la complexité des fonctions d'approximation. Cependant, le principal inconvénient des bornes établies par Vapnik est qu'elles ne sont valables que pour des variables aléatoires indépendantes identiquement distribuées. Or dans cet article, nous traitons des séries temporelles et nous ne sommes pas dans un cadre i.i.d. De plus, la dimension de Vapnik-Chernovensky n'est connue que pour les MLP ayant des fonctions indicatrices sur la couche cachée. Dans le cas où les fonctions d'activation sont des tangentes hyperboliques, il n'existe que des bornes supérieures de cette dimension. C'est pourquoi, pour réduire la sur-paramétrisation de nos modèles, nous utiliserons plutôt un terme de pénalisation qui dépend du nombre de paramètres et du nombre de données. La philosophie d'une telle approche (principe de parcimonie) est assez similaire au principe du SRM, mais le cadre théorique est différent, ainsi que les résultats qui en découlent.

3.2. *Identification presque-sûre*

Nous supposons ici que nous avons le choix entre plusieurs modèles pour expliquer le processus observé. Comment choisir convenablement un modèle? Le choix doit être parcimonieux (le moins de paramètres possibles) mais fournissant un bon ajustement (des paramètres en nombre suffisant). En utilisant les théorèmes existant sur la sélection de modèles au moyen de contrastes pénalisés (Guyon, 1995, [GUY 95]), nous montrons l'identification presque sûre, grâce à un terme de pénalisation, lorsqu'il y a un nombre fini de modèles possibles, ayant tous un modèle dominant commun.

Plus précisément, supposons qu'il existe une borne supérieure M pour toutes les dimensions possibles pour le modèle. Cette supposition, bien que standard pour des méthodes de pénalisation, peut paraître forte en théorie. Elle n'a cependant pas de conséquence pratique puisqu'on se limite toujours à une architecture maximale ne serait-ce qu'à cause des capacités limitées de calcul et de mémoire des ordinateurs.

Soit $\mathcal{W} \subset \mathbb{R}^M$ et $F_{\max}$ un modèle dominant, dont le vecteur de paramètres est $W_{\max} = (w_1, w_2, \ldots, w_M)$. Considérons la famille finie

$$\mathcal{F} = \{(w_1, w_2, \ldots, w_M)/\text{où certaines composantes peuvent être nulles}\} .$$

Pour $F \in \mathcal{F}$, sous-modèle de $F_{\max}$, on note $m(F)$ le nombre des paramètres non nuls, c'est-à-dire la dimension du vecteur paramètre W, et $\mathcal{W}_F$ l'ensemble des valeurs possibles de W. On suppose que le vrai modèle est un sous-modèle de $F_{\max}$, on le note F_0 et la vraie valeur du vecteur paramètre est notée W_0 de dimension $m(F_0)$.

Soit $\hat{W}_{T,F}$ l'estimateur des moindres carrés de W restreint à F,

$$\hat{W}_{T,F} = \text{Arg} \min_{W \in \mathcal{W}_F} S_T(W) \quad ,$$

et $S_T(F)$ (au lieu de $S_T(\hat{W}_{T,F})$) le minimum correspondant de la fonction erreur. Soit aussi $(c(t))$ une suite de nombres réels positifs. Le *contraste des moindres carrés pénalisés, Contrast With Penalty* en anglais, avec le taux de pénalisation $(c(t))$ prend la forme

$$\text{CWP}(T, F) = \frac{S_T(F)}{T} + \frac{c(T)}{T} m(F). \qquad [6]$$

Soit $\hat{F}_T = \text{Arg}\min_{F \in \mathcal{F}} \text{CWP}(T, F)$ le modèle estimé, qui est le résultat de deux minimisations successives pour un T fixé : une minimisation sur un espace continu, pour calculer $\hat{W}_{T,F}$ et $S_T(F)$, et une minimisation sur un espace fini, pour calculer $\hat{F}_T$.

A partir de ces définitions, on montre le résultat suivant dont la preuve complète se trouve dans (Mangeas and Yao, 1997, [MAN 97d]).

