Apprentissage de réseaux de neurones à fonctions radiales de base avec un jeu de données à entrée-sortie bruitées

Learning radial basis function neural networks with noisy input-output data set

par Abd-Krim SEGHOUANE, Gilles FLEURY

École Supérieure d'Électricité, Service des Mesures, plateau de Moulon, 3 rue Joliot Curie, 91192 Gif-sur-Yvette cedex, France

Abd-krim.Seghouane@supelec.fr, Gilles.Fleury@supelec.fr

résumé et mots clés

Cet article traite du problème de l'apprentissage des réseaux de neurones à fonctions radiales de base pour l'approximation de fonctions non linéaires L_2 de $\mathcal{R}^d$ vers $\mathcal{R}$. Pour ce type de problème, les algorithmes hybrides sont les plus utilisés. Ils font appel à des techniques d'apprentissage non supervisées pour l'estimation des centres et des paramètres d'échelle des fonctions radiales, et à des techniques d'apprentissage supervisées pour l'estimation des paramètres linéaires. Les méthodes d'apprentissage supervisées reposent généralement sur l'estimateur (ou le critère) des moindres carrées (MC). Cet estimateur est optimal dans le cas où le jeu de données d'apprentissage $(\mathbf{z}_i, y_i)_{i=1,2,..,q}$ est constitué de sorties y_i, $i = 1, .., q$ bruitées et d'entrées $\mathbf{z}_i$, $i = 1, .., q$ exactes. Cependant lors de la collecte des données expérimentales il est rarement possible de mesurer l'entrée $\mathbf{z}_i$ sans bruit. L'utilisation de l'estimateur des MC produit une estimation biaisée des paramètres linéaires dans le cas où le jeux de données d'apprentissage est à entrées et sorties bruitées, ce qui engendre une estimation erronée de la sortie. Cet article propose l'utilisation d'une procédure d'estimation fondée sur le modèle avec variables entachées d'erreurs pour l'estimation des paramètres linéaires (pour l'apprentissage supervisé) dans le cas où le jeux de données d'apprentissage est à entrées et sorties bruitées. L'interprétation géométrique du critère d'estimation proposé est établie afin de mettre en évidence son avantage relativement au critère des moindres carrés. L'amélioration des performances en terme d'approximation de fonctions non linéaires est illustrée sur un exemple.

Réseaux de neurones à fonctions radiales de base, modèle avec variables entachées d'erreurs, approximation non linéaire.

abstract and key words

This paper deals with the problem of learning radial basis function neural networks to approximate non linear L_2 function from $\mathcal{R}^d$ to $\mathcal{R}$. Hybrid algorithms are mostly used for this task. Unsupervised learning techniques are used to estimate the center and width parameters of the radial functions and supervised learning techniques are used to estimate the linear parameters. Supervised learning techniques are generally based on the least squares (LS) estimates (or criterion). This estimator is optimal when the training set $(\mathbf{z}_i, y_i)_{i=1,2,..,q}$ is composed of noisy outputs y_i, $i = 1, .., q$ and exactly known inputs $\mathbf{z}_i$, $i = 1, .., q$. However, when collecting the experimental data, it is seldom pos-

sible to avoid noise when measuring the inputs $\mathbf{z}_i$. The use of least squares estimator produces a biased estimation of the linear parameters in the case of noisy input output training data set, which leads to an erroneous output estimation. This paper proposes the use of an estimation procedure based on the error in variables model to estimate the linear parameters (for supervised learning) when the training set is made up of input and output data corrupted by noise. The geometrical interpretation of the proposed estimation criterion is given in order to illustrate its advantage with respect to the least squares criterion. The improved performances in non linear function approximation is illustrated with a simulation example.

Radial basis function neural networks, error in variables model, non linear approximation.

1. introduction

Les réseaux de neurones dits « Feedforward » (RNF) et les réseaux de neurones à fonctions radiales de base (RNFRB) constituent deux classes de modèles paramétriques largement utilisées en identification de systèmes non linéaires [1]. En effet, ces réseaux avec une seule couche cachée peuvent approximer n'importe quelle fonction continue ayant un nombre fini de discontinuités sur tout compact [2][3]. Un net regain d'intérêt pour les RNFRB à été constaté ces dernières années dans divers domaines d'application, tel que le traitement du signal [4], le contrôle [5] et le diagnostic d'erreurs [6]. En effet, ils offrent deux avantages majeurs par rapports aux RNF habituellement utilisés. Pour un problème donné, l'utilisation d'un RNFRB conduit généralement à une structure de modèle (nombre d'unités de la couche cachée) moins complexe que celle produite par un RNF [7]. La complexité de calcul induite par leur apprentissage est moindre que celle induite par l'apprentissage des RNF grâce à l'existence d'algorithmes hybrides [8]. Les performances d'un tel réseau dépendent, pour un choix de fonctions de base [1], du nombre de fonctions constituant la base de fonctions radiales (nombre d'unités de la couche cachée) et de l'estimation des paramètres du réseau. Le choix du nombre optimal d'unité de la couche cachée et l'estimation des paramètres du réseau sont effectués lors d'une phase d'apprentissage au cours de laquelle un ensemble de paires entrée-sortie expérimentales $(\mathbf{z}_i, y_i)_{i=1,2,..,q}$ est utilisé pour permettre aux RNFRB d'acquérir une relation entrée-sortie non linéaire. Les algorithmes hybrides utilisent des techniques d'apprentissage non supervisées pour l'ajustement des paramètres de la couche cachée (et dans certain cas pour le choix du nombre d'unités de la couche cachée) et des techniques d'apprentissage supervisées (fondées sur l'estimateur des moindres carrés) pour l'ajustement des paramètres reliant la couche cachée à la couche de sortie. Ces techniques d'apprentissage supervisées sont optimales dans le cas où uniquement la sortie expérimentale y_i est bruitée. Cependant, ceci est clairement une supposition fausse car toute donnée produite expérimentalement est bruitée. Ces techniques pourront être utilisées dans le cas où les erreurs sur les entrées expérimentales sont négligeables ou dans le cas d'un rapport signal sur bruit en entrée élevé. Dans le cas contraire, elle produiront inévitablement une estimation fortement erronée. Il est donc nécessaire de réfléchir à des algorithmes d'apprentissage qui tiennent compte du bruit qui affecte l'entrée expérimentale $\mathbf{z}_i$. Ce problème est assez nouveau pour la communauté des traiteurs de signaux [9][10] bien qu'étant trés ancien pour la communauté des statisticiens [11]. Il n'a été pris en compte que très récemment dans la communauté des réseaux de neurones. Dans [12], un algorithme pour l'apprentissage des RNF a été proposé. Inversement, il est à noter que l'injection volontaire d'un bruit gaussien aux entrées du jeu de données expérimentales lors de la phase d'apprentissage des RNF améliore leurs performances de généralisation [13][14]. En effet il a été montré que ceci est équivalent à une régularisation [15]. Dans [16] une procédure pour l'apprentissage des RNFRB à partir d'un jeu de données expérimentales à entrées-sorties bruitées a été proposée. Cette procédure s'inspire des développements établis pour interpréter l'amélioration des performances de généralisation par injection de bruit ; l'entrée perturbée étant modélisée par $\mathbf{z} \rightarrow \mathbf{z} + \delta\mathbf{z}$. Dans cet article on s'intéressera à la partie supervisée des algorithmes hybrides lorsque les entrées-sorties expérimentales sont toutes deux bruitées. Plus exactement à l'ajustement des paramètres reliant la couche cachée à la couche de sortie de manière à maximiser les performances des RNFRB pour une base de fonctions choisie. Les paramètres de la couche cachée, étant obtenus par des méthodes d'apprentissage non supervisées, ne sont pas affectés par le bruit d'entrée car ils ne font pas intervenir la relation entrée-sortie recherchée [19].

Le reste de cet article s'organise comme suit, dans la section suivante une formulation du problème est présentée. La section 3 est consacrée à la description de l'architecture et des algorithmes

d'apprentissage des RNFRB. Le critère pour l'estimation des paramètres linéaires (paramètres reliant la couche cachée à la couche de sortie) est développé en section 4. Afin de mettre en évidence la différence du critère proposé avec l'estimateur des moindres carrés classique, son interprétation géométrique est établie en section 5. Un exemple de simulation illustrant les performances du critère proposé pour l'apprentissage des RNFRB en régression non linéaire est donné en section 6.

2. formulation du problème

On s'intéresse à l'estimation par RNFRB d'une relation entre deux variables continues $\mathbf{z} \in \Gamma \subseteq \mathcal{R}^d$ et $y \in \mathcal{R}$ à partir d'un ensemble d'observations expérimentales $(\mathbf{z}_i, y_i)_{i=1,...,N}$. En inférence paramétrique, les couples d'observations sont généralement décrits par l'équation :

$$y_i = f_{RN}(\mathbf{z}_i, \theta) + \epsilon_i, \quad i = 1, ..., q, \tag{1}$$

où $\epsilon_i \in \mathcal{R}$ est échantillonné d'une variable aléatoire $\epsilon \propto N(0, \sigma_\epsilon^2)$, $f_{RN}(., \theta)$ est la fonction décrite par le RNFRB et $\theta \in \Theta$ est le vecteur contenant l'ensemble des paramètres du RNFRB. Ce modèle suppose que les entrées expérimentales $\mathbf{z}_i$ ne sont pas bruitées. Or ceci est une supposition fausse (ou incomplète) car toute donnée produite expérimentalement est bruitée (bruit dû aux appareils de mesure). Il est donc nécessaire de tenir compte du bruit sur les entrées expérimentales et de sa propagation aux données de sortie lors de l'estimation des paramètres du RNFRB. L'utilisation dans ce cas du modèle (1) induit un biais sur les paramètres estimés du modèle et donc sur la sortie prédite. Un modèle tenant compte de ce bruit sur l'entrée expérimentale, donc plus complet et permettant ainsi d'appréhender le problème de la présence d'incertitude à l'entrée, est le modèle avec variables entachées d'erreurs [11][24]. Il décrit les couples d'observations $(\mathbf{z}_i, y_i)_{i=1,...,N}$ par le biais du couple d'équations :

$$\begin{cases} y_i = f_{RN}(\mathbf{x}_i, \theta) + \epsilon_i, \quad i = 1, ..., q, \\ \mathbf{z}_i = \mathbf{x}_i + \boldsymbol{\eta}_i, \end{cases} \tag{2}$$

où $\epsilon_i \in \mathcal{R}$ et $\boldsymbol{\eta}_i \in \mathcal{R}^d$ sont respectivement échantillonnés des variables aléatoires $\epsilon \propto N(0, \sigma_\epsilon^2)$ et $\eta \propto N(0, \sigma_\eta^2 I)$. La première équation du couple (2) décrit la régression $E(y_i/\mathbf{x}_i) = f_{RN}(\mathbf{x}_i, \theta)$ et la seconde décrit l'erreur en la variable $\mathbf{x}_i$. La matrice de covariance du couple $(\epsilon_i, \boldsymbol{\eta}_i)$ sera tout au long de cet article supposée connue et se présentera sous la forme :

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \sigma_{\boldsymbol{\eta}}^2 I \end{pmatrix}, \quad \forall i = 1, ..., q. \tag{3}$$

Ceci traduit pratiquement des mesures indépendantes effectuées dans des conditions et avec des appareils de mesure, identiques d'une mesure à l'autre mais, différents à l'entrée et à la sortie. En pratique, cette matrice pourra être estimée à partir de la répétition des expériences à chaque $i = 1, ..., q$ et dans ce cas le modèle (2) correspondra au modèle moyen [17]. Si M correspond au nombre de répétitions d'expérience, la matrice de covariance du couple de variables aléatoires à partir duquel est échantillonné le couple $(\epsilon_i, \boldsymbol{\eta}_i)$ aura pour matrice de covariance $\boldsymbol{\Sigma} = M^{-1}\boldsymbol{\Phi}$ où $\boldsymbol{\Phi}$ est une matrice fixe définie positive. Le problème dans ce cas consiste à estimer les paramètres du RNBFR ainsi que le vecteur des entrées expérimentales non observées $(\mathbf{x}_i)_{i=1}^q$.

3. architecture et apprentissage

3.1. architecture des RNBFR

Une fonction continue $f : \mathcal{R}^d \longrightarrow \mathcal{R}$, avec $f \in L_2$ peut être décrite par une combinaison linéaire de fonctions élémentaires appelées noyaux. Cette décomposition peut être générée par un réseau de neurones à deux couches où chaque noyau est implanté par une unité de la première couche dite couche cachée, comme représenté sur la figure 1. Chaque noyau de la couche cachée est associé à une région de l'espace d'entrée $\mathcal{R}^d$ appelée région d'action. Dans le cas des RNFRB, la réponse des noyaux dépend de la distance entre l'entrée et un paramètre interne, appelé centre, modulée par un paramètre d'échelle :

$$\phi_l(\mathbf{x}; \mathbf{c}_l, \alpha_l) = \phi\left(\frac{\| \mathbf{x} - \mathbf{c}_l \|^2}{\alpha_l^2} \right). \tag{4}$$