Théorème 3 *On suppose que les conditions du théorème (2) sont vérifiées. On suppose aussi que le taux de pénalisation $c(T)$ est tel que*

$$\lim_T \frac{c(T)}{T} = 0, \quad \textit{et} \quad \liminf_T \frac{c(T)}{2 \ln \ln T} > \sigma^2 \frac{\Lambda}{\lambda} \qquad [7]$$

où Λ (resp. λ) est la plus grande (resp. la plus petite) valeur propre de la matrice Σ_0. Alors, le couple $(\hat{F}_T, \hat{W}_{T,\hat{F}_T})$ converge presque sûrement vers la vraie valeur (F_0, W_0), quand T tend vers ∞.

On remarque que là encore il est indispensable que la matrice Σ_0 soit inversible puisque sa plus petite valeur propre λ doit être différente de zéro. La condition d'identifiabilité du modèle (cf Théorème 1) est fondamentale puisque, si elle n'est pas vérifiée, cette valeur propre est nulle.

A partir du théorème (3), on peut donc proposer une méthodologie d'identification presque sûre pour déterminer le *vrai modèle* à l'intérieur de l'ensemble des sous-modèles de $F_{\max}$.

En effet, il suffit de trouver un modèle dont à la fois l'erreur de prévision et son terme de pénalisation soient suffisamment petits. Ce théorème nous guide essentiellement pour choisir une pénalisation convenable.

Ainsi, supposons que γ est une constante positive. Un taux de pénalisation logarithmique de la forme $c(t) = \gamma \ln t$ vérifie clairement les conditions (7). Avec un tel taux de pénalisation, on obtient un *critère des moindres carrés pénalisé* noté BIC* utilisé pour choisir le modèle :

$$\text{BIC}^* = \text{BIC}^*(T, F) = \frac{S_T(F)}{T} + \gamma \frac{\ln T}{T} m(F) \qquad [8]$$

En pratique, on constate qu'il faut choisir γ du même ordre de grandeur que la variance σ^2, car dans ce cas, le critère BIC* ne dépend pas de l'échelle des termes de la série, voir Mangeas, 1997, [MAN 97c]. On peut aussi optimiser la valeur de γ (Mangeas, 1997, [MAN 97a]), mais le gain obtenu sur le modèle ne justifie pas les calculs supplémentaires.

En fait, on peut démontrer les mêmes résultats en utilisant les propriétés de l'estimateur $\tilde{W}_T$ et dans ce cas le critère avec pénalisation utilisé est exactement le critère BIC usuel donné par :

$$\text{BIC} = \ln \frac{S_T(F)}{T} + \frac{\ln T}{T} m(F). \qquad [9]$$

Celui-ci est associé à l'estimateur du maximum de vraisemblance lorsque le bruit est gaussien, et il a l'avantage qu'il n'est pas nécessaire d'introduire une constante γ, car le premier terme en logarithme est assez insensible à la variance du bruit. Par contre ce critère n'est utilisable que si $E\ \varepsilon_t^{1\,2} < \infty$.

On a ainsi obtenu deux critères (proches, mais différents) dont la minimisation conduit en théorie presque sûrement vers le vrai modèle, pour autant qu'on parte d'un modèle dominant qui soit un sur-modèle du vrai modèle.

4. Recherche pratique du vrai modèle

A partir du théorème (3), on peut donc proposer la méthode suivante de détermination du *vrai modèle*.

Pour initialiser l'architecture, nous commençons par prendre toutes les entrées pertinentes (comme on les obtiendrait à partir d'un modèle linéaire de base) et une seule unité cachée. Ensuite on ajoute progressivement des unités dans la couche cachée, en calculant à chaque étape le critère BIC*(resp. BIC), tant que la valeur de ce critère décroît. Quand le critère BIC*(resp. BIC) reste stable ou commence à croître, on arrête la recherche de modèle, et on prend le dernier modèle comme modèle dominant noté $F_{\max}$. On remarquera qu'on suppose avec cette méthodologie que le critère BIC*(resp. BIC) est une fonction convexe du nombre d'unités cachées. C'est d'ailleurs ce qui est fait classiquement lorsqu'on utilise des critères de type BIC sur des modèles linéaires. Il paraît difficile de justifier cette hypothèse théoriquement, cependant cette méthode a l'avantage d'être simple à mettre en oeuvre et donne empiriquement de bons résultats comme le montre l'exemple traité dans la section suivante et les études faites sur données réelles (voir Mangeas, [MAN 97b]).

L'initialisation des paramètres est faite de manière très simple. Avec une seule unité cachée, les coefficients (β_{i1}) sont pris égaux aux valeurs issues du modèle linéaire, le coefficient α_0 est pris égal à la valeur moyenne de la série, les autres sont petits et aléatoires (par exemple entre -0.5 et 0.5).