La distance généralement utilisée est la norme euclidienne sur $\mathcal{R}^d$. La région d'action associée au noyau ϕ_l et donc caractérisée par les paramètres $\mathbf{c}_l$ et α_l. L'ensemble des centres et des paramètres d'échelle représentent les paramètres de la couche cachée. La réponse d'un RNFRB à une entrée $\mathbf{x} \in \mathcal{R}^d$ est donnée par la relation affine :

$$\begin{aligned} \hat{f}_{RN}(\mathbf{x}; \mathbf{C}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) &= \lambda_0 + \sum_{l=1}^{L} \lambda_l \phi\left(\frac{\| \mathbf{x} - \mathbf{c}_l \|^2}{\alpha_l^2} \right) \\ &= \lambda_0 + \sum_{l=1}^{L} \lambda_l \phi_l(\mathbf{x}; \mathbf{c}_l, \alpha_l) = \boldsymbol{\lambda}^t \phi, \end{aligned} \tag{5}$$

où $\mathbf{C} = [\mathbf{c}_1, .., \mathbf{c}_L]$ est la matrice dont les colonnes sont les centres du RNFRB, $\alpha = [\alpha_1, .., \alpha_L]^t$ est le vecteur des paramètres d'echelle, $\boldsymbol{\lambda} = [\lambda_0, .., \lambda_L]^t$ est le vecteur des paramètres

linéaires et $\phi = [1, \phi_1, .., \phi_L]^t$ est le vecteur des fonctions radiales de base. Divers types de fonctions peuvent être utilisés comme noyaux ou fonctions de base [1], la fonction gaussienne reste cependant la plus utilisée. Des investigations théoriques ont montré qu'un paramètre d'échelle unifome pour chaque unité cachée est suffisant pour l'approximation non linéaire [3]. Les performances d'un tel réseau en approximation dépendent donc du nombre de centres et par conséquent du nombre de noyaux, de leurs positions, de la valeur des paramètres d'échelle et de la méthode utilisée pour l'apprentissage de la relation entrée-sortie recherchée (ajustement du vecteur des paramètres linéaires $\boldsymbol{\lambda}$).

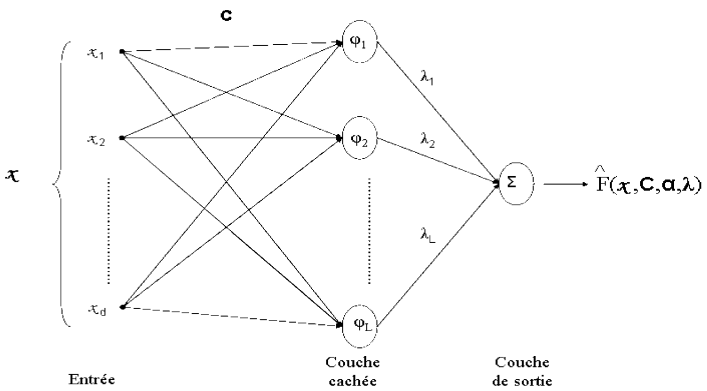

Figure 1. – Réseau de neurones à fonctions radiales de base, RNFRB.

3.2. apprentissage des RNBFR

La matrice des paramètres d'un $RNFRB$, $\mathbf{P} = [\mathbf{C}^t, \boldsymbol{\alpha}, \boldsymbol{\lambda}]$ est de taille $(L+1) \times (d+2)$. Elle est constituée des centres, des paramètres d'échelles et des paramètres linéaires, elle est généralement ajustée de manière à minimiser la fonction coût :

$$J(\mathbf{P}) = \frac{1}{q} \sum_{i=1}^{q} (y_i - f_{RN}(\mathbf{x}_i, \mathbf{P}))^2. \tag{6}$$

La matrice des paramètres optimaux est :

$$\mathbf{P}^* = \arg\min_{\mathbf{P}} J(\mathbf{P}). \tag{7}$$

Deux stratégies sont proposées dans la littérature pour la recherche de $\mathbf{P}^*$. La première se base sur des méthodes supervisées ou directes utilisant des algorithmes coûteux en temps de calcul tel que l'algorithme du gradient [18] pour determiner $\mathbf{P}^*$. La seconde adopte un schéma hybride (moins coûteux en temps de calcul) pour déterminer les composantes de $\mathbf{P}^*$.

3.3. algorithmes hybrides

3.3.1. partie non supervisée

Les paramètres de la couche cachée ($\mathbf{C}$ et $\boldsymbol{\alpha}$) sont obtenus dans cette partie de l'algorithme. L'algorithme hybride le plus utilisé en pratique [8] fait appel à l'algorithme des « centres mobiles » pour l'ajustement des centres $\mathbf{c}_l$ et à l'algorithme des « M-plus proches voisins » pour l'ajustement du vecteur des paramètres d'échelle $\boldsymbol{\alpha}$. D'autres types de techniques d'apprentissage non supervisées peuvent cependant être envisagées [19] (et ses références). Cette partie a pour objectif de déterminer complétement la base de fonctions radiales utilisée pour l'approximation. Les performances d'un RNFRB dépendent de la dimension de cette base et donc du nombre de centres. Diverses approches ont été proposées dans la littérature pour le choix de la dimension optimale de la base de fonctions radiales [20]. Ce problème ressemble au problème du compromis biais-variance traité dans [21]. Cette partie de l'algorithme ne fait pas intervenir la relation entrée sortie recherchée [19]. L'estimation produite à partir d'entrées expérimentales bruitées n'affectera donc pas les performances du RNFRB.

3.3.2. partie supervisée

Les paramètres linéaires propagent les sorties des fonctions radiales formant la base de manière à ce qu'elles soient combinées linéairement en sortie. Leur estimation est généralement obtenue par l'estimateur des moindres carrées minimisant ainsi l'erreur au sens L_2 sur le jeu de données expérimentales $(\mathbf{z}_i, y_i)_{i=1,..,q}$. Afin de simplifier les notations, $f_{RN}(\mathbf{x}_i, \mathbf{P})$ sera noté $f_{RN}(\mathbf{x}_i, \boldsymbol{\lambda})$ (la matrice $\mathbf{C}$ et le vecteur des paramètres $\boldsymbol{\alpha}$ sont déterminés dans la partie précédente de l'algorithme). Pour le jeu de données $(\mathbf{z}_i, y_i)_{i=1,..,q} = (\mathbf{Z}^t, \mathbf{y}^t)$, la matrice des centres $\mathbf{C}$ et le vecteur des paramètres d'échelles $\boldsymbol{\alpha}$, l'estimation $\hat{\boldsymbol{\lambda}}$ de $\boldsymbol{\lambda}$ est

$$\begin{aligned}
\hat{\boldsymbol{\lambda}} &= \arg\min_{\boldsymbol{\lambda}} J(\mathbf{P}) \\
&= \arg\min_{\boldsymbol{\lambda}} \frac{1}{q} \parallel \mathbf{y} - \hat{f}_{RN}(\mathbf{Z}; \mathbf{C}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) \parallel_2^2 \\
&= \arg\min_{\boldsymbol{\lambda}} \parallel \mathbf{y} - \boldsymbol{\Phi}^t(\mathbf{Z}, \mathbf{C}, \boldsymbol{\alpha})\boldsymbol{\lambda} \parallel_2^2 \\
&= [\boldsymbol{\Phi}(\mathbf{Z}, \mathbf{C}, \boldsymbol{\alpha})\boldsymbol{\Phi}^t(\mathbf{Z}, \mathbf{C}, \boldsymbol{\alpha})]^{-1}\boldsymbol{\Phi}(\mathbf{Z}, \mathbf{C}, \boldsymbol{\alpha})\mathbf{y},
\end{aligned}$$

où $\boldsymbol{\Phi}$ est une matrice de taille $(L+1) \times q$ dont les (éléments) colonnes sont les réponses des unités de la couche cachée. Hormis [16] où il est proposé une méthode de régularisation, l'ensemble des algorithmes proposés dans la littérature pour l'estimation de $\boldsymbol{\lambda}$ supposent que le jeu de données expérimentales est décrit par le modèle de régression (1). Ils ne tiennent donc pas compte du bruit pouvant affecter l'entrée. Ils produiront donc dans ce cas une estimation erronée de $\boldsymbol{\lambda}$ [22].