En fait, beaucoup de chercheurs ont proposé des méthodes et des astuces pour initialiser correctement les paramètres, mais il reste à proposer une méthode efficace qui permettrait de s'approcher d'un minimum global. Nous avons fait des essais en utilisant l'algorithme du recuit simulé, mais les temps de calcul sont très longs et on peut plus simplement répéter plusieurs initialisations différentes et garder la meilleure ! (voir la thèse de Joseph Rynkiewicz [RYN 00]).

Rappelons que $W_{\max} = (w_1, w_2, \ldots, w_M)$ est le vecteur paramètre associé au modèle $F_{\max}$. En principe, pour estimer le vrai modèle, nous devrions explorer exhaustivement la famille finie de tous les sous modèles F du modèle dominant $F_{\max}$ et calculer le BIC* (resp. le BIC) pour chacun. Mais le nombre de ces sous-modèles est exponentiellement grand (comme 2^M) et il est impossible de le faire en pratique. Alors comme en régression linéaire, (Draper and Smith, 1981, [DRA 81]), nous proposons une *méthode statistique pas à pas : Statistical Stepwise Method, SSM*) pour guider la recherche dans $\mathcal{F}$. Une telle stratégie de descente est basée sur la normalité asymptotique de l'estimateur $\hat{W}_T$. Voir [COT 95] pour des présentations antérieures de l'algorithme SSM, avec de nombreux exemples.

Décider la suppression ou non d'un poids w_l est équivalent à construire le test de l'hypothèse nulle "$w_l = 0$" contre l'hypothèse alternative "$w_l \neq 0$", c'est-à-dire un test de Student sur w_l, (en fait un test Gaussien puisque la normalité de l'estimateur du poids est assurée seulement lorsque T est grand).

Décider les suppressions successives des poids $w_{l_1}, w_{l_2}, \ldots, w_{l_L}$ est équivalent à tester itérativement une suite de modèles emboîtés $F_{\max}, F^{l_1}_{\max}, F^{l_1,l_2}_{\max}, \ldots$, où $w_{l_1} = 0, w_{l_1} = w_{l_2} = 0, \ldots$.

Nous utiliserons donc les statistiques de Student comme une aide à l'exploration de la famille $\mathcal{F}$ tout en suivant une trajectoire décroissante du critère BIC* (resp. BIC).

Considérons la suppression d'un poids w_l, où F est le modèle courant et F_l le sous-modèle de F obtenu en posant le paramètre w_l égal à 0 (en supposant que w_l est un paramètre non nul de F). Pour tester F_l contre F, c'est-à-dire "$w_l = 0$" contre "$w_l \neq 0$", on calcule la statistique de Student

$$Q_l = \frac{\hat{w}_l}{\hat{\sigma}(\hat{w}_l)}$$

où

$$\hat{\sigma}(\hat{w}_l) = \frac{\hat{\sigma}}{\sqrt{T}} \sqrt{(\hat{\Sigma}_0^{-1})_{l,l}}$$

est l'écart-type estimé de $\hat{w}_l$.

On accepte "$w_l = 0$" si la statistique Q_l est inférieure en valeur absolue à une valeur seuil lue dans la table de la loi gaussienne et qui dépend du niveau du test voulu. Par exemple, en fixant ce seuil à 1, on risque de se tromper 3 fois sur 10 quand on garde w_l alors qu'il ne fallait pas, mais le risque de se tromper en supprimant w_l est rapidement très petit dès que w_l est différent de 0.

On sait que ce test de Student est rigoureusement équivalent au test de F_l contre F, qui utilise classiquement une statistique de Fisher donnée par :

$$Q_l^2 = \frac{S_T(F_l) - S_T(F)}{S_T(F)/(T - m(F))}. \qquad [10]$$

Mais comme on s'est placé dans un cadre asymptotique, $S_T(F)/(T - m(F))$ converge presque sûrement vers σ^2 et peut être considéré comme une constante pour T grand. Q_l^2 est donc proportionnel à $S_T(F_l) - S_T(F)$.