4. critère pour la partie supervisée de l'algorithme

Le critère développé ici peut être utilisé pour la phase d'apprentissage supervisé dans une méthode hybride. Le but étant la recherche d'une estimation du vecteur des paramètres linéaires $\boldsymbol{\lambda}$ qui maximisent les performances du RNFRB sur un jeu de données expérimentales à entrée et sortie bruitées. Étant donné le modèle à variables entachées d'erreurs (2) et soit $\mathbf{X} = (\mathbf{x}_1, .., \mathbf{x}_q)$ la matrice des vecteurs d'entrées expérimentales non observés. Dans ce cas, en plus du vecteur des paramètres $\boldsymbol{\lambda}$, la matrice $\mathbf{X}$ est inconnue. Il sera supposé tout au long de cet article que le vecteur composite $(\mathbf{X}, \boldsymbol{\lambda})$ est tel que

$$(\mathbf{X}, \boldsymbol{\lambda}) \in \Xi = \Gamma^q \times \Lambda, \quad \Gamma^q \subset \mathcal{R}^{d \times q}, \Lambda \subset \mathcal{R}^{L+1},$$

et que

$$\Gamma^q = \Gamma_1 \times .... \times \Gamma_q, \quad \Gamma_i \subset \mathcal{R}^d.$$

L'estimateur au sens du maximum *a posteriori* est défini, sous l'hypothèse d'un *a priori* $\wp(\mathbf{X}, \boldsymbol{\lambda})$ uniforme, par la maximisation de la densité des paires de données d'observations $(\mathbf{z}_i, y_i)_{i=1,...,q} = (\mathbf{Z}, \mathbf{y})$ générées par le modèle (2)

$$\begin{aligned} \wp(\mathbf{Z}, \mathbf{y}/\mathbf{X}, \boldsymbol{\lambda}) &= \wp(\mathbf{Z}/\mathbf{X}, \boldsymbol{\lambda}) \wp(\mathbf{y}/\mathbf{X}, \boldsymbol{\lambda}) \\ &= \wp(\mathbf{Z}/\mathbf{X}) \wp(\mathbf{y}/\mathbf{X}, \boldsymbol{\lambda}) \\ &= \prod_{i=1}^{q} \wp(\mathbf{z}_i/\mathbf{x}_i) \wp(y_i/\mathbf{x}_i, \boldsymbol{\lambda}) \\ &= \prod_{i=1}^{q} \wp_\eta(\mathbf{z}_i - \mathbf{x}_i) \wp_\epsilon(y_i - f_{RN}(\mathbf{x}_i, \boldsymbol{\lambda})), \end{aligned} \tag{8}$$

où $\wp_\epsilon$ et $\wp_\eta$ représentent respectivement les lois des densités du bruit en sortie et du bruit en entrée qui sont supposés gaussiennes de moyenne nulle, de variance σ_ϵ^2 et covariance $\sigma_\eta^2 I_d$. Cette quantité traduit la vraisemblance des données et l'estimateur du maximum *a posteriori* correspond dans ce cas à l'estimateur du maximum de vraisemblance. L'opposée de la log vraisemblance est dans ce cas

$$\begin{aligned} L(\mathbf{X}, \boldsymbol{\lambda}) &= \sum_{i=1}^{q} \left(\left(\frac{y_i - f_{RB}(\mathbf{x}_i, \boldsymbol{\lambda})}{\sigma_\varepsilon} \right)^2 + \left(\frac{\| \mathbf{z}_i - \mathbf{x}_i \|}{\sigma_\eta} \right)^2 \right) \\ &= \sum_{i=1}^{q} [y_i - f_{RB}(\mathbf{x}_i, \theta), \mathbf{z}_i - \mathbf{x}_i] \boldsymbol{\Sigma}^{-1} [y_i - f_{RB}(\mathbf{x}_i, \theta), \mathbf{z}_i - \mathbf{x}_i]^t \\ &= \sum_{i=1}^{q} l(y_i, \mathbf{z}_i; \mathbf{x}_i, \boldsymbol{\lambda}). \end{aligned} \tag{9}$$

L'estimateur $\hat{\boldsymbol{\lambda}}$ de $\boldsymbol{\lambda}$ s'obtient donc par la minimisation de $L(\mathbf{X}, \boldsymbol{\lambda})$

$$\hat{\boldsymbol{\lambda}} = \arg \min_{\boldsymbol{\lambda} \in \Lambda} \min_{\mathbf{X} \in \Gamma^q} L(\mathbf{X}, \boldsymbol{\lambda}). \tag{10}$$

La difficulté essentielle de l'estimation proposée réside dans l'optimisation du critère proposé.

4.1. procédure d'estimation

L'application direct du critère d'estimation présenté ci-dessus est délicate car elle necessite une optimisation sur un grand nombre de variables. L'utilisation de l'hypothèse :

$$\Xi = \Gamma^q \times \Lambda,$$

et les résultats de [23] permettent la construction d'une procédure itérative d'estimation utilisant une approximation au premier ordre en $\mathbf{x}$. Soit $\overline{\mathbf{X}}$ une estimation initiale des entrées non observées et soit $\overline{\boldsymbol{\lambda}}$ l'estimation initiale de $\boldsymbol{\lambda}$ obtenue par :

$$\overline{\boldsymbol{\lambda}} = \arg \min_{\boldsymbol{\lambda} \in \Lambda} L(\overline{\mathbf{X}}, \boldsymbol{\lambda}). \tag{11}$$

Un développement en série de Taylor au premier ordre de $f(\mathbf{x}, \overline{\boldsymbol{\lambda}})$ au voisinage de $\overline{\mathbf{x}}$ s'écrit :

$$f_{RN}(\mathbf{x}, \overline{\boldsymbol{\lambda}}) = f_{RN}(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) + f_{\mathbf{x}}(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) \Delta \mathbf{x}, \tag{12}$$

où $\Delta \mathbf{x} = \mathbf{x} - \overline{\mathbf{x}}$ et $f_x(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) = \dfrac{\partial f_{RN}(\mathbf{x}, \overline{\boldsymbol{\lambda}})}{\partial \mathbf{x}}|_{\mathbf{x}=\overline{\mathbf{x}}}$.
Si l'on pose