Or on peut remarquer que la différence des valeurs des BIC* pour ces deux modèles F_l et F vaut :

$$\text{BIC}^*(F_l) - \text{BIC}^*(F) = \frac{1}{T}\{S_T(F_l) - S_T(F)\} - \gamma \frac{\ln T}{T}. \qquad [11]$$

Comme F_l est un sous-modèle de F, $S_T(F_l)$ est toujours supérieure à $S_T(F)$, et donc pour maximiser la décroissance de BIC* il faut minimiser la différence $S_T(F_l) - S_T(F)$, et donc *supprimer si possible le poids w_l tel que Q_l^2 soit minimum*. C'est la règle d'élagage et en pratique, on supprime un poids dès que $Q_l^2 \leq 1$ et/ou BIC* décroît.

Notons que Q_l est connu dès que $\hat{W}_{T,F}$ est estimé dans le modèle F et peut être calculé sans re-estimation des paramètres dans le sous-modèle F_l. Il n'est donc pas

nécessaire d'essayer tous les F_l possibles pour ré-estimer les paramètres et calculer toutes les sommes de carrés $S_T(F_l)$ pour choisir le meilleur sous-modèle F_l.

Les remarques précédentes s'appliquent précisément au cas de la minimisation du critère BIC* mais elles s'appliquent aussi d'une manière approchée au critère BIC. Il n'est pas équivalent de minimiser une différence ou une différence de logarithmes, mais ce n'est pas très différent !

En résumé, la procédure de recherche du vrai modèle est comme suit :

1. Déterminer $F_{\max}$, comme expliqué à la Section 4 et estimer les paramètres (entraîner le réseau).

2. Calculer pour chaque l, le rapport $Q_l = \hat{w}_l / \hat{\sigma}(\hat{w}_l)$.

3. Trouver l_1 réalisant le minimum de ces rapports en valeur absolue.

4. Accepter l'élimination de w_{l_1} seulement si le critère BIC* (resp. BIC) a diminué.

5. En cas de rejet, arrêter le processus d'élimination, et garder le modèle précédent. En cas d'acceptation, ré-estimer les paramètres correspondant au modèle $(F^{l_1}_{\max})$, et répéter l'étape 2), à partir du modèle $(F^{l_1}_{\max})$, pour chercher un autre indice l_2, etc.

La *règle d'arrêt* est tout à fait naturelle : on continue le processus d'élimination tant que le critère BIC* (resp. BIC) décroît. Ceci fournit une critère objectif, bien fondé, reposant sur les propriétés statistiques des estimateurs des poids. On est en réalité capable de décider ce qu'est un "petit" poids qui peut être éliminé.

L'algorithme *SSM* appartient à la famille des *stepwise backward algorithms* largement utilisé en *analyse de régression* (Draper and Smith, 1981, [DRA 81]). Remarquons qu'il est aussi possible d'utiliser une *ascending stepwise method*, pour guider la recherche d'un minimum global du critère BIC* (resp. CIP), voir (Jutten and Chentouf, 1995, [JUT 95]).

La méthode d'élagage SSM est proche de l'algorithme OBD défini par (Le Cun et al., 1990, [CUN 90]), car ils choisissent de la même manière le paramètre candidat à l'élimination. Mais leur algorithme ne fournit pas de critère d'arrêt du processus d'élagage, il a besoin d'une évaluation de la performance qui se fait en dehors sur un ensemble de données externes. Cela exige d'avoir une stratégie de découpage des données en ensemble d'apprentissage et ensemble de validation, et sacrifie un certain nombre de données au détriment de la qualité d'évaluation du modèle. Ici, grâce aux résultats sur l'identification presque sûre du modèle, nous obtenons un critère d'arrêt théorique : le critère BIC* (ou BIC). Le principe est d'arrêter la suppression de paramètres dès que le critère BIC* (ou BIC) croît. Notons qu'avec cette méthode nous gardons donc un maximum de données pour l'apprentissage, ce qui permet d'exploiter au mieux l'information fournie par celles-ci.

5. Exemple

Nous allons tester sur une simulation l'efficacité du critère BIC*. La véritable architecture, le vrai vecteur paramètre W_0 et la variance du bruit gaussien associé sont connus. Le modèle choisi (voir figure 1) est un perceptron à 2 entrées, 2 unités cachées et une sortie, et 8 paramètres (une des unités cachées n'est pas reliée à une des entrées. La fonction f_{W_0} du vrai modèle s'écrit

$$X_t = \tanh(-0.5X_{t-1} - 1.5X_{t-3} + 0.5) + \tanh(X_{t-3} - 0.5) + 0.5 + \varepsilon_t.$$

Le bruit utilisé (ε_t) est i.i.d. Gaussien avec une variance $\sigma_0^2 = 0.1$.

Notre but est de retrouver la vraie architecture ainsi que les bons paramètres. Puisque ce modèle comporte peu de paramètres, on peut comparer la recherche exhaustive avec la méthodologie SSM. La recherche sera effectuée parmi les sous-modèles du modèle dominant décrit figure 2.

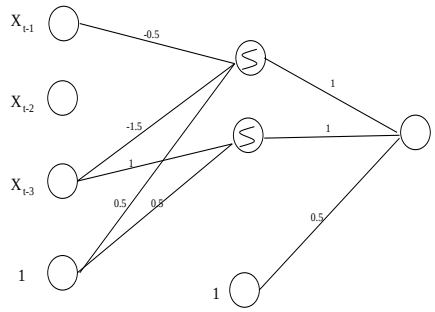

Figure 1. *Vraie architecture*

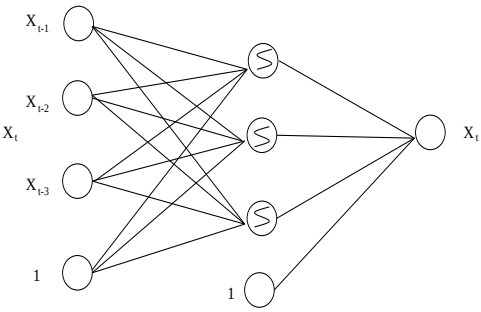

Figure 2. *Architecture dominante*

On simule 50 fois et de façon indépendante une suite de 1000 points. L'architecture du vrai modèle avec 8 connexions est montrée figure 1. Le modèle dominant utilisé, avec 16 connexions, est montré figure 2. Pour le critère BIC*, nous fixons $\gamma = \sigma_0^2 = 0.1$, un choix qui a l'avantage de rendre le poids du terme de pénalisation invariant du niveau du bruit du processus étudié et qui donne en moyenne les meilleurs résultats pratiques (cf [MAN 97b]). Afin d'éviter les minima locaux, l'estimation des paramètres est la meilleure obtenue parmi 10 estimations partant d'initialisations aléatoires différentes.

Le Tableau 1 montre la proportion des trois architectures: T,A,B (minimisant le critère BIC*) retenues par la recherche exhaustive sur 50 simulations. Comme espéré la vraie architecture T apparaît pour 73% des cas. Les autres gagnantes A et B sont très proches de l'architecture T mais avec, respectivement, une connexion de plus et une connexion de moins. Elles apparaissent, respectivement, dans 12% et 10% des cas. Le tableau 2 montre les trois meilleures architectures sélectionnées par SSM pour 50 simulations. On peut remarquer que cette stratégie fournit les mêmes architectures et que la vraie architecture est trouvée dans 62% des cas.

Architecture finale	Pourcentage sur 50 simulations
T	0.73
A	0.12
B	0.10

Tableau 1. *performances de la recherche exhaustive.*

Architecture finale	Pourcentage sur 50 simulations
T	0.62
A	0.22
B	0.16

Tableau 2. *Performances de SSM.*

6. Conclusion

Les principaux résultats de ce papier peuvent être étendus en plusieurs directions. Tout d'abord on peut ajouter des variables exogènes comme entrées du réseau et étudier ce qu'on appellera un modèle NARX. Ensuite, on peut introduire des retards de l'erreur en entrée pour définir un modèle non linéaire de type "ARMA". Enfin, on peut étendre tous les résultats au cas de sorties multi-dimensionnelles. Dans son travail de thèse, Joseph Rynkiewicz a montré qu'il fallait minimiser dans ce cas le déterminant de la matrice de covariance empirique de l'erreur [RYN 00].

Tous ces résultats ont été établis puis testés sur des exemples simulés et sur des données réelles, par exemple la série des tâches solaires, la consommation électrique journalière (Cottrell et al., 1995, [COT 95]), différentes séries réelles ou simulées, (Mangeas, [MAN 97b]), mais aussi plus récemment sur des données de pollution en niveau d'ozone à Paris (cf Rynkiewicz, [RYN 00]).

Attention, tout ceci ne s'applique que pour des séries temporelles stationnaires et l'utilisation de réseaux de neurones ne supprime pas la nécessité des pré-traitements classiques, il faut absolument enlever les tendances et périodicités.