- $a_i = y_i - f(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$,
- $\mathbf{b}_i = \mathbf{z}_i - \overline{\mathbf{x}}_i$ et
- $\Delta y_i = f(\mathbf{x}, \overline{\boldsymbol{\lambda}}) - f(\overline{x}, \overline{\theta}) = f_x(\overline{\mathbf{x}}_i, \overline{\boldsymbol{\lambda}}) \Delta \mathbf{x}_i$,

l'approximation locale de $L(\mathbf{X}, \boldsymbol{\lambda})$ au premier ordre s'écrit :

$$\begin{aligned} L(\mathbf{X}, \overline{\boldsymbol{\lambda}}) &= \sum_{i=1}^{q} l(y_i, \mathbf{z}_i; \mathbf{x}_i, \overline{\boldsymbol{\lambda}}) \\ &\simeq \sum_{i=1}^{q} [a_i - \Delta y_i, \mathbf{b}_i - \Delta \mathbf{x}_i] \boldsymbol{\Sigma}^{-1} [a_i - \Delta y_i, \mathbf{b}_i - \Delta \mathbf{x}_i]^t \\ &= (\mathbf{a} - \mathbf{F} \Delta \mathbf{X}) \boldsymbol{\Sigma}_\varepsilon^{-1} (\mathbf{a} - \mathbf{F} \Delta \mathbf{X})^t + (\mathbf{B} - \Delta \mathbf{X}) \boldsymbol{\Sigma}_\eta^{-1} (\mathbf{B} - \Delta \mathbf{X})^t, \end{aligned} \tag{13}$$

où $\mathbf{a}$ est un vecteur de dimension q, $\mathbf{B}$ est une matrice de dimension $q \times d$, $\mathbf{F}$ est une matrice diagonale de dimension $q \times q$ où les éléments diagonaux sont les dérivées $f_x(\overline{\mathbf{x}}_i, \overline{\boldsymbol{\lambda}})$, $i = 1..., q$, $\boldsymbol{\Sigma}_\varepsilon^{-1} = \sigma_\varepsilon^{-1} I_q$ et $\boldsymbol{\Sigma}_\eta^{-1} = \sigma_\eta^{-1} I_d$. La matrice $\Delta \widehat{\mathbf{X}} = [\Delta \hat{\mathbf{x}}_1, ..., \Delta \hat{\mathbf{x}}_q]^t$ minimisant cette somme s'exprime selon :

$$\Delta\widehat{\mathbf{X}} = (\mathbf{F}^t\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{F} + \boldsymbol{\Sigma}_\eta^{-1})^{-1}(\mathbf{F}^t\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{a} + \boldsymbol{\Sigma}_\eta^{-1}\mathbf{B}), \quad (14)$$

elle permet l'actualisation de l'estimation de $\mathbf{X}$ selon :

$$\widehat{\mathbf{X}} = \overline{\mathbf{X}} + \Delta\widehat{\mathbf{X}}, \quad (15)$$

et donc celle de $\boldsymbol{\lambda}$. L'estimateur initial de $\mathbf{X}$ le plus approprié est $\mathbf{Z}$ [17] (page 25). L'inconvénient majeur de la procédure d'estimation proposée, est que le modèle et donc la base de fonctions radiales devra être convenablement choisi, car un passage par toutes les données expérimentales (ce qui est mauvais en régression) produit un $\Delta\widehat{\mathbf{X}}$ nul. Les propriétés statistiques de la procédure d'estimation ainsi proposée on été établies dans [17] (aux pages 42-47).

5. interprétation géométrique

Afin de mettre en évidence l'apport du modèle à variables entachées d'erreurs (2) relativement au modèle de régression classique (1) sur l'estimation par maximum de vraisemblance, une interprétation géométrique de l'estimateur obtenu est décrite ici. De l'hypothèse :

$$\Gamma^q = \Gamma_1 \times \Gamma_2 \times ..... \times \Gamma_q, \quad \Gamma_i \subset \mathbf{R}^d,$$

l'estimateur du maximum de vraisemblance est obtenu par minimisation relativement à $\boldsymbol{\lambda}$ selon :

$$\min_{\mathbf{X}\in\Gamma^q} L(\mathbf{X}, \boldsymbol{\lambda}) = \sum_{i=1}^{q} \min_{\mathbf{x}_i\in\Gamma_i} \left(\frac{1}{\sigma_\varepsilon^2}(y_i - f_{RN}(\mathbf{x}_i, \boldsymbol{\lambda}))^2 + \frac{1}{\sigma_\eta^2}(\mathbf{z}_i - \mathbf{x}_i)^2 \right). \quad (16)$$

Dans le cas théorique, chaque vecteur de l'estimation de la matrice des entrées non observées :

$$\widehat{\mathbf{X}}(\boldsymbol{\lambda}) = (\hat{\mathbf{x}}_1(\boldsymbol{\lambda}), ..., \hat{\mathbf{x}}_q(\boldsymbol{\lambda}))$$

(obtenue par maximisation de la vraisemblance) satisfait l'équation normale :

$$\frac{1}{\sigma_\varepsilon^2}(y_i - f_{RN}(\hat{\mathbf{x}}_i(\boldsymbol{\lambda}), \boldsymbol{\lambda}))f_{\mathbf{x}}(\hat{\mathbf{x}}_i(\boldsymbol{\lambda}), \boldsymbol{\lambda}) + \frac{1}{\sigma_\eta^2}(\mathbf{z}_i - \hat{\mathbf{x}}_i(\boldsymbol{\lambda})) = 0, \quad (17)$$

où

$$f_{\mathbf{x}}(\hat{\mathbf{x}}_i(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = \frac{\partial f_{RN}(\hat{\mathbf{x}}(\boldsymbol{\lambda}), \boldsymbol{\lambda})}{\partial x}|_{\mathbf{x}=\hat{\mathbf{x}}_i(\boldsymbol{\lambda})}.$$

La tangente $t(., \boldsymbol{\lambda})$ de la courbe d'estimation $f_{RN}(., \boldsymbol{\lambda})$ de la régression à l'entrée estimée $\hat{\mathbf{x}}_i(\boldsymbol{\lambda})$ s'écrit

$$t(\mathbf{x}, \boldsymbol{\lambda}) = f_{RN}(\hat{\mathbf{x}}_i(\boldsymbol{\lambda}), \boldsymbol{\lambda})) + (\mathbf{x} - \hat{\mathbf{x}}_i(\boldsymbol{\lambda}))f_{\mathbf{x}}(\hat{\mathbf{x}}_i(\boldsymbol{\lambda}), \boldsymbol{\lambda}). \quad (18)$$

Soit $t_a(., \boldsymbol{\lambda})$ la tangente ajustée de la courbe d'estimation $f_{RN}(., \boldsymbol{\lambda})$ de la régression à l'entrée estimée $\hat{\mathbf{x}}_i(\boldsymbol{\lambda})$:

$$t_a(\mathbf{x}, \boldsymbol{\lambda}) = f_{RN}(\hat{\mathbf{x}}_i(\boldsymbol{\lambda}), \boldsymbol{\lambda})) + (\mathbf{x} - \hat{\mathbf{x}}_i(\boldsymbol{\lambda}))\frac{\sigma_\eta^2}{\sigma_\varepsilon^2}f_{\mathbf{x}}(\hat{\mathbf{x}}_i(\boldsymbol{\lambda}), \boldsymbol{\lambda}). \quad (19)$$