Un logiciel appelé REGRESS, écrit par Joseph Rynkiewicz, permet d'estimer les paramètres d'un réseau de neurones pour un modèle de régression ou d'analyse de séries temporelles. Il utilise la méthode SSM (basée au choix sur les critères BIC* ou BIC) pour choisir la meilleure architecture, et il est disponible à l'adresse internet : "http://panoramix.univ-paris1.fr/SAMOS". Il inclut aussi la possibilité d'estimer des modèles comprenant une chaîne de Markov cachée.

7. Bibliographie

[COT 95] COTTRELL M., GIRARD B., GIRARD Y., MANGEAS M., MULLER C., « Neural modeling for time series : a statistical stepwise method for weight elimination », *IEEE Tr. on Neural Networks*, , n° 6, 1995, p. 1355-1364.

[CUN 90] CUN Y. L., DENKER J., SOLLA S., « Optimal Brain Damage », *Advances in neural Information Processing Systems 2*, Morgan Kaufman, 1990, p. 598-605.

[DRA 81] DRAPER N., SMITH H., *Applied regression analysis*, John Wiley and Sons, New York, 1981.

[DUF 90] DUFLO M., *Méthodes récursives aléatoires*, Masson, 1990.

[FUN 89] FUNAHASHI K., « On the approximate realization of continuous mappings by neural networks », *Neural Networks*, , n° 2, 1989, p. 183-192.

[GUY 95] GUYON X., *Random fields on a network-modeling : Modelling, statistics, and applications*, Probability and its applications, Springer-Verlag, 1995.

[HAN 88] HANNAN E., DEISTLER M., *The statistical theory of linear systems*, Wiley series in probability and mathematical statistics, John Wiley & Sons, 1988.

[HOR 89] HORNIK. K., STINCHCOMBE M., WHITE H., « Multilayer feedforward networks are universal approximators », *Neural Networks*, vol. 4, 1989, p. 359-366.

[JUT 95] JUTTEN C., CHENTOUF R., « A new scheme for incremental learning », *Neural Processing Letters*, vol. 2, n° 1, 1995, p. 1-4.

[MAN 97a] MANGEAS M., « Neural model selection: How to Determine the Fittest Criterion », *Proc. of ICANN'97*, Lausanne, 1997, Springer-Verlag, p. 987-992.

[MAN 97b] MANGEAS M., « Propriétés statistiques des modèles paramétriques non-linéaires de prévision de série temporelles : Etude des réseaux de neurones à propagation directe », Thèse de doctorat, Université de Paris 1, Panthéon-Sorbonne, 1997.

[MAN 97c] MANGEAS M., COTTRELL M., YAO J., « New criterion of identification in the multilayered perceptron modelling », *Proc. of ESANN'97*, Bruges, 1997, DiFacto.

[MAN 97d] MANGEAS M., YAO J., « Sur l'estimateur des moindres carrés des modèles autoregressifs fonctionnels », *CRAS de Paris*, vol. 324-I, 1997, p. 471-474.

[MEY 94] MEYN S., TWEEDIE R., *Markov chains and stochastic stability*, Springer-Verlag, 1994.

[MOO 92] MOODY J., « The effective number of parameters: An Analysis », *Advances in neural Information Processing Systems 4*, Morgan Kaufman, 1992.

[MUR 94] MURATA N., YOSHIZAWA S., AMARI S., « Network Information Criterion-Determining the Number of Hidden Units for an Artificial Neural Network Model », *IEEE Tr. on Neural Networks*, vol. 5, n° 6, 1994, p. 865-872.

[REE 93] REED R., « Pruning algorithms - a survey », *IEEE Tr. on Neural Networks*, vol. 4, n° 5, 1993, p. 740-747.

[RYN 00] RYNKIEWICZ J., « Modèles hybrides intégrant des réseaux de neuronnes artificiels à des modèles de chaînes de Markov cachées : application à la prédiction de séries temporelles. », Thèse de doctorat, Université de Paris 1, Panthéon-Sorbonne, 2000.

[SUS 92] SUSSMANN H., « Uniqueness of the weights for minimal feedforward nets with a given input-output Map », *Neural Networks*, vol. 5, 1992, p. 589-593.

[VAP 95] VAPNIK V., *The nature of statistical learning theory*, Springer, 1995.

[WEI 93] WEIGEND A., GERSHENFELD N. A., *Time series prediction*, Addison-Wesley, 1993.