La multiplication de l'équation normale précédente (17) par $(\mathbf{x} - \hat{\mathbf{x}}_i(\boldsymbol{\lambda}))$ permet de déduire :

$$\begin{pmatrix} \mathbf{z}_i \\ y_i \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{x}}_i(\boldsymbol{\lambda}) \\ f_{RN}(\hat{\mathbf{x}}_i(\boldsymbol{\lambda}), \boldsymbol{\lambda}) \end{pmatrix} \perp \begin{pmatrix} \mathbf{x} \\ t_a(\mathbf{x}, \boldsymbol{\lambda}) \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{x}}_i(\boldsymbol{\lambda}) \\ t_a(\hat{\mathbf{x}}_i(\boldsymbol{\lambda}), \boldsymbol{\lambda}) \end{pmatrix}; \quad \forall\boldsymbol{\lambda} \in \Lambda \text{ et } \mathbf{x} \in \Gamma. \quad (20)$$

Pour $\sigma_\eta^2 = \sigma_\varepsilon^2$, on en déduit que le segment de droite joignant le point expérimental $(\mathbf{z}_i, y_i)$ à la courbe d'estimation de la régression à l'entrée estimée $\hat{\mathbf{x}}_i(\boldsymbol{\lambda})$ est orthogonale à la tangente à cette courbe en ce point. Ceci est illustré sur la figure 2. Il est à noter que cette orthogonalité est respectée pour tous les vecteurs de paramètres estimés. L'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\lambda}}$ (qui dans notre cas est obtenu avec des erreurs gaussiennes) est celui qui minimise la somme des distances orthogonales entre les paires de données expérimentales et la courbe d'estimation. Dans le cas où σ_η^2 et σ_ϵ^2, sont différents entre eux, le rapport $\frac{\sigma_\epsilon^2}{\sigma_\eta^2}$ carac-

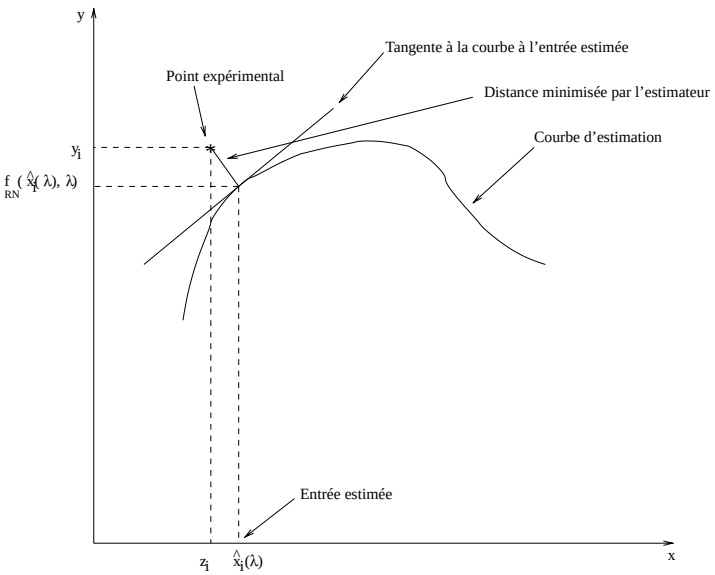

Figure 2. – Illustration géométrique.

térise l'angle entre le segment joignant le point expérimental $(\mathbf{z}_i, y_i)$ à la courbe d'estimation à l'entrée estimée $\hat{\mathbf{x}}_i$ et la tangente en ce point. Il est à noter que l'utilisation naïve de l'estimateur des moindres carrées produit une minimisation de la somme des distances verticales entre les paires de données expérimentales et la courbe d'estimation aux entrées mesurées $\mathbf{z}_i$, $i = 1, ..., q$, ce qui engendre une estimation biaisée. Cette interprétation correspond à la maximisation de la vraisemblance pour une modélisation des données par le modèle (1).

6. application à l'estimation des paramètres linéaire d'un RNFRB

Afin d'illustrer la méthode d'apprentissage développée dans cet article, nous allons utiliser un RNFRB pour l'approximation de la fonction $f : \mathcal{R} \longrightarrow \mathcal{R}$ définie par,

$$y = f(x) = (1 + 2x - x^2)e^{-x^2}, \tag{21}$$

avec $x \in [-4, 4]$, à partir d'un jeu de données d'apprentissage généré par le modèle

$$\begin{cases} y_i = f(x_i, \theta) + \epsilon_i, & i = 1, ..., 20, \\ z_i = x_i + \eta_i, \end{cases} \tag{22}$$

où $\eta_i \sim N(0, \sigma_e^2)$, $\epsilon_i \sim N(0, \sigma_S^2 = 0, 1)$, et x est uniformément réparti sur $[-4, 4]$. Afin d'effectuer une comparaison des résultats, deux méthodes d'apprentissage hybrides sont utilisées. La partie non supervisée des deux méthodes est la même. Ces méthodes diffèrent par leur partie supervisée. Dans la première méthode, l'estimateur des moindres carrés est utilisé pour l'ajustement des paramètres linéaires. Dans la seconde, c'est la procédure d'estimation proposée qui est utilisée. La valeur des centres des fonctions radiales est prise parmi les entrées du jeu de données d'apprentissage et leur nombre est choisi de manière adaptative. Le nombre de centres (et par conséquent le nombre d'unités de la couche cachée) est augmenté (en commençant à 1) jusqu'à obtenir une erreur acceptable sur le jeu de données d'apprentissage (de l'ordre de σ_S^2) avec la première méthode d'apprentissage. Un mauvais choix du nombre d'unités de la couche cachée (conduisant à un sur-apprentissage du RNFRB) provoque un blocage de la procédure d'estimation proposée. En effet, un passage de la courbe d'estimation par toutes les données expérimentales engendre un $\Delta\widehat{\mathbf{X}}$ nul. Les paramètres d'échelle ont été pris de valeurs identiques égales à 0.5. Cette valeur a été choisie après avoir testé toutes les valeurs entre 0 et 1 avec un pas de 0.1. L'erreur quadratique entre la fonction cible f et la sortie du RNFRB sur 801 points régulièrement espacés entre $[-4, 4]$ a été prise comme mesure de la généralisation du RNFRB. Dans le tableau ci-dessous, est reportée la moyenne de cette erreur $E(f, f_{RNi}), i = 1, 2$ sur 100 jeux de données générés par le modèle (22) avec différentes variances d'entrée σ_η^2.

Table 1. – Erreurs de généralisation.

σ_η^2	0.2	0.175	0.15	0,125	0.1	0.075	0.05	0.025
$E(f, f_{RN1})$	41,59	28,88	21,72	21,89	11,29	10,74	8,91	6,52
$E(f, f_{RN2})$	26,28	23,08	17,14	16,03	9,99	9,23	7,39	5,26

La figure 3 illustre la fonction cible, un exemple de jeu de données d'apprentissage produit avec $\sigma_\eta^2 = 0, 2$ et les estimations produites par la sortie du RNFRB dont les paramètres ont été ajustés par les méthodes 1 et 2. Les erreurs quadratiques sont respectivement dans ce cas (sur 801 points régulièrement espacés entre $[-4, 4]$) de $34, 48$ et de $12, 45$. Sur la figure 4 il a été reporté les histogrammes des erreurs commises sur les entrées d'apprentissage non observées avant et après application de la procédure d'estimation proposée pour le jeu de la figure 3. On remarque bien sur cette figure l'augmentation du nombre des erreurs proches de zéro et donc l'amélioration que produit la procédure d'estimation proposée sur l'estimation des entrées d'apprentissage non observées. L'évolution et la convergence de l'estimation des paramètres linéaires obtenus par la méthode d'estimation proposée (pour le jeu de la figure 3) sont illustrées sur la figure 5. La figure 6 illustre quant à elle la décroissance de la norme L_2 du vecteur $\Delta\mathbf{x}$ en fonction du nombre d'itération.

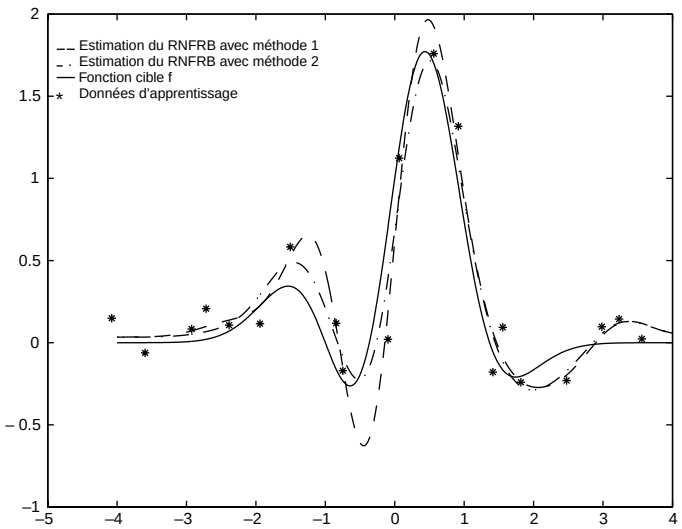

Figure 3. – Exemple de résultats d'estimation.

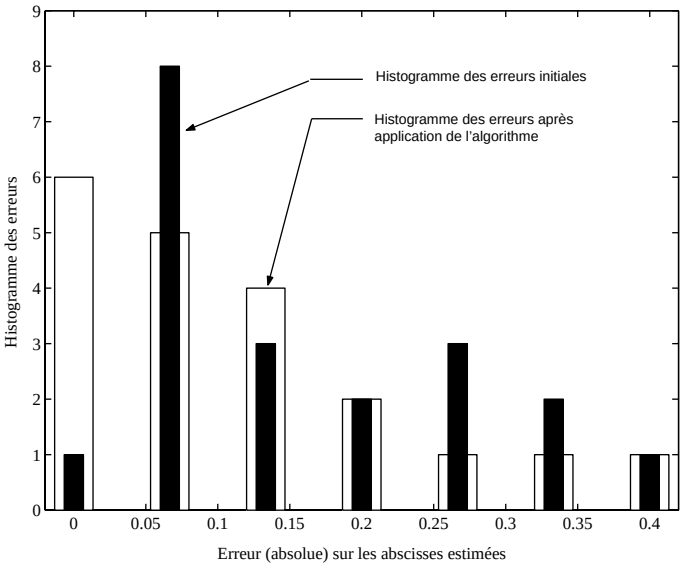

Figure 4. – Exemple de l'estimation des entrées d'apprentissage non observées.

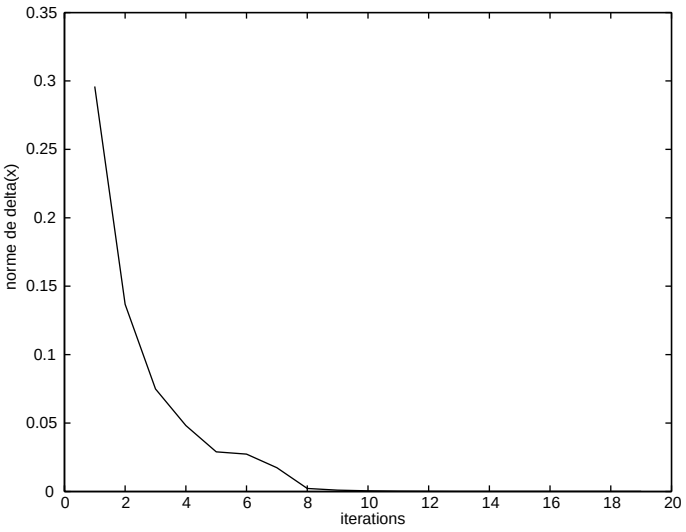

Figure 6. – Évolution et convergence de la norme L_2 du vecteur $\Delta \mathbf{x}$.

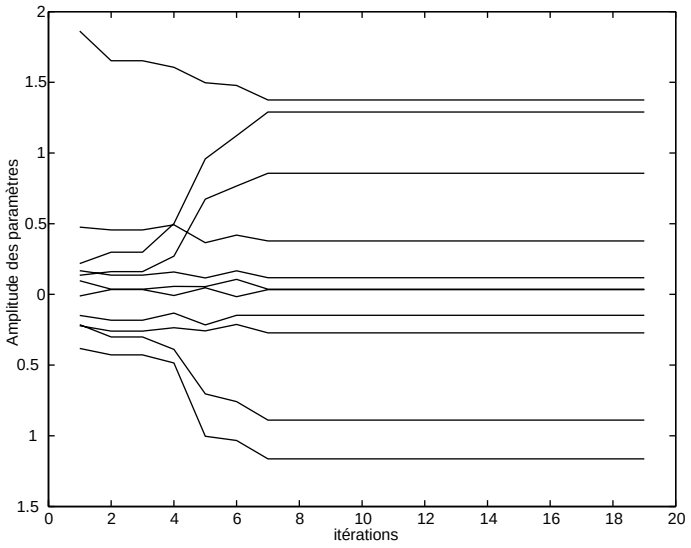

Figure 5. – Évolution et convergence de l'estimation des paramètres par la procédure proposée.

7. conclusion

L'identification des paramètres d'un RNFRB à partir d'un jeu de données expérimentales constitué d'entrées et de sorties bruitées a été étudiée. L'utilisation des algorithmes hybrides existant ne permet pas l'obtention d'une bonne estimation de la réponse du RNFRB pour une entrée donnée. En effet, ces algorithmes utilisent l'estimateur des moindres carrés pour l'ajustement des paramètres linéaires, ne tenant ainsi pas compte de l'erreur présente sur les entrées expérimentales et de leur propagation aux sorties expérimentales. Ceci va donc générer une erreur sur l'estimation des paramètres linéaires du RNFRB, l'estimation de la valeur des centres et des paramètres d'échelles ne faisant pas intervenir la relation entrée-sortie recherchée. Pour la réduction de cette erreur, nous avons indroduit une méthode d'apprentissage supervisé reposant sur la minimisation d'un critère fondé sur le modèle avec variables entachées d'erreurs. L'estimation nécessite dans ce cas la connaissance des variances (matrice de covariance dans le cas multidimensionnel) du bruit à l'entrée σ_η^2 et à la sortie σ_ϵ^2. Une estimation de ces grandeurs peut être obtenue en pratique par la répétition des mesures. L'implémentation du critère proposé nécessite une optimisation sur un grand nombre de variables, car en plus des paramètres linéaires, les entrées expérimentales non observées doivent être estimées. Afin de mettre en évidence la différence entre l'estimation produite par la minimisation du critère proposé et l'estimation produite par moindres carrés, son interprétation géométrique a été établie et comparée à celle de l'estimateur des moindres carrés. Le gain obtenu par l'utilisation de cette méthode est la réduction du biais sur les paramètres linéaires [17] (pages 37-38) et donc l'amélioration des performances d'approximation. Un exemple de simulation montrant cette amélioration lors de l'utilisation de ce critère dans la phase supervisée de l'apprentissage a été présenté. La modification de la valeur des centres et des paramètres d'échelle à chaque actualisation de l'estimation des entrées expérimentales pourrait aussi être effectuée.

RÉFÉRENCES

[1] S. Chen, and S. A. Billings, « Neural networks for nonlinear system modelling and identification », *International Journal of Control*, Vol. 56, pp. 319-346, 1992.

[2] K. Hornick, M. Stinchcombe, and H. White, « Multilayer feedforward networks are universal approximators », *Neural Networks*, Vol. 2, pp. 359-366, 1989.

[3] J. Parks, and I. W. Sandberg, « Universal approximation using radial-basis function networks », *Neural Computation*, Vol. 3, pp. 246-257, 1991.

[4] S. chen, « Nonlinear time series modeling and prediction using Gaussian RBF networks with enhanced clustering and RLS learning », *IEEE Signal Processing Letters*, Vol. 31, pp. 117-118, 1995.

[5] S. Fabri and V. Kadirkamanathan, « Dynamic structure neural networks for stable adaptive control of nonlinear systems », *IEEE Transaction on Neural Networks*, Vol. 7, pp. 1151-1167, 1996.

[6] D. L. Yu, J. B. Gomm and D. Williams, « Sensor fault diagnosis in a chemical process via RBF neural networks », *Control Engineering and Practice*, Vol. 7, pp. 49-55, 1999.

[7] S. Lee and R .M. Kil, « A Gaussian potential function network with hierarchically self-organizing learning », *Neural Networks*, Vol. 4, pp. 207-224, 1991.

[8] J. E. Moody, and C. J. Darken, « Fast learning in networks of locally-tuned processing units », *Neural Computation*, Vol. 1, pp. 281-294, 1989.

[9] J. K. Tungnait and Y. Ye, « Stochastic system identification with noisy input-output measurements using polyspectra », *IEEE Transaction on Signal Processing*, Vol. 40, pp. 670-683, 1995.

[10] V. Krishnamurthy and A. Logothetis, « Iterative and recursive estimators for hidden markov errors in variables models », *IEEE Transaction on Signal Processing*, Vol. 44, pp. 639-629, 1996.

[11] W. Abraham, « The fitting of straight lines if both variables are subject to error », *Annals of Mathematical Statistics*, Vol. 11, pp. 284-300, 1940.

[12] J. Van Gorp, J. Schoukens and R. Pintelon, « Learning neural networks with noise inputs using the errors in variables approach », *IEEE Transaction on Neural Networks*, Vol. 11, pp. 402-414, 2000.

[13] J. Sietsma and R. Dow, « Creating artificial neural networks that generalize », *Neural Networks*, Vol. 4, pp. 67-79, 1991.

[14] K. Matsuoka, « Noise injection into inputs in back-propagation learning », *IEEE Transactions on Neural Networks*, Vol. 22, pp. 436-440, 1992.

[15] C. M. Bishop, « Training with noise is equivalent to Tikhonov regularization », *Neural Computation*, Vol. 7, pp. 108-116, 1995.

[16] N. W. Townsend and L. Tarassenko, « Estimations of error bounds for RBF networks », *IEE Artificial Neural Networks*, Vol. 7, pp. 227-232, 1997.

[17] A. K. Seghouane, « Choix de structures de modèles pour traitement robuste », *Thèse de Doctorat*, Université Paris Sud, Orsay, décembre 2002.

[18] N. B Karayiannis, « Reformulated radial basis neural networks trained by gradient descent », *IEEE Transactions on Neural Networks*, Vol. 10, pp. 657-671, 1999.

[19] Z. Uykan, C. Guzelis, M. E. Celebi and H. N. Koivo, « Analysis of input-ouput clustering for determining centers of RBFN », *IEEE Transactions on Neural Networks*, Vol. 11, pp. 851-858, 2000.

[20] J. Barry, D. Li Yu, « Selecting radial basis function network centers with recursive orthogonal least squares training », *IEEE Transactions on Neural Networks*, Vol. 11, pp. 306-314, 2000.

[21] S. Geman, E. Bienenstock and R. Doursat, « Neural networks and the bias/variance dilemma », *Neural Computation*, Vol. 4, pp. 1-58, 1992.

[22] S. D. Hodges and P. G. Moore, « Data uncertainties and least squares regression », *Applied Statistics*, Vol. 21, pp. 185-195, 1972.

[23] J. Pfanzagl, « On the measurability and consistency of minimum contrast estimates », *Metrika*, Vol. 14, pp. 247-276, 1969.

[24] P. M. Reilly, and H. P. Leal, « A Bayesian study of the error-in-variables model », *Technometrics*, Vol. 23, pp. 221-231, 1981.

Manuscrit reçu le 2 juillet 2001

LES AUTEURS

Abd-Krim SEGHOUANE

Abd-Krim SEGHOUANE est né en 1973. Il a reçu le diplôme d'ingénieur en électronique de l'Institut d'Électronique de l'Université Mouloud Mammeri, Tizi-ouzou, Algérie en 1996, le diplôme de Magistère en traitement du signal de l'École Militaire Polytechnique d'Alger en 2000 et le titre de docteur en sciences de l'Université Paris XI, Orsay en 2002. Ses centres d'intérêts sont l'analyse de données et l'apprentissage statistique.

Gilles A.FLEURY

Gilles A.FLEURY est né à Bordeaux le 8 janvier 1968. Il a reçu en 1990 le diplôme d'ingénieur de l'École Supérieure d'Électricité, Gif-sur-Yvette, et a obtenu en juillet 1994, le titre de docteur en sciences de l'Université de Paris XI, Orsay. Il est actuellement professeur au sein du service des mesure de Supélec. Après des travaux portant sur les problèmes inverses et l'optimisation d'instrument, ses recherches se sont orientées vers la modélisation optimale et le traitement des données à échantillonnage non uniforme